Historical Gazetteer Integration:
CHGIS, Regnum Francorum & GeoNames

Working Digitally with Historical Maps
AAG 2012

Merrick Lex Berman & Johan Åhlfeldt
Harvard University & Regnum Francorum
Geographic Info Retrieval  [GIR]

Search for Place Names, features, and / or locations

Queries to existing gazetteer services **work** for:

names, feature types, footprints, administrative districts…

GeoTemporal Info Retrieval  [GTIR]

Search for Place Names, events, and / or locations with dates

Queries to existing gazetteers **don't work** because:

dates are optional or entirely missing
Extend GeoNames Schema By Adding Dates

<table>
<thead>
<tr>
<th>geoname</th>
</tr>
</thead>
<tbody>
<tr>
<td>geonameid int</td>
</tr>
<tr>
<td>name varchar(200)</td>
</tr>
<tr>
<td>asciiName varchar(200)</td>
</tr>
<tr>
<td>alternatenames varchar(6000)</td>
</tr>
<tr>
<td>latitude float</td>
</tr>
<tr>
<td>longitude float</td>
</tr>
<tr>
<td>fclass char(1)</td>
</tr>
<tr>
<td>fcode varchar(10)</td>
</tr>
<tr>
<td>country varchar(2)</td>
</tr>
<tr>
<td>cc2 varchar(80)</td>
</tr>
<tr>
<td>admin1 varchar(20)</td>
</tr>
<tr>
<td>admin2 varchar(80)</td>
</tr>
<tr>
<td>admin3 varchar(20)</td>
</tr>
<tr>
<td>admin4 varchar(20)</td>
</tr>
<tr>
<td>population bigint</td>
</tr>
<tr>
<td>elevation int</td>
</tr>
<tr>
<td>gtopo30 int</td>
</tr>
<tr>
<td>timezone varchar(40)</td>
</tr>
<tr>
<td>moddate date</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>alternate name</th>
</tr>
</thead>
<tbody>
<tr>
<td>alternatenameId int</td>
</tr>
<tr>
<td>geonameid int</td>
</tr>
<tr>
<td>isoLanguage varchar(7)</td>
</tr>
<tr>
<td>alternateName varchar(300)</td>
</tr>
<tr>
<td>isPreferredName boolean</td>
</tr>
<tr>
<td>isShortName boolean</td>
</tr>
<tr>
<td>isHistoricalName boolean</td>
</tr>
<tr>
<td>existDate date</td>
</tr>
<tr>
<td>endDate date</td>
</tr>
<tr>
<td>srcCiteName varchar(300)</td>
</tr>
<tr>
<td>srcCiteLink varchar(300)</td>
</tr>
</tbody>
</table>
Testbed Methodology

- Automate process for matching placenames via open API
- Test methods for geonomial and geospatial matching
- Discover problems in current datasets, schemas
- Evaluate matching results
- Propose a schema for exchange of historical placename data
- Recommend the types of attributes needed to refine matching
Existing Gazetteers as Web Services

**Google Maps API Web Services**

- The Google Geocoding API
  - What is Geocoding?
  - Audience
  - Usage Limits
  - Geocoding Requests
  - Geocoding Responses
    - JSON Output Formats
    - XML Output Formats
    - Status Codes
    - Results
    - Address Component Types
  - Reverse Geocoding

**GeoNames WebServices overview**

<table>
<thead>
<tr>
<th>WebService</th>
<th>XML</th>
<th>JSON</th>
<th>RDF</th>
<th>CSV</th>
<th>TXT</th>
<th>RSS</th>
<th>KML</th>
</tr>
</thead>
<tbody>
<tr>
<td>children</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cities</td>
<td>XML</td>
<td>JSON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>countryCode</td>
<td>XML</td>
<td>JSON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TXT</td>
</tr>
<tr>
<td>countryInfo</td>
<td>XML</td>
<td>JSON</td>
<td></td>
<td>CSV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>findNearbyPlaceName</td>
<td>XML</td>
<td>JSON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gtopo30</td>
<td>XML</td>
<td>JSON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TXT</td>
</tr>
<tr>
<td>hierarchy</td>
<td>XML</td>
<td>JSON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neighbourhood</td>
<td>XML</td>
<td>JSON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>search</td>
<td>XML</td>
<td>JSON</td>
<td></td>
<td></td>
<td>RDF</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Placemaker**

**PLACES V. PLACE NAMES**

Yahoo! Placemaker aims to capture all forms of how a place is called, and disambiguate the place-name to its canonical form. The platform identifies and disambiguates every place-name to a specific place concept, referenced by its unique identifier, the Where-on-Earth ID (WOEID). WOEIDs always reference a place, not a place name. For example, "New York", "New York City", "NYC", and "the Big Apple" are all variant names for WOEID 2459115. If Placemaker find these variants in the text, it will understand them to be multiple appellations of the same place.
Existing Historical Gazetteers

Gazetteer of 50,000+ historical placenames
Administrative Hierarchy
Points have description of “present location”

Gazetteer of 10,000+ historical placenames
Administrative Hierarchy
Includes explicit links to GeoNames ID

Gazetteer of 120,000+ placenames
Administrative Unit Ontology
Query API or tabular datasets?
State of the Art - Locations in Classical Texts and Atlases

Perseus Project

Pleiades

Google Ancient Places

Barrington Atlas

Digital Atlas of Roman and Medieval Civilization
State of the Art - Web Services for historical placenames

Unlock: geoparsing (place name text mining and mapping)

Chalice: creating linked data historic gazetteer through text mining

DEEP: extend Chalice by digitizing the 86 volumes of the English Place Name Survey

Unlock web service only - UK only

Chalice Schema not published

DEEP plans to develop schema for historical place name queries
Gazetteer Augmentation

Queries directed to different resources

Results augment existing records

Geofeature Set
Sample data for testbed: CHGIS

Historical Place Details:

- **name:** Tengyue Ting
- **vernacular:** 腾越厅
- **alternate:** 腾越廂
- **feature type:** independent sub-prefecture 直隶厅 Zhiltian
- **date range:** 1820 to 1820
- **present location:** 云南騰沖县

**admin hierarchy SUPERIOR units:**
from 1820 to 1820 Tengyue Ting was part of: 云南 Yunnan

**admin hierarchy SUBORDINATE units**

- **coordinate type:** point location
- **latitude:** 25.024349 N
- **longitude:** 98.49498 E

[Google 地图 - get Google Map - get CHGIS Map]

Historical placename: Tengyue Ting

Present placename: 云南騰沖县

Romanized present placename: yunnantengchongxian

Time series county seats: 403 placename records
Sample data for testbed: Regnum Francorum, France

### Historical placename:
Éclance

### GeoNames ID:
3020485

### France places in RFO:
4164 placename records

<table>
<thead>
<tr>
<th>num</th>
<th>oid</th>
<th>oname</th>
<th>lng</th>
<th>lat</th>
<th>insee</th>
<th>geon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11051</td>
<td>Ébreuil</td>
<td>3.0882</td>
<td>46.1161</td>
<td>3107</td>
<td>3020569</td>
</tr>
<tr>
<td>2</td>
<td>7856</td>
<td>Échemines</td>
<td>3.8315</td>
<td>48.3886</td>
<td>10134</td>
<td>3020517</td>
</tr>
<tr>
<td>3</td>
<td>8958</td>
<td>Éclance</td>
<td>4.6354</td>
<td>48.3058</td>
<td>10135</td>
<td>3020485</td>
</tr>
<tr>
<td>4</td>
<td>8607</td>
<td>Écoivres</td>
<td>2.6781306</td>
<td>50.3445124</td>
<td>62283</td>
<td>3020470</td>
</tr>
<tr>
<td>5</td>
<td>3664</td>
<td>Écommoy</td>
<td>0.2757986</td>
<td>47.8300763</td>
<td>72124</td>
<td>3020462</td>
</tr>
<tr>
<td>6</td>
<td>2351</td>
<td>Écouen</td>
<td>2.3791247</td>
<td>49.0184969</td>
<td>95205</td>
<td>3020440</td>
</tr>
</tbody>
</table>
GIS sampling of CHGIS names and GeoNames

Distribution of study sample

2km and 8km buffers + GeoNames

Sample Placenames
403 historical features
GIS sampling of France names and GeoNames

Distribution of study sample

2km buffers + GeoNames

Accuracy of France Matches often within 250M
Matching process

**REQUEST**

Sample Placename [from Source Gazetteer]

- **Geospatial Filter**
  - 2km
  - 8km
  - Buffer Results

- **Augment Results**
- **Reverse Geocoding**

**Geospatial AND Geonomial Processing**

- **Genominal Filter**
- **Preprocess Source & Target Strings to Match**
- **Source to Target Y/N**
- **Target to Source Y/N**
- **Geonomial Results**

**Augment AND Publish**

- **Results Ranking**
- **Augment Source**
- **Publish XML**
Why trim to five characters?  Why match in both directions?

"Henansheng Puyang Shi Puyang Xian Qinghetou Xiang"  to  "Puyang"

After special chars trim:

"HenanshengPuyangShiPuyangXianQinghetouXiang" to  "Puyang"

After first five chars trim:

Source to Target:

"Henan" to  "Puyang"  no match

"Puyan" to  "HenanshengPuyangShiPuyangXianQinghetouXiang" match
Results based on distance filtering

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total Names</th>
<th>2km Hits</th>
<th>Avg &lt; 2km</th>
<th>8km Hits</th>
<th>Avg &lt;8km</th>
<th>Avg All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yunnan</td>
<td>403</td>
<td>297</td>
<td>1.15 km</td>
<td>1994</td>
<td>4.67 km</td>
<td>6.07 km</td>
</tr>
<tr>
<td>France</td>
<td>4164</td>
<td>17705</td>
<td>0.84 km</td>
<td>n.a.</td>
<td>n. a.</td>
<td>1.02 km</td>
</tr>
</tbody>
</table>
Results based on string matching within buffer distances

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total Names</th>
<th>One-way match</th>
<th>Two-way match</th>
<th>Source to Target</th>
<th>Target to Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>4164</td>
<td>97%</td>
<td>86%</td>
<td>97%</td>
<td>192%</td>
</tr>
<tr>
<td>Yunnan</td>
<td>403</td>
<td>74%</td>
<td>18%</td>
<td>27%</td>
<td>87%</td>
</tr>
</tbody>
</table>

Main discovery:

Historical placenames not often found in GeoNames alternates

Much more likely for current GeoName to be found in historical gazetteer record when “present location” is attested
Schema for the augmented gazetteer
Schema for the augmented gazetteer

Linda Hill - ADL
Augmented gazetteer for faceted geo-temporal queries

- Placenames
- Dates of existence (or related time periods)
- Administrative jurisdictions (past and present)
- Alternative footprints
Existing XML schema for historical placenames (for CHGIS)

<query>

<results>

<item>
  <placename>
  <feature_type>
  <temporal>
  <spatial>
  <part_of>
  <evidenced_by>
  <preceded_by>
  <links>
Old schema - reorganized for multiple placename attestations

<placename>
  <name_romanized>Fu Zhou
  <name_vernacular>抚州
  <name_alternate>撫州
<temporal>
  <begin_year>1161
  <begin_year_rule>4
  <end_year>1276
  <end_year_rule>4
<part_of>
  <parent_name>江南西路
  <part_of_begin>1134
  <part_of_end>1276
New schema - need to allow for multiple placename attestations

<placename>
  <name>撫州</name>
  <lang>zh</lang>
  <lang_class>traditional Chinese</lang_class>
  <variant>N</variant>
  <begin>1161</begin>
  <begin_attestation>jurisdiction area increased</begin_attestation>
  <end>1276</end>
  <end_attestation>placename changed</end_attestation>
</placename>

<parent_address>江南西路</parent_address>
<parent_begin>1134</parent_begin>
<parent_end>1276</parent_end>

<source>CHGIS</source>
<verified>Y</verified>
<name_rank>default</name_rank>
<counterclaim>N</counterclaim>
New schema - need to allow for multiple placename attestations

<placename>
    <name>Linchuan</name>
    <lang>zh</lang>
    <lang_class>transliteration</lang_class>
    <variant>N</variant>
    <begin>20120212</begin>
    <begin_attestation>exists</begin_attestation>
    <end>20120212</end>
    <end_attestation>exists</end_attestation>

    <parent_address>Linchuan, Fuzhou, Jiangxi, China</parent_address>
    <parent_begin>20120212</parent_begin>
    <parent_end>20120212</parent_end>

    <source>Google</source>
    <verified>N</verified>
    <name_rank>default</name_rank>
    <counterclaim>N</counterclaim>
Publications & Resources

http://fas.harvard.edu/~chgis/gazetteer