Georeferencing Historical Placenames and Tracking Changes Over Time

Merrick Lex Berman
Project Manager, China Historical GIS
Harvard Center for Geographic Analysis

Presented at the Georeferencing Workshop, Harvard University

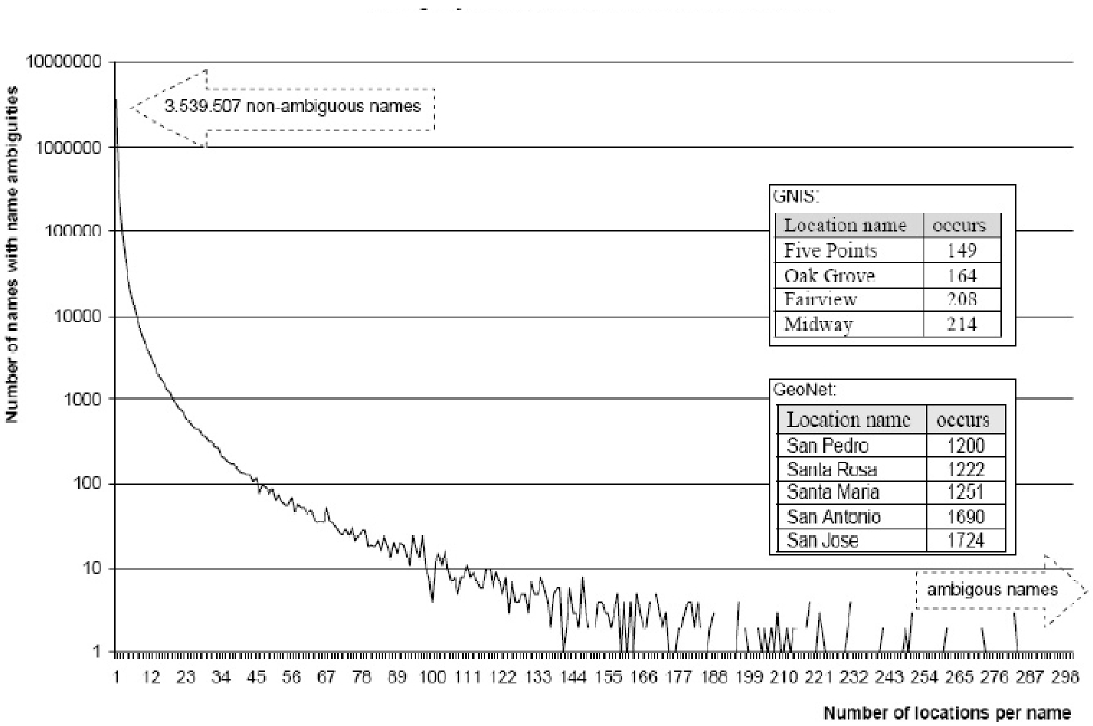
Mar 2008

Abstract:

Working with a specific gazetteer of Chinese historical placenames, the rates of ambiguous vs. non-ambiguous placenames are quantified in relation to larger corpus of spellings derived from the GNS and GNIS datasets. The case is made for time as a salient factor for disambiguating placename strings as compared to feature types or administrative jurisdiction. On the one hand, handling time values is practical because generic date handling libraries already exist. On the other hand, trying to match spellings to feature types and administrative divisions is highly dependent on language use and other variables which have neither universally accepted controlled vocabularies nor, for that matter, any semantic constraints. When the problem of semantic interoperability is compounded across languages, political systems, and historical eras, the use of time element in placename disambiguation seems all the more practical. This suggests that adding valid time spans to existing gazetteer databases would greatly enhance their usefulness for georeferencing.

1.1 Measuring percentage of ambiguous placenames

Recent work on disambiguation of placenames using an ontology-based scheme has been proven dramatically effective, raising the overall percentage of non-ambiguous matches against a large corpus of placenames from 59% to over 90%. **[Volz *et al*, 2007]** To accomplish this an ontology framework was built from feature types and parent countries, and then used to narrow the range of candidates for matching against any given placename string. An interesting graph showing the ratio of name occurrences to the number of cases in which the same name was shared is shown in Figure 1.



Location name	occurs
Five Points	149
Oak Grove	164
Fairview	208
Midway	214

Location name	occurs
San Pedro	1200
Santa Rosa	1222
Santa Maria	1251
San Antonio	1690
San Jose	1724

Fig 1: Ambiguity of Geographic Names in GNS and GNIS [from Volz, et al 2007]

This graph sparked my curiosity to measure the percentage of non-ambiguous placenames in the China Historical GIS dataset **[CHGIS]** based on spellings.

It should be noted that in the graph shown in Figure 1, each location has only one unique placename spelling. That is to say, when the Location Name “San Jose” occurs 1,724 times in the GeoNet database, those 1,724 cases represent unique locations around the world.

However, in the case of CHGIS data, we must account for multiple identical spellings related to any particular pair of x, y coordinates. This is necessary because in the CHGIS dataset, historical instances are created not only for each change in placename, but also for changes in feature type or location. **[Berman, 2003]** This results in a time series of historical instances, for example:

instance 1:	St. Petersburg	(Russia)	Begin=1703	End=1913
instance 2:	Petrograd	(Russia)	Begin=1914	End=1923
instance 3:	Leningrad	(Soviet Union)	Begin=1924	End=1990
instance 4:	St.Petersburg	(Russian Fed)	Begin=1991	

Just because the placename spelling is identical in the first and fourth instances does not mean that they are the same place historically. Tsarist Russia of 1913 is simply not the same as the Russian Federation of 1991. And more importantly, if we are trying to match placenames against historical gazetteer records, we really don’t want to dissolve instances based on spelling alone. We want to preserve all the possible matches, then filter them based on a time value.

Let us take a first pass over the placename spellings stored in the CHGIS search engine, which is comprised of multiple gazetteer sources (the China Names downloaded from GNS **[GNS]** , CHGIS placenames, China in Time And Space placenames **[CITAS]**, and Russia Academy of Science placenames extracted from historic maps **[RAS]**). The ratio of name occurrences to number of places sharing the same name is shown in Figure 2.

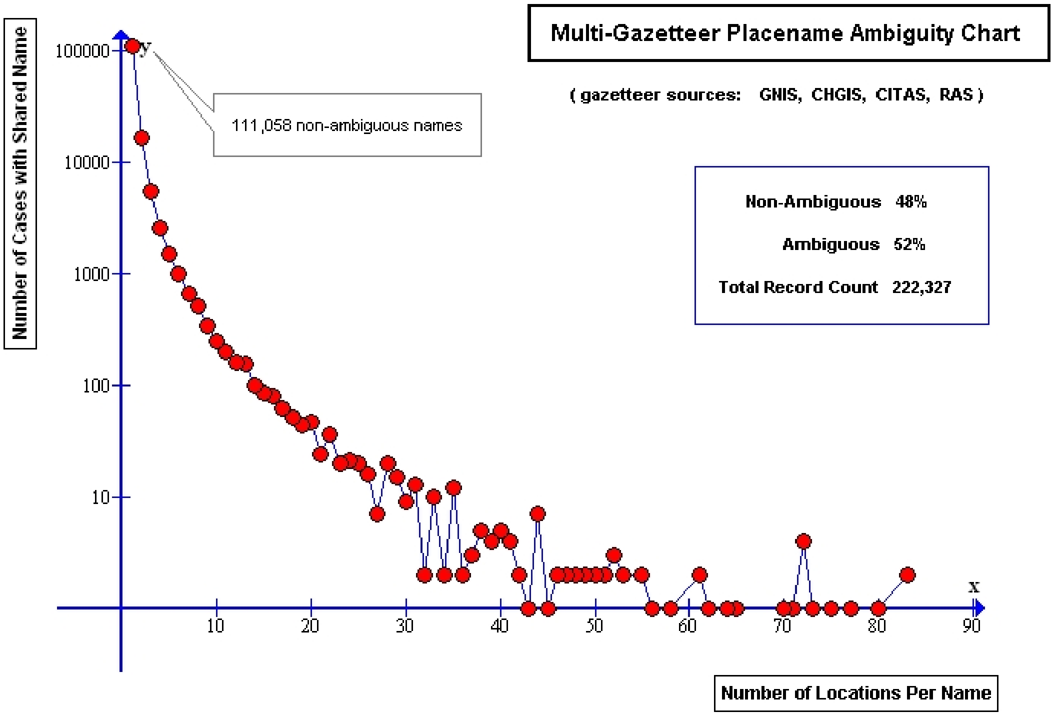


Fig 2: Ambiguity of Placenames in CHGIS Search Engine – Inclusive of all Records

We could run similar queries on subsets of the CHGIS search engine, to show all of the CHGIS historical records (including Time Series and Time Slices for 1820 and 1911), or limiting the query to show only Time Series records. Time Series records are those which represent historical instances as they change over time, and therefore have begin and end dates that are not the same value. Time Slice records are those which represent a single snapshot in time, and have the same begin and end date values. The results can be plotted on a single graph, as shown in Figure 3.

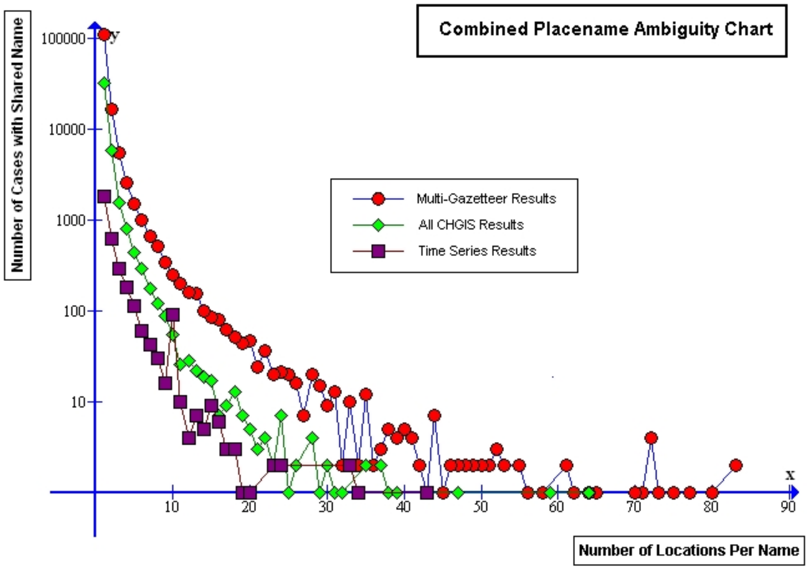


Fig 3: Ambiguity of Placenames in Datasets - All Records, All CHGIS Records, Time Series Only

When looking at this graph, I began to question it's purpose. After all, it is not very mysterious for there to be a high number of non-ambiguous records, then a smaller number where the same name was shared by a total of two records, and yet a smaller number shared by three records, and so on. The phenomenon of distribution, in this case, is not very instructive. Therefore, I decided to look at the raw percentages of non-ambiguous placenames instead, and to compare them with the results of the ontology-based scheme from Volz, et al. The comparison is shown in Figure 4.

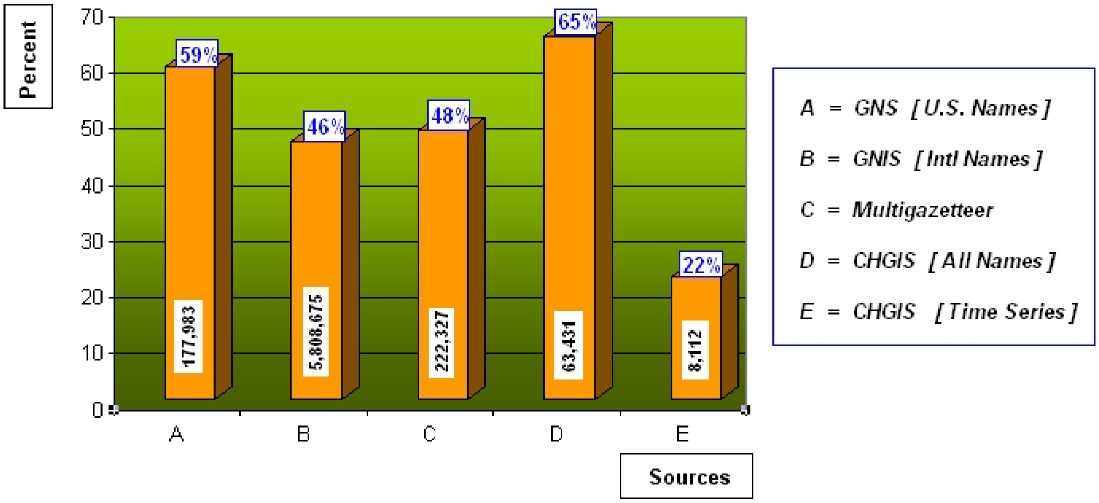


Fig 4 : Percentages of Non-Ambiguous Placenames Overall

As you can see, the non-ambiguous names for the GNIS International Names dataset (46%) tracks quite closely with the CHGIS Multigazetteer dataset (48%), since the largest component of the latter is drawn from the former. Note the increase in non-ambiguous names for the CHGIS datasets overall (65%), which I believe is explained by the ability to group by Chinese Character placenames, rather than romanized placenames. Also, please note the dramatic reduction in non-ambiguous placenames when testing the CHGIS Time Series in isolation (22%). This is due to the fact that the Time Series includes many historical instances of the same spelling, as mentioned earlier in the case of St. Petersburg. The next question, then, is what occurs if we dissolve instances with identical spellings in the Time Series data when they occur at the same location but at different times?

After dissolving multiple occurrences of identical spellings based on location, we end up with the same number of point locations to be disambiguated spatially, but far fewer instances of spellings. Indeed, when identical spellings are collapsed into one **spelling-plus-location** combination, the percentage of non-ambiguous names is dramatically increased, as shown in Figure 5, columns E and F.

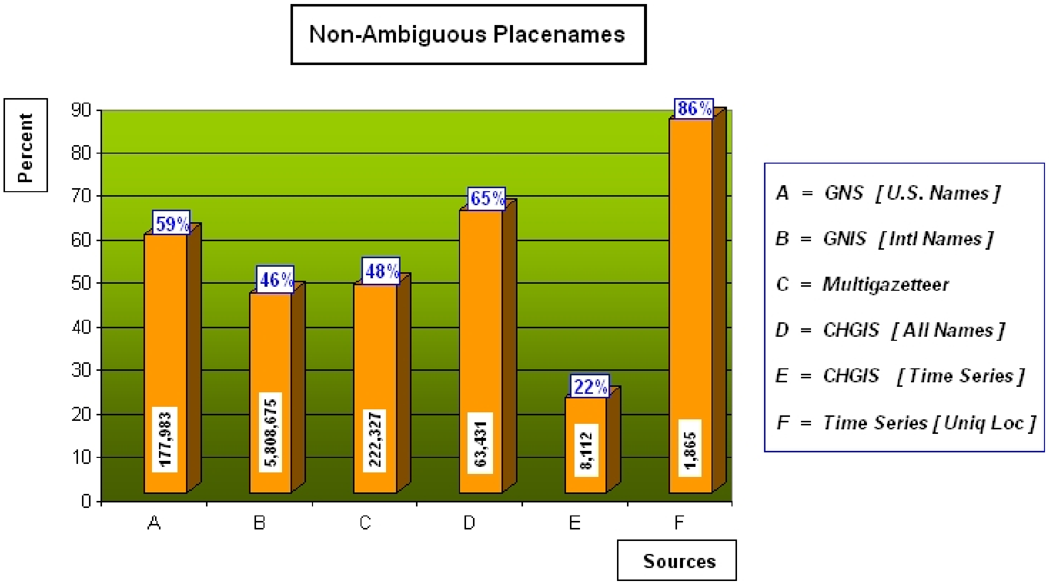


Fig 5: Percentages of Non-Ambiguous Placenames Overall

This suggests that the addition of dates to the core elements of our reference gazetteers and, at the same time, combining dates with each toponym that is being queried in the gazetteer will have a significant effect on disambiguating the results.

1.2 Evaluating Historical Gazetteers as Georeferencing Resources

With the exception of several national historical GIS projects, such as GBHGIS **[GBHGIS]**, CHGIS **[CHGIS]**, and NHGIS **[NHGIS]**, the majority of available digital gazetteer sources contain no date of validity information for placename records. This includes the largest available datasets: GNS **[GNS]** and GeoNames **[GeoNames]**. At best, the user knows the date of publication, but nothing about the dates of occurrence for each placename. This gap was partially addressed by the inclusion of temporal status as a core element in the Alexandria Digital Library Gazetteer Standard. **[ADL , 2004]** However, the only mandatory attribute for the temporal status is to indicate whether the historical instance is *current, former, or proposed*. Actual dates, which could fix gazetteer records into a searchable timeline, are optional. Furthermore, very few cases of implementation of the ADL standard for historical gazetteers exist. Which leaves us with very little to work with, or to test programmatically, in order to study the georeferencing of historical placenames.

An entirely different approach, and a promising one, is to create collaborative tools which automate the creation of standardized metadata and allow multiple users to conduct research on historical places, people, and events. This technique has been developed at HeuristScholar.org, **[Heurist]** where both individual researchers and workgroups can populate a database with specific placenames, spatial objects, historical events, dates of occurrence, and bibliographic references, and can also map relationships among any of the records that have already been entered. The result is an extensible framework of spatio-temporal information, allowing for fuzziness and for multiple interpretations of all the objects in the database and their relationships. **[Mostern, Johnson]** Of course, the openness of Heurist is also problematic, since it allows for the markup of any historical instances (*and their relationships!*) without semantic constraints on how they are being described or defined. Though the interlocking contributions of the participants can develop into a fascinating conglomeration of data about historical places and events, nonetheless there remains a serious risk that navigating and understanding the complexity of the relationships will neither be easy to interpret, nor suitable for the task of georeferencing. The Heurist project itself is a process of georeferencing historical data, but is not yet a tool for georeferencing of that data.

A similar collaborative research project takes a narrower approach and focuses primarily on ancient geography. **[Pleiades]** Arising from the Barrington Atlas of the Greek and Roman World **[Barrington, 2000]** and the Ancient World Mapping Center, **[AWMC]** the Pleiades Project provides a platform for the examination of Ancient Names and Ancient Places (where a “place” is equal to a “name” plus an attested location). The satisfying aspect of the Pleiades approach is that places can be given temporal attestations, which are divided into neat groups, and that all attestations are documented. The unsatisfying aspect of Pleiades, so far, is that both place and temporal attestations are optional and that the only downloadable version of the dataset (in KML format) presents all the temporal information as formatted text rather than dates. Of course, we can easily map the time periods used in Pleiades to dates (*eg* Archaic = Pre-550 BC, Classical 550-330 BC), which will allow us to test the georeferencing of historical placenames. Unfortunately, the existing downloadable dataset from Pleiades is less than 500 records, which is not adequate for an experiment yet.

Perhaps the largest corpus of geographically annotated digital texts is the Perseus Project Digital Library of Classical Literature. **[Perseus]** Perseus paved the way for the disambiguation of toponyms in the academic world, much as Metacarta has done in the private sector. **[Metacarta]** Perseus uses a complex process based on natural language parsing and then prunes candidate matches based on a series of spatial standard deviations away from clusters of related points. **[Smith *et al*, 2001]** The results of the Perseus toponym disambiguation process are especially impressive, since the percentage of non-ambiguous placenames in the Perseus digital source texts is only 8%, and after processing non-ambiguous matches are increased to approximately 90%. It is interesting to contemplate what can be accomplished if we combine temporal information as part of the disambiguation process with the Perseus toponym disambiguation algorithms. By point of comparison, see Figure 6.

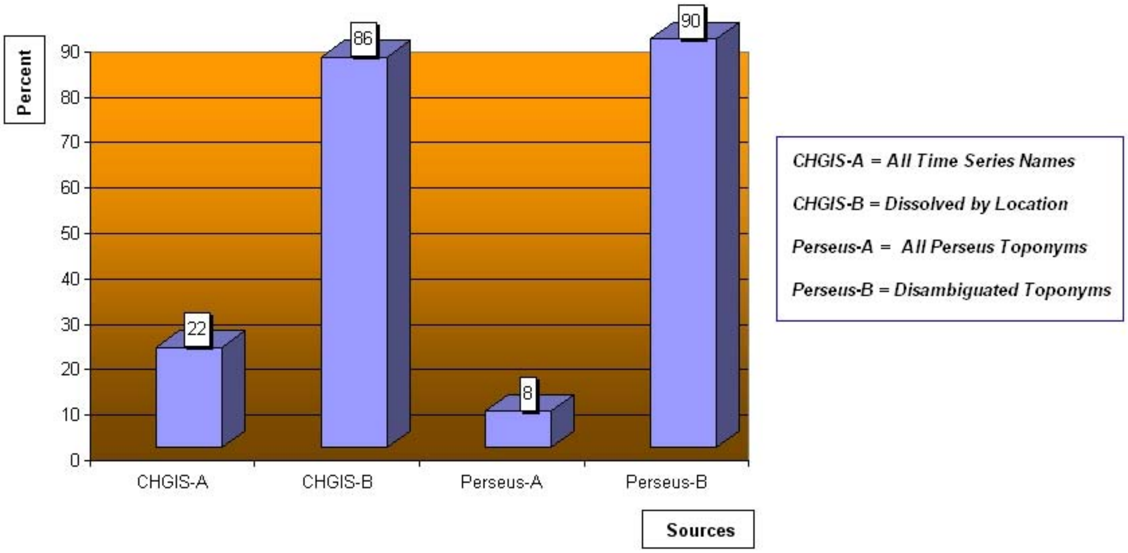


Fig 6 : Gain in Non-Ambiguous Placenames

It is clear that the addition of dates to the disambiguation process will boost the percentage of non-ambiguous names, but until we have a core gazetteer to test which contains date information we will not be able to measure the performance increase.

The latest incarnation of Perseus Digital Library **[Perseus Version 4]** includes the ability to download entire sections of the digital library as XML formatted text, and also provides an API for extracting specific well-formed XML fragments based on query parameters. However, there has yet to be made available a gazetteer derived from the Perseus projects digital library and disambiguation algorithms, which seems to be a serious gap. Clearly the leader in the field of disambiguating historical toponyms, it would be extremely valuable for others to obtain the latest snapshot of named geographic entities derived from the Perseus Digital Library. Until then, we struggle on in darkness!

1.3 <u>Chronologies as the Backbone of Global Gazetteers and Global Historical GIS</u>

But there may yet be a way through the jungle of geographic names. And it might not be in the direction of building a universal ontology of geographic classifiers. As the Perseus team has demonstrated in their recent findings, mapping of classifiers from one system to another, and across languages or domains of specialization brings up a host of problems. **[Babeu *et al*, 2007]** The work being done to map multiple digital collections to CIDOC CRM **[CIDOC]** offers us some hope for this process, but does not offer any immediate means of integrating such diverse geographic toponym domains as those found in Chinese History, the historical peoples and states of the African continent, or the far flung diaspora of Oceania.

Instead of dealing at the outset with the complexity of multilingual and spatio-temporal semantic classifications systems, why not begin with a temporal framework, in which chronologies may be immediately related to each other using the ISO 8601 Standard for representation of Date and Time? **[ISO 8601]** In other words, let us collect and integrate what we know about human history into a single standard timeline. For the CHGIS project we published a basic chronology of all the Chinese reign periods from 249 BCE to 1911 CE. Each entry contains the Period Name, Sub-period Name, Reign Name, Begin Year and End Year. **[CHGIS Chronology]** Other named Time Periods have been extracted from the Library of Congress and University of California MELVYL catalog subject headings. **[Feinberg 2003] [Petras *et al*, 2005]** It will be a straightforward task to create groupings of such named time periods, similar to the period attestations of the Pleiades project mentioned earlier, and then publish them under open source licenses for all to use.

Why go to such trouble over chronologies? Precisely because the localized date information becomes the best identifier for disambiguation of historical entities, geographic or otherwise. To give an example, if we attempt to disambiguate the toponym "Fu Zhou" by spelling alone, we would match 13 records in the CHGIS search engine. If we were to add the parent attribute "China" to our query it would gain us nothing, nor would knowing the parent ADM1 level administrative unit "Fujian" help us much. However, if we were to request either the year 970 CE or the reign period "Kai Bao" which lasted from 968 CE to 979 CE, we would get exactly two matches if searching for the romanized string, and exactly one match if searching for the Chinese Character string. A direct hit. It makes no difference at all that the parent unit of Fu Zhou during the year 970 was of the administrative unit type "*jiedushi*" (military commission). All we needed was the vernacular script string "Fu Zhou" and year "970."

In addition to the advantages gained in searching and geocoding processes, chronologies offer the most practical means of dating historical placenames. Whereas a query, automated or otherwise may not know the exact year of validity for a particular toponym, it could contextually derive period information. A placename search for the toponym "Valencia" will result in numerous results (especially in Spain and the Phillipines), but only one if related to the "Black Death," which could be mapped to the temporally to the years 1347 to 1351 CE, and geographically to the coast of Spain. Even more interesting would be to map all the objects associated with the same period, and then begin to mark them up with more accurate date attestations. In this way spatio-temporal networks can be developed to test and improve our existing knowledge of historical geography.

In conclusion I would suggest that building upon the existing infrastructure of named time periods within a unified schema (such as the Historical Event Markup and Linking Project) **[HEML]** or the MIT Simile Timeline RDF format **[Simile Timeline]** we could quickly move towards an integrated historical chronology standard, which would serve not only to improve the results of toponym disambiguation processes, but to enhance the development of scalable global historical geographic information systems.

Notes

ADL Alexandria Digital Library Gazetteer Content Standard, 2004. http://www.alexandria.ucsb.edu/gazetteer/

AWMC Ancient World Mapping Center http://www.unc.edu/awmc/

Babeu *et al*, 2007 Named Entity Identification and Cyberinfrastructure, Alison Babeu, David Bamman, Gregory Crane, Robert Kummer, and Gabriel Weaver. Preprint for ECDL 2007.

Barrington, 2000 Barrington Atlas of the Greek and Roman World, Richard J. A. Talbert (ed). Princeton Univ Press, 2000.

Berman, 2003 A Data Model for Historical GIS: The CHGIS Time Series, Merrick Lex Berman. CHGIS, 2003. http://www.fas.harvard.edu/~chgis/work/docs/papers/tech_specs.html

CHGIS China Historical GIS http://www.fas.harvard.edu/~chgis

CHGIS Chronology Search Engine and Downloadable table of Chinese Reign Periods. http://www.people.fas.harvard.edu/~chgis/work/downloads/faqs/chronology.html

CIDOC CIDOC Conceptual Reference Model http://cidoc.ics.forth.gr/

CITAS China in Time and Space http://citas.csde.washington.edu/

Feinberg 2003 Application of Geographic Gazetteer Standards to Named Time Periods, Melanie Feinberg in consultation with Ruth Mostern, Susan Stone, and Michael Buckland. http://ecai.org/imls2002/time_period_directories.pdf

GBHGIS Great Britain Historical GIS http://www.visionofbritain.org.uk/

GNS GEOnet Names Server http://earth-info.nga.mil/gns/html/

GeoNames GeoNames http://www.geonames.org/

HEML Historical Event Markup Language http:// **heml**.mta.ca/

Heurist Heurist Scholar Collaborative Knowledge Space http://heuristscholar.org/

Metacarta http://metacarta.com/

Mostern, Johnson From Named Place to Naming Event: Creating Gazetteers for History, Ruth Mostern and Ian Johnson, International Journal on GIS (forthcoming).

NHGIS National Historical GIS of the United States http://nhgis.org/

Perseus Perseus Digital Library, Tufts University http://www.perseus.tufts.edu/

Perseus Version 4 http://www.perseus.tufts.edu/hopper/

Petras *et al*, 2005 Leveraging Library of Congress Subject Headings to Improve Search for Events – A Time Period Directory, Vivien Petra, Matt Meiske, Ray Larson, Jeanette Zernecke, Kim Carl and Michael Buckland, 2005. http://metadata.sims.berkeley.edu/tpd/TPD-report.pdf

Pleiades Pleiades Project http://pleiades.stoa.org/

RAS Russian Historical Maps of China, Russian Academy of Sciences, 2003. http://www.fas.harvard.edu/~chgis/data/rus_geo/

Simile Timeline MIT DHTML Timeline http://simile.mit.edu/timeline/

Smith *et al*, 2001 Disambiguating Geographic Names in a Historical Digital Library, David A. Smith and Gregory Crane. Perseus Project, 2001. http://www.cs.jhu.edu/~dasmith/geodl01.pdf

Volz *et al*, 2007 Towards ontology based disambiguation of geographical identifiers, Raphael Volz, Joachim Kleb, and Wolfgang Mueller. Presented at the 16th Intl WWW Conference (Banf, Canada), 2007. http://www2007.org/workshops/paper_132.pdf