Historical Gazetteer System Integration:
CHGIS, Regnum Francorum, and GeoNames

Merrick Lex Berman *(CGA, Harvard University)* and Johan Åhlfeldt *(Regnum Francorum Online)*

23 Feb 2012

Abstract: Integration of digital gazetteers, involving the disambiguation of unique places and conflation of duplicates or variant placeneames, has been the focus of many theoretical papers in recent years. The challenge of mapping between historical instances of placenames is also an ongoing concern for several important projects dealing with ancient placenames. Here the matching of historical placenames from two unrelated datasets to gazetteer web services is undertaken using a simple geospatial and geonomial algorithm. The quantitative results of the matching trials are considered, problems in dealing with vernacular scripts considered, and practical implications for integrating historical gazetteers discussed.

1. Existing Gazetteer Web Services and Digital Historical Gazetteers

The importance of online gazetteer services as authorities for geocoding placenames, or retrieval of placenames based on queries containing real world coordinates (reverse geocoding) has already been demonstrated. For example, the expanding interconnections of LinkedOpenData [1] on the semantic web have consistently placed GeoNames [2] at or near the center of the semantic web's social graph. The centrality of GeoNames is largely due to three factors: first, it is a free and open API; second, the API is simple and easy-too-use; and third, it is currently the only global geographic resource with stable URIs. Traffic for the GeoNames web service has topped 20 million requests per day, and since half of these are from smart phones, [3] it is clear that geographic information retrieval [GIR] is being built into many new location-based applications for hand-held devices.

Another major GIR web service is provided through the GoogleMaps Geocoding API,[4] which provides free geocoding and reverse geocoding web services. But the GoogleMaps web service, unlike GeoNames, provides no standard URI or unique identifier with their query results, which explains why there is no GoogleMaps presence on the OpenLinkedData cloud. Even so, the general explosion of webmaps and geocoding applications based on the GoogleMaps is clear to be seen. In 2010, Google declared that more than 350,000 websites were using the service, and that:

> "Google Maps API has established itself as the most popular Google API and the most deployed service-based API on the web." [5]

Clearly, the demand for accurate, automated GIR has become an essential part of the Internet experience for a rapidly growing audience.

The emergence of these robust gazetteer web services -- GeoNames, GoogleMaps Geocoding API, Yahoo Placemaker [6] -- provides an interesting testbed for GIR research. They have clearly outstripped their predecessor, the Alexandria Digital Library [ADL] gazetteer content standard and protocol, in terms of performance.[7] And while ADL established the basic principles of digital gazetteers, [8] the new breed of gazetteer web services simply appear as operational APIs, with technical documentation on query and response parameters but no theoretical underpinnings at all. Therefore it is quite interesting to see the ways in which GIR research is taking advantage of these gazetteer web services, by trying out new methodologies for integrating digital gazetteers, as well as exploring new theoretical aspects of GIR on the semantic web. [9]

One prospect for integration of digital gazetteers, is to augment the existing gazetteers with temporal attributes, turning them into spatio-temporal gazetteers, and enabling Geo-Temporal Information Retrieval [GTIR]. An obvious place to begin with this task, would be to establish links from the dated placename attestations in existing historical gazetteers to the undated placenames found in the authoritative gazetteer web services mentioned above. Examples of such a linking process have already been developed for the purpose of parsing placenames within historical texts, such as the pioneering Perseus Project,[10] and more recently the Google Ancient Places [GAP] project.[11] In both these cases, natural language parsing of placenames, tokenized within digital texts, is combined with geocoding the ancient placenames. For this purpose, Google Ancient Places can leverage placenames recorded by the Pleiades Project, which has a consistent means of accessing attestations about specific historical places in the ancient Euro-Mediterranean world. [12]

Despite these interesting achievements, there has yet to emerge a standard way to create linkages between historical instances of placenames and the current gazetteer authorities such as GeoNames. One way to approach this will be to "time-enable" current gazetteer services by establishing explicit links from them to the major digital historical gazetteers such as the Great Britain Historical GIS and its Administrative Unit Ontology,[13] the China Historical GIS gazetteer [CHGIS],[14] or the Regnum Francorum Online [RFO][15] collection of resources on the Merovingian and Carolingian Frankish kingdom.

The Regnum Francorum project, for example, has already implemented a number of cross-linked resources involving historical gazetteers. These include GeoNames, Wikipedia, the Princetone Encyclopedia of classical sites,[16] topographical dictionaries,[17] and links to numerous archaeological sites found in GoogleEarth and OpenStreetMap.[18] In this way, Regnum Francorum Online [RFO] provides a unique point of integration for disparate types of information, such as passages from historical texts, images of artifacts, important historical events, and a variety of spatial correlates, such as vector objects (points, lines and polygons) or high-resolution satellite views. The key point of association is the historical placename, therefore the connections that RFO has established between these resources is both an exemplar and practical demonstration of how the semantic web can be utilized for the compilation of historical gazetteers on a regional scale.

When working with printed resources such as the *Barrington Atlas of the Greek and Roman World*,[19] it is readily apparent that the features being depicted at a regional scale are not going to align perfectly with actual sites on the ground. Moreover, the digitization of the Barrington Atlas sheets [DARMC][20] required the help of several dozen graduate students over the course of more than two years to georeference scanned maps and to align vectorized map features with their positions on the contemporary GoogleEarth basemap. Nor can these GIS features be considered the last word on the subject, because new evidence or new interpretations about historical sites will inevitably follow.

The Pleiades Project, which can be seen as a "next generation" gazetteer, addresses this problem by incorporating multiple attestations about historical places. These attestations may be on a case-by-case basis, or may be the result of a batch integration, as was done with nearly 20,000 placename records derived from the DARMC project. However, neither DARMC nor Pleiades placenames contain references to the historical or present administrative hierarchy for their locations, a key factor in disambiguation of historical placenames. For integration of historical gazetteers to succeed, we must first augment them with meaningful attributes, such as relationship of named places to the administrative system, (both current and ancient,) and to provide attestations about their known dates of existence.

The major issues related to development of historical gazetteers were discussed during three days of meetings held at the American Association of Geographers [AAG] Annual Meeting in Seattle (2011),[21] out of which a proposal for a global historical gazetteer testbed emerged. In this paper, we will explore the prospects of historical gazetteer integration by implementing this testbed, and running placename matching algorithms between historical source gazetteers and online gazetteer web services. We will attempt to match records from CHGIS and RFO with GeoNames, and at the same time augment the records with results from the Google Maps API reverse geocoder, and then compare the results.

2. String Matching Methods

Before starting on the placename matching algorithm, we evaluated current research that makes use of web services (such as Yahoo, Google, and GeoNames) for placename integration. A variety of approaches were found, including the use of Soundex scores,[22] Levenshtein distance,[23] and adminstrative hierarchy checks.

In the case of Soundex, both source and target phonemes are given pronunciation codes, which can assist to find matches of similar spellings where an identical spelling match would otherwise fail. Unfortunately, there is currently no Soundex equivalent for working with Chinese characters, and for placenames based on spellings of romanized strings, the Soundex score cannot be reliably calculated, because Chinese words have both very high degree of ambiguity in phonemes per written character, as well as a variety of possible romanized spellings.

For example, Kuangtung (K52352) and Kwangtung (K52352), (respectively the Wade-Giles and a common variant) would match, and their Soundex scores are quite close to Guangdong (G52352), which is the official Pinyin romanization. However, the same does not hold true for the Pinyin spelling: Tianjin (T525), and it's Wade-Giles form: Tientsin (T5325), where the "*j*" and "*ts*" introduce a difference that is not based on the actual pronunciation of the Chinese word, but only on the spelling variations used by the two romanization systems. Of course, Soundex can only help with identicial or phonologically similar placenames, but never for alternate placenames, such as Canton (C535) or Cantão (C530). Based on these considerations, we have not implemented Soundex scores in our testbed matching algorithm.

The principle of Levenshtein distance, which calculates the minimum number of deletions or changes in a target string that would result in an exact match to a source string, is useful in theory for matching Pinyin spellings of Chinese words, though the Levenshtein distance is thrown off by the actual instances of Chinese placenames contained in our samples. For example, placenames in Chinese are generally considered to have two components: a *toponym* [zhuan ming], and a *classifier* [tong ming]. For example, Tengchong Xian [Tengchong County], is composed of the *toponym* = "Tengchong," and the *classifier* = "Xian."

One problem with using Levenshtein distance method is that the authority gazetteers (including GeoNames), may contain a placename "Tengchong" that lacks the *classifier* "Xian". Therefore, if we were to compare the two strings "Tengchong" and "Tengchong Xian" the Levenshtein distance will include the steps of deleting the blank space and the *classifier* "Xian," and **increase** the distance value; when, in fact, it is precisely the meaning of the classifier, "Xian," which might have semantically disambiguated that match from other candidates.

A recent study (by Gang Cheng, *et al*) specifically applying Levenshtein distance to Chinese placenames, takes this aspect into consideration, and claims to improve the completeness and accuracy of the placename matching algorithm, when compared to string match only.[24] Though we are intrigued by the methodology, their results were unconvincing. For example, the authors provide a table of similarity scores based on Levenshtein distance comparing "Puyang City" to "Puyang Country" (*sic*), in which:

> "special name" (*toponym*) similarity = 1
> "generic term" (c*lassifier*) similarity is 0.6561
> comprehensive similarity = 0.8968
> string match similarity = 0.6667 (ie, Levenshtein distance).

The overall conclusion being that a similarity of 0.8968 is much better than 0.6667.

However, if we were to simply match the first five letters of the two strings ("Puyan" = "Puyan"), the Levenshtein Distance would be a perfect 1. Also, as described below, it is possible to avoid complications of *toponym* & *classifier* bound forms in Chinese by limiting the string match to the minimum number of letters possible for two Chinese syllables. This also has the counter-intuitive advantage of finding more, rather than fewer, matches within longer address strings. To illustrate this concept, consider an attempt to match the strings "Henansheng Puyang Shi Puyang Xian Qinghetou Xiang" and "Puyang" using the method described by Gang Cheng, et al. Note that in the first string, there are four *classifiers*, some of them separated from their related toponyms with a blank space. The parsing method to determine toponym from classifier used by Gang *et al* would fail in the first string, since their inputs require a single *toponym* & *classifier* pair. On the other hand, if we simply stripped out all blank spaces, and matched the first five letters from each resulting placename string to the other, ***in both directions***, we would get a 0 match in one direction and a 1 match in the other:

HenanshengPuyangShiPuyangXianQinghetouXiang compared to *Puyang*

Henan != Puyang [0]
Puyan = Henansheng***Puyan***gShiPuyangXianQinghetouXiang [1]

The advantage of this method is that within any heterogenous list of placenames, regardless of completeness or incompleteness of *toponym* & *classifier*, and regardles of completeness in the hierarchical address, we are able to find positive matches. Of course, it's true that the match occurred at a higher level of aggregation in the specific string being searched *against,* and did not match exactly equivalent placenames. If this geonomial method is used by itself, all of the locations with "Puyang Shi" in their address would be found, which appear to be mis-matches. Ideally, we could use administrative jurisdiction information to complete the match, but this information is lacking in the first string, "Puyang," therefore we have no complete "address" to compare. However, if we add a geospatial buffer process, selecting for only those matches that occur within a certain threshold distance, then matchin relevance increases.

Owing to the special case of romanized Chinese, we stipulate that Levenshtein distance calculation is not appropriate for string matching process. Similarly, as our previous example demonstrates, the presence or lack of administrative hierarchy information, in the form of a complete "address" spelling, is not consistent enough in the samples for consistent parent / hierarchical matching. Therefore, our first objective is to develop a placename matching process to establish preliminary matches that will allow us to enhance the sample records by leveraging information from one gazetteer to another. Ultimately, the goal is to set up semi-automated processes, enabling temporal and ontological information to be cross-harvested between digital gazetteer sources.[25]

3. Strategy for Gazetteer Integration

Gazetteer **augmentation**, is in our view, by far the most practical approach to pursue for digital gazetteer integration. That is to say, an accumulation of information about particular historical places can be built up by harvesting factual elements from one or more gazetteer sources and then storing those elements, *or links to them*, in a target placename record. In this way, facts can be leveraged and cross-checked between gazetteer authorities.

This approach is very well formalized by Smart *et al*,[26] who have devised a Toponym Ontology as both a means of cross-checking and integrating placenames and related information about places from multiple sources. The toponym ontology relies on the use of a GeoFeature Augmenter, which enhances records from the original sources with related attributes from target gazetteers. For example, a placename record in one source, which has only a location and toponym, may query a second gazetteer to discover and store the current administrative district of that location. Or the query may consult another gazetteer to discover and store a feature classification.

Placename being queried: Tengchong Xian
Query to target gazetteer one: retrieves parent jurisdiction = Baoshan District, Yunnan Province
Query to target gazetteer two: retrieves feature classifications = County, Administrative Unit

In Smart, *et al*, the cumulative results comprise a Geofeature Set, which serves as a container object to bind the elements from different sources together. In this way, the GeoFeature Set can be considered as a means to define a "place" based on queries about toponyms or locations.

```
GeoFeature Set:
<name>Tengchong Xian</name>
  <class src=gaz2>County</class>
  <class src=gaz2>Adminstrative Unit</class>
  <parent src=gaz1>Baoshan District, Yunnan Province</parent>
```

Another advantage of the GeoFeature Augmenter concept, is that each gazetteer can be categorized for type of information that it is able to provide. In this way, incoming queries to the proposed system can be vetted with a Resource Selection Policy, and sent to the optimal source, depending on the type of related information being requested.

A concrete example of an augmented gazetteer are seen in Regnum Francorum Online [RFO] placename records, each of which contains official placenames, alternate placenames, administatrive districts (both historical and contemporary), as well as external identifiers for linking out to other web resources (GeoNames, Pleiades, GoogleMaps, OSM, Wikipedia, etc). An important finding of the partly automated augmentation process in RFO was that proximity filters (such as bounding boxes) when querying the Google Geocoding API for placenames resulted in quite ambiguous results, but when providing complete administrative information (ie. a complete "address" such as *Holzhausen, Niedersachsen, Germany* vs. *Holzhausen, Oetwil am See, Switzerland*), the system almost always produces an exact match. Therefore, from the outset we should attempt to leverage existing historical administrative jurisdiction information for inclusion in the basic augmentation process.

The augmented gazetteer idea was also central to Connecting Historical Authorities with Linked data, Contexts and Entities [the Chalice Project].[27] Chalice set out to use the Unlock Places API [28] to geocode gazetteers published by the English Place Names Society as well as other historical texts. Following Chalice, a new project, called Digital Exposure of England's Placenames [DEEP] [29] has now been launched with plans to complete the development of the historical placename gazetteer begun by Chalice. It should be noted that the DEEP project plan states that the conceptual model for their XML schema has yet to be finalized:

> "A major conceptual weakness identified by CHALICE was the lack a formal structure for the final form which the Survey-derived XML should take. It was concluded that simply reflecting the existing Survey content structure in XML did not exploit fully the flexibility and richness that working in the digital medium can bring to a complex historical/linguistic resource. This meant that the exemplar interface provided, while successfully demonstrating how linking the data worked and added some new functionality (such as being able to group placename forms by the century in which they are attested), was inevitably limited."[30]

Therefore, we should keep in mind that even the leading projects now underway in the field of historical gazetteers have yet to establish their own schemas, and the results are very much in a state of discovery and experimentation.

3. Gazetteer Integration Testbed

Our testbed aims to integrate placenames in an existing digital historical gazetteer with the unique IDs used by GeoNames. In this way, we can take advantage of the centrality of GeoNames in the LinkedOpenData web, and conduct future experiments to harvest linked data on the semantic web that is related to particular historical placenames *via* the common GeoNames ID. The overview of the testbed system is shown in Figure 1.

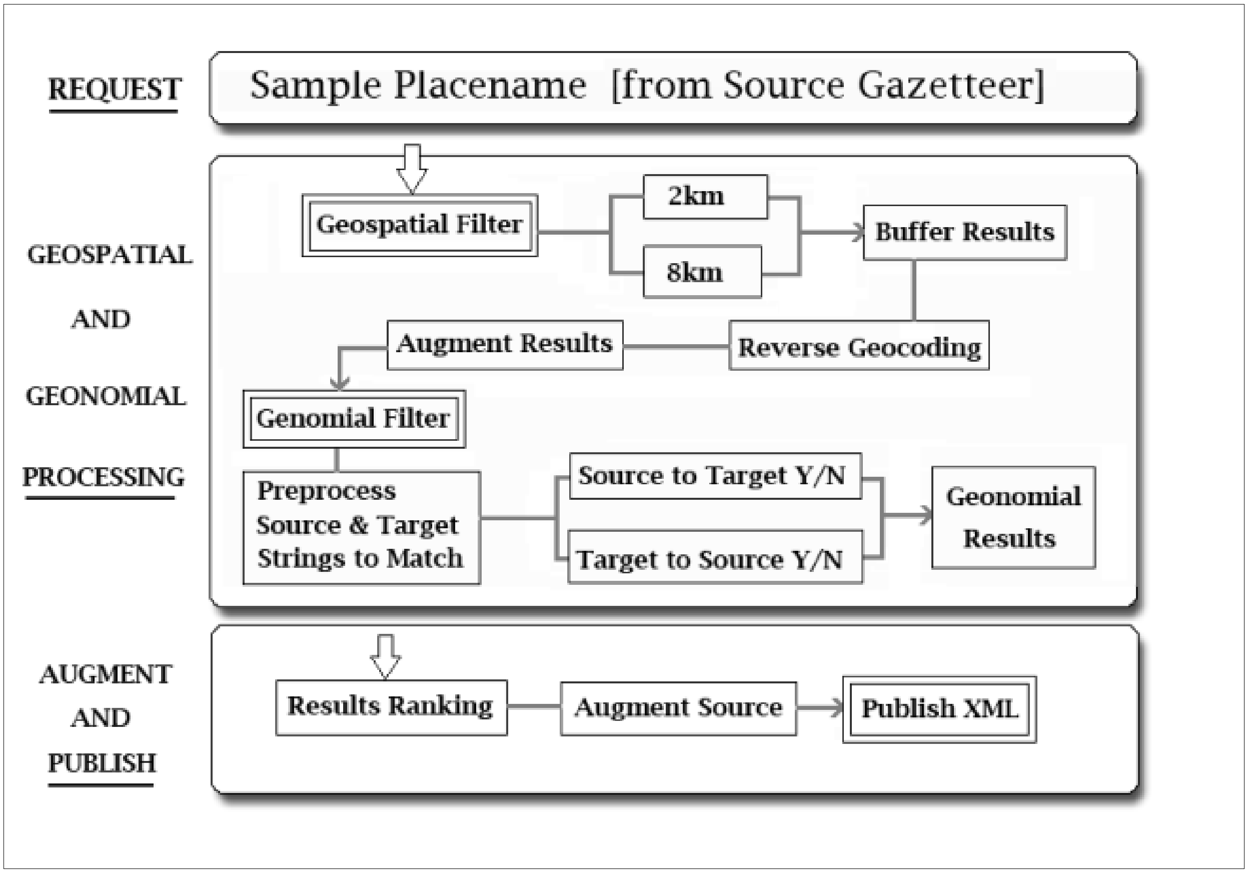

Figure 1: Overview of the gazetteer integration testbed system architecture

As a starting point, we developed an algorithm to iterate through a subset of China Historical GIS [CHGIS] placenames, (the historical county seats of Yunnan Province), and to augment them with GeoNames IDs by conducting geonomial and geospatial matching tests on each. To speed up the process, the matching tests were not processed via the GeoNames API, but were run using SQL queries upon a downloaded version of the GeoNames dataset for China and stored locally in a MySQL database. The results will be published in XML files conforming to a new historical placename schema, adapted from one currently in use by the CHGIS XML Web Service and expanded to include elements used in the RFO places schema. [31]

As a point of comparison, an identical test will be run using the same algorithm on historical placenames of France found in RFO database. The RFO placename records already contain GeoNames identifiers, hand-coded by the editor, and cross-checked with Google Geocoding API. Therefore, the matching of RFO names to GeoNames results will provide an interesting control dataset, to show similarities or differences in the hand-coded vs. auto-generated matches.

It is tempting to rely completely on GIS spatial analysis for the integration of digital historical gazetteers, especially when both source and target datasets contain mappable x, y coordinates. For example, it is trivial to produce areal buffers around the CHGIS historical county seats of Yunnan and then overlay the GeoNames points to see which GeoNames locations fall within those buffers. In this way, a number of buffer distances can be quickly tested to find the preferred radii to use for our basic distance filtering process. [See Figure 2]

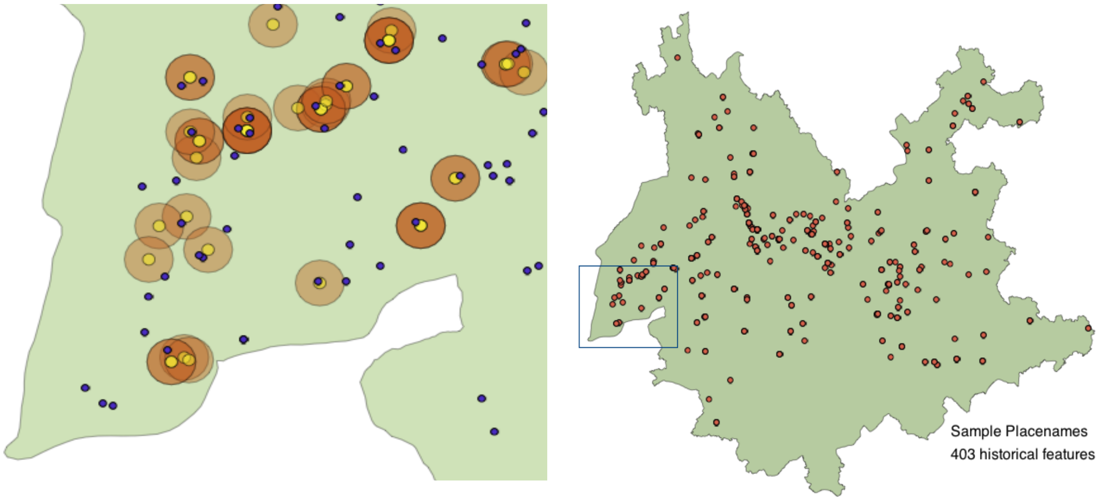

Figure 2: Yunnan Placenames Sample [left - showing zoom area and overlay of blue GeoNames points]

As a case in point, when testing the Yunnan study sample, we discovered that placename matches occur most frequently at break points within 2km and 8km, (a rather large planimetric error to account for.) Another interesting consequence of creating buffers from instances of historical placenames, is that, when symbolizing the buffers as semi-transparent, the darker hues indicate the points where more historical instances are recorded, an indicator of greater number of changes over time.

Since the source of many Chinese GeoNames records is the USGS GNS China files, their locational accuracy should be attributed to legacy errors inherited from the the source data, and not to subsequent processing by GeoNames. However, this reminds us that even though the spatial join in GIS is useful as a visual check, our main objective is to conduct string matching and distance checks as queries sent to web-based APIs, rather than to develop models to run in GIS applications.

Regarding the planimetic accuracy of points in GeoNames, it is important to know that regional variations are dramatically different. While point buffers for CHGIS points in Yunnan typically needed 2 to 8km to find reasonable matches, matches between the RFO points in France with their GeoNames counterparts never exceeded the 2km buffer, and often were found with Haversine calculated distance of less than 250 meters. [Figure 3]

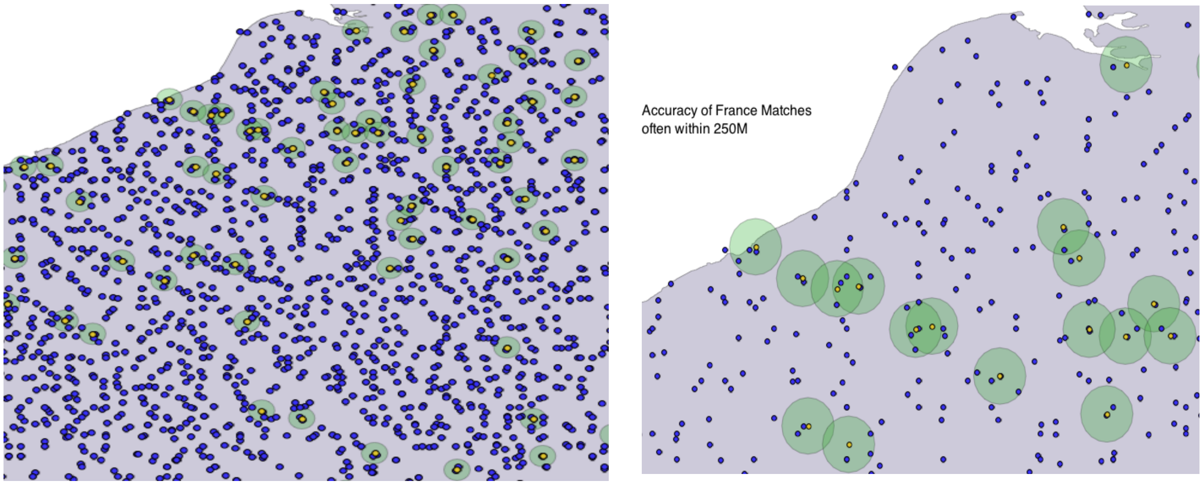

Figure 3: France Placenames Sample [right - showing overlay of blue GeoNames points within 2km zones]

4. Geospatial Matching

Although geonomial (string matching) of placenames was the primary task for our algorithm, it is much more efficient to filter by Haversine distance before conducting the geonomial function. Working from the findings of the preliminary GIS buffering, radii of 2km and 8km were calculated for the Yunnan sample, and 2km radii for the France sample. Note that Smart, *et al,* worked with a 500m radius, which would have been slightly too constraining to match all of the the France data. In the following chart, the total number of source names is shown, and the total number of matches made purely on calculated Haversine distance from source point to any possible GeoNames point, within 2km or 8km buffer zones. The Average distance within the zones is also shown, along with the total average distance of all points found. [Figure 4]

Sample	Total Names	2km Hits	Avg < 2km	8km Hits	Avg <8km	Avg All
Yunnan	403	297	1.15 km	1994	4.67 km	6.07 km
France	4164	17705	0.84 km	n.a.	n. a.	1.02 km

Figure 4: Number of matches found within 2km and 8km buffer zones, with average distances

For CHGIS Yunnan points, the 2km radius was clearly not enough to capture the majority of matches, while the density of GeoNames hits with 2km of RFO data was sufficiently large to capture all the candidates.

Familiarity with the sample data led us to believe that many of the Chinese historical placenames would not successfully match geonames placename spellings, so we implemented two other geospatial matching functions in the test algorithm. The first was a reverse geocoding request to the Google Geocoding API, based on the source placename's geographic coordinates. [32] The second was to geocode the locality name returned in the reverse geocoding response. For example, when reverse geocoding the CHGIS location for Yongping Xian, the top return (ie. the closest location found in Google API) was a "rooftop" located at Number 1 Bonan East Road, Yongping, Dali, 672600. Since this "address" found is not optimal for our placename string matching checks, we first parse the "locality" element from the the first response (which, in this case = "Yongping"), then do a straight placename geocoding request for the locality name.[33] In this way, regardless of the placenames stored in the CHGIS gazetteer, we have established a simple cross-check to find the nearest Google API locality (*ie* "Yongping" rather than "Number 1 Bonan East Road, Yongping, Dali, 672600").

5. Geonomial Matching

The geonomial matching algorithm, which checks for toponym string matches, is designed to take the simplest path through the obstacles to Chinese placename parsing mentioned previously. The CHGIS placenames, consist of the historical placenames (in UTF8 encoded Chinese characters, and romanized Pinyin), as well as a Present Location description (in UTF8 encoded Chinese and an auto-generated Pinyin transliteration). By contrast, the GeoNames placenames consist of a default toponym, and Alternate Names (consisting of a comma-delimited array of mixed vernacular scripts using UTF8 encoding). The difficulty in matching the CHGIS to GeoNames placenames is primarily due to the fact that typically only one part of the Present Location address string (of the CHGIS record) will match any part of the GeoNames record.

As an expedient to enable matching in either direction, we used a preprocess to strip all blank spaces from the source placenames, and then trim each sample string down to the first five characters. In this way we have a "source" string consisting of only five roman letters that we can check against the complete addrees string in the target gazetteer.

The reason for limiting the string to only five letters is based on the problem related to bound forms of Chinese placenames, mentioned above. The majority of Chinese syllables in romanized spellings rarely contain less than three letters. For example, the Pinyin syllable "Ou" may sometimes be spelled "O" in Wade-Giles, and very rarely are one letter syllables in toponyms possible, such as "E," as in "Emei Shan." In addition, there are no syllables at all containing more than six romanized letters, ie "Zhuang" or "Shuang." The majority of syllables fall into groups spelled with two to four letters, meaning that six letters should be enough to capture the majority of possible Chinese placenames that might consist of one *toponym* syllable and one *classifier* syllable.

An example might be: "Ai Xian." There might be cases of a single letter syllable *toponym* and a three letter *classifier*, something like "O Cun," that would be a rare occurrence indeed. [34] It would not be at all uncommon, on the other hand, to encounter romanized *toponym* / *classifier* combinations such as: "Yi cun," "Ao cun," "An Xian," etc. Even more typical for Chinese placenames would be a two syllable *toponym* and a one syllable *classifier*, on the order of: "Mengba Xian," "Fujian Sheng," or "Shuangcheng Xian." Based on these possible syllable length combinations, the minimum number of letters for an intellible match was deemed to be five letters. This will be a practical method for handling Chinese placenames until a more reliable means of comparing romanized Chinese placenames with their vernacular Chinese characters can be developed. [35]

Stripping out all blank spaces was necessary because an auto-generated romanization was used for the Present Location address in the CHGIS source data, and those values contained no spaces between syllables. By stripping out all blank spaces in both source and target strings, even the extreme case of short syllable spellings such as "aicun" resulted in postive matches. Similarly, a long name such as "Shuangbai Xian" stripped to "shuan" resulted in positive matches.

6. Quantitative Study of Matching Historical Placenames to Contemporary Gazetteers

When combined with initial filtering based on the haversine distance, the overall successful matching rates for the algorithm can be summarized in the following table. [Figure 5]

Sample	Total Names	One-way match	Two-way match	Source to Target	Target to Source
France	4164	97%	86%	97%	192%
Yunnan	403	74%	18%	27%	87%

Figure 5: Perecentage of matches found with algorithm

As we can see from the France source, even though we know in advance that each name has a corresponding GeoNames ID, the actual match between RFO placename spellings and GeoNames spellings in both directions was only 86%. If we check for only one match, in either direction, we find a 97% match rate. Because matching from source to target is one-to-many relationship, the percentage of matches from target back to source can be higher than 100%; for the France data this figure was 192%. The multiple matches from GeoNames back to RFO names is mostly an indication of ambiguity; more than one matching RFO placename string (within 2km) was found in GeoNames, which is not surprising. Although we have not examined the complete set of placenames that failed to match, a small sample were retrieved (based on the known GeoNames ID in the RFO gazetteer), which revealed 80% of the inability to correctly match was due to inconsistent character set encodings of accent marks, and another 20% due to actual difference in placenames or variant spellings.

The Yunnan matching results tell a somewhat different story. Not only were there only 74% one-way matches, but the two way match was dramatically lower, at 18%. This indicates that there were **very few** cases in which the historical placename spellings in the source gazetteer were direct matches with the target gazetteer placenames. In fact, for the Yunnan data, we find that source to target matches were only 27%, meaning the occurrence of historical placenames in the the GeoNames Alternative Placenames field were relatively low. By contrast, 87% of the one way matches were found when matching GeoNames to the Present Location placename string in CHGIS. This indicates that the match was three times **more** likely to occur when checking in the direction of FROM the contemporary gazetteer TO the historical gazetteer record (as long as the historical gazetteer contains an attestation of the present location), vs. checking FROM the historical gazetteer TO the "alternate" names of the contemporary gazetteer.

7. Augmented Gazetteers: the Convergence of Historical and Modern Gazetteers

Having examined the preliminary matching results, we can now harvest the actual placenames into an augmented gazetteer for public consumption. As mentioned above, there has yet to emerge a "standard" metadata schema for exchange of historical placename data. The Alexandria Digital Library Gazetteer Content Standard remains, in our view, the best-organized historical placename standard. The simplified version implemented in RelaxRNG format for the CHGIS XML Web Service, however, was ready at hand, so we have opted to exend it slightly to allow for multiple sources, for explicit language and encoding elements, and to propose a more generic method of linking to other resources via URIs.

Here is a snippet from the existing CHGIS XML response for the query:
http://chgis.hmdc.harvard.edu/xml/placename/腾越

```
<item id="9750">
  <placename>
    <name_romanized>Tengyue Ting</name_romanized>
    <name_vernacular>腾越厅</name_vernacular>
    <name_alternate>騰越廳</name_alternate>
  </placename>
  <feature_type>
    <type_english>independent sub-prefecture</type_english>
    <type_romanized>Zhiliting</type_romanized>
    <type_vernacular>直隶厅</type_vernacular>
    <type_id>888</type_id>
  </feature_type>
 </item>
```

Our proposed augmented gazetteer, will contain repeatable elements for the linked data found by matching the GeoNames gazetteer, for example:

```
<item id="9750">
  <placename>
    <name_romanized  lang=zh form=py charset=ansi>Tengyue Ting</name_romanized>
    <name_vernacular  lang=zh form=simp charset=utf8>腾越厅</name_vernacular>
    <name_alternate  lang=zh form=trad charset=utf8>騰越廳</name_alternate>
  </placename>
  <linkedData>
          <link>
             <linkName lang=zh form=py charset=ansi>Tengyue</linkName>
             <linkType>gazetteer</linkType>
              <linkSrc>GeoNames</linkedSrc>
             <linkLat>24.99492</linkLat>
             <linkLong>98.51276</linkLong>
```

```
           <linkAddress>Tengyue, Yunnan, China</linkAddress>
           <linkID>1279891</linkID>
           <linkDate>20110520</linkDate>
           <linkURI>http://www.geonames.org/1279891/</linkURI>
          </link>
          <link>
            <linkName lang=zh form=py charset=ansi>Tengchong</linkName>
            <linkType>gazetteer</linkType>
            <linkSrc>GoogleMapsAPI</linkedSrc>
            <linkLat>25.020617</linkLat>
            <linkLong>98.490002</linkLong>
            <linkAddress>Tengchong, Baoshan, Yunnan, China</linkAddress>
            <linkID>Fengshan Rd, Tengchong, Baoshan, Yunnan, China</linkID>
            <linkDate>20110902</linkDate>
           <linkURI>http://maps.googleapis.com/maps/api/geocode/json?
latlng=25.020617,98.490002&sensor=true</linkURI>
          </link>
          <link>
            <linkName lang=zh form=py charset=ansi>云南省图</linkName>
            <linkType>map</linkType>
            <linkSrc>Maptown CN</linkedSrc>
            <linkLat>25</linkLat>
            <linkLong>17.7</linkLong>
            <linkID>TwoEightFourFourZero</linkID>
            <linkDate>1925</linkDate>
           <linkURI>http://www.mapd.cn/Map/YunNan/DetailTwoEightFourFourZero.html</linkURI>
          </link>
  </linkedData>
  </item>
```

In this case, the original CHGIS record has been augmented with the match found in GeoNames, as well as the Google Maps API reverse geocoding "locality" name, Tengchong. Each placename attestation is defined with a language, format, and character set encoding, as well as a source, identifier, and (hopefully) long-lasting URI. Where available the administrative hierarchy is captured in the complete "address."

Before finalizing the XML schema for publication, we plan to experiment by augmenting the gazetteer entries with links to other potentially useful cross-references, such as DBpedia pages, WOEIDs,[36] etc.

In conclusion, we find that the use of "alternate names" in the current gazetteer authorities are not well-enough defined to enable consistent matching of historical placenames. On the other hand, there is mugh higher degree of reliability in matching from the selection of nearby gazetteer authority placenames obtained from reverse geocoding, when an attestation of "present location" is provided within historical gazetteer placename records. Based on these findings, our recommendation is augment the historical gazetteers with explicit links to their contemporary locations using permanent identifiers (such as GeoNames IDs), and persistent URIs. In this way, the augmented historical gazetteers functionality is extended to provide programmatic methods for retrieval of both historical and contemporary location-based information. As a means of bootstrapping historical placnames into the realm of LinkedOpenData, the potential use cases for historical gazetteer resources will be expanded, and also make it possible for practical tests of the next wave of gazetteer research: Geo-Temporal Information Retrieval.

References

1. Cyganiak, Richard. *The Linking Open Data Cloud Diagram*. Sep 2011. online: http://richard.cyganiak.de/2007/10/lod/

2. *GeoNames*. online: http://www.geonames.org/export/web-services.html

3. Wick, Marc. *Application Identification for Free GeoNames Web Services*. Jan 2011. online: http://bit.ly/hWX7Ef

4. *GoogleMaps Geocoding API*. online: http://code.google.com/apis/maps/documentation/geocoding/

5. Google Geo Developers Blog. *Google Maps API turns 5!* online: http://bit.ly/9yRPMi

6. *Yahoo Placemaker*. online: http://developer.yahoo.com/geo/placemaker/

7. Alexandria Digital Library Project. *Guide to the ADL Gazetteer Content Standard*. Feb 2004. online: http://bit.ly/wzTdSe

8. Hill, Linda. *Georeferencing, the Geographic Assocations of Information*. 2006.

9. See, for example, the *Pelagios Project*. online: http://pelagios-project.blogspot.com/
See also "geographic" topics in the *Sematic Web Journal*, edited by Pascal Hitzler and Krzysztof Janowicz. online: http://www.semantic-web-journal.net/search/node/geographic

10. Smith, David A. and Gregory Crane. "Disambiguating Geographic Names in a Historical Digital Library," in *Proceedings of European Conference on Digital Libraries*, 2001: 127-136. online: http://www.perseus.tufts.edu/~ababeu/geodl01.pdf

11. Rabinowitz, Nick. "Designing a Visual Interface for Google Ancient Places Texts," *GoogleAncientPlaces Blog*, Nov 2011. online: http://googleancientplaces.wordpress.com/

12. Gillies, Sean. "Does Pleiades Have an API?" *Sean Gillies Blog*, Dec 2011. online: http://sgillies.net/blog/1101/does-pleiades-have-an-api/

13. Aucott, Paula, Kupca, V., Lagrelius, J., Von Luenen, Alexander, Palm, F. and Southall, Humphrey. *QVIZ: report and schema for the administrative unit ontology*. 2008. online: http://eprints.port.ac.uk/5083/1/QVIZ_Admin_Unit_Ontology_Report.pdf

14. *China Historical GIS* [CHGIS]. online: http://www.fas.harvard.edu/~chgis

15. Åhlfeldt, Johan. *Regnum Francorum Online : Interactive Maps and Sources of Late Antique and Early Medieval Europe*. online: http://www.francia.ahlfeldt.se/
see also: http://static.ahlfeldt.se/pelagios_rfo.pdf

16. Stillwell, Richard, William L. MacDonald, Marian Holland McAllister, Stillwell, Richard, MacDonald, William L., McAlister, Marian Holland, Ed. *The Princeton Encyclopedia of Classical Sites* . Princeton University Press, 1976. Digitized by the Perseus Project with funding from the National Endowment for the Humanities. online: http://bit.ly/y1bDcg

17. Gysseling, Maurits. *Toponymisch woordenboek von België, Nederland, Luxemburg, Frankrijk en West-Duitsland (vóór 1226)* Brussels, 1960. [A scholarly reference which includes source evidence, placenames (both modern and historical) and administrative jurisdiction.] online: http://www.wulfila.be/tw/

See also: *Dictionnaire topographique de la France*, 33 vols (1861-2008) online: http://cths.fr/topo/accueil.php

18. Åhlfeldt, Johan. *Compiling Sources of Ancient History*. Feb, 2012. online: http://early-medieval-gis.blogspot.com/2012/02/compiling-sources-of-ancient-history.html

19. Talbert, Richard J. A. *Barrington Atlas of the Greek and Roman World*. Princeton University Press, 2000. online: http://www.unc.edu/awmc/batlas.html

20. McCormick, Michael, Guoping Huang, Kelly Gibson, et al. *Digital Atlas of Roman and Medieval Civilization [DARMC]*. online:

21. Symposium on Space-Time Integration in Geography and GIScience, Special Track on Historical Gazetteers. *AAG Annual Meeting 2012*, Seattle, Washington. online: http://www.fas.harvard.edu/~chgis/gazetteer/

22. *Soundex*. [phonetic algorithm for indexing names by sound, as pronounced in English.] online: http://en.wikipedia.org/wiki/Soundex

23. *Levenshtein Distance*. [a caculation to determine the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character.] online: http://en.wikipedia.org/wiki/Levenshtein_distance

24. Cheng, Gang, Fei Wang, Haiyang Lv, and Yinling Zhang. "A New Matching Algorithm for Chinese Place Names." in *Geomatics 2011 Proceedings*, June 2011.

25. Snae, Chakrit. "A comparison and Analysis of Name Matching Algorithms," in *World Academy of Science Engineering and Technology*, v25, 2007: 252-257. online: http://www.waset.org/journals/waset/v25/v25-47.pdf

26. Smart, Philip D., Christopher Jones, and Florian Twaroch. "Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service," in *Lecture Notes in Computer Science*, 2010, Volume 6292/2010, 234-248. online: http://www.springerlink.com/content/4225546vp3123801/

27. *Connecting Historical Authorities with Linked data, Contexts and Entities* [the Chalice Project] 2011. online: http://chalice.blogs.edina.ac.uk/

28. *Unlock Places Project*. 2010. online: http://unlock.edina.ac.uk/places/introduction

29. *Digital Exposure of England's Placenames* [DEEP]. 2011-2013. online: http://englishplacenames.cerch.kcl.ac.uk/

30. *Digital Exposure of England's Placenames [DEEP], JISC Project Plan*, 2011: Appendix B: Data Model Development. online: http://www.jisc.ac.uk/media/documents/programmes/digitisation/econtent/econtent11_13/englishplacenamesprojectplan.pdf

31. Berman, Merrick Lex. *China Historical GIS XML Web Service*. 2008. online: http://chgis.hmdc.harvard.edu/xml/

32. Google Maps API Version 3, example Reverse Geocoding request:
http://maps.googleapis.com/maps/api/geocode/json?latlng=25.464684,99.54124&types=locality&sensor=true

33. Google Maps API Version 3, example Geocoding request:
http://maps.googleapis.com/maps/api/geocode/json?address=永平县+云南&sensor=true

34. Tsai, Chih-Hao. *Zhuyin, Hanyu Pinyin, and Tongyong Pinyin Cross-Reference Table*. 2000. online: http://research.chtsai.org/papers/pinyin-xref.html

35. Kwok, Kui Lam, and Qiang Deng. "GeoName: a system for back-transliterating pinyin place names," in *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 2003. online: http://acl.ldc.upenn.edu/W/W03/W03-0104.pdf

36. *Where On Earth Identifiers* [WOEID]. online: http://developer.yahoo.com/geo/geoplanet/guide/concepts.html