Mapping Text: Automated Geoparsing and Map Browser for Electronic Theses and Dissertations

Kathy Weimer
Professor and Curator of Maps

Naga R. Modala
Research Assistant

James Creel
Sr. Software Applications Developer

Rohit Gargate
Research Assistant

Texas A&M University Libraries
Overview

- Background
- Project concept
- Map based interface
- Geoparser
- Lessons learned
- Future plans
University Background & ETDs

• Founded in 1876 as land-grant university
  – Land-, sea and space-grant university
  – Formerly military college
• 50,000 student body
• 240 Masters and PhD programs
  – Ranks in Top 10 universities in the number of science and engineering doctorates produced
  – Ranks in Top 20 in number of doctoral degrees awarded to minorities
• 2004 = mandate for digital T&D
• Now = > 10,000 born digital theses & dissertations in repository
Why Map a Textual Collection?

- Increase attention and access to the collection
- Presents a unique context
- Visualize interconnections in the locations of study
- Interactive & visual format appeals to users
- Fills conceptual gaps in traditional cataloging of places
- Increasing amount of place based queries (Ahlers)
- Benefits of spatial queries (Larson) for adjacency, proximity, etc.
Project Aims and Scope

To create tools for and increase understanding of:

• Geoparsing
• Automated Metadata Creation
• Map Based Search Interfaces for Digital Collections
• Use of Digital Gazetteers
Collaborations

• TAMU Map & GIS Library
  – Created an early prototype of map showing T&D locations of study
  – AMIGOS Fellowship (Weimer)

• TAMU Library Digital Initiatives
  – Staff support
  – IT expertise

• TAMU Thesis & Dissertation Office
  – Provided sample set

• Texas Digital Library (TDL)
  – Holds collection in DSpace
  – Enhance collection access

• TAMU Initiative for Digital Humanities, Media and Culture
  – Interest and support for base methodology and wider applications
Geoparsing Enables a Map Based Interface

**Goal is to automate geocoding**

- Match toponym in text against gazetteer
- Protocol for place name disambiguation
- Obtain geographic coordinates from gazetteer
- Encode coordinates and other item metadata in KML
- Render KML in a specialized map with link to ETD in repository
Desired Map Functionality

- Read KML output from geoparser
- Base map: GoogleMaps, OpenLayers, OpenStreetMaps
- Marker clustering and List of placemarks
- Dropdown menu for countries and states
- Dropdown menu for departments grouped by college
- Search by author
- Time range slider (by year)
- Use the University Brand color palette
## Metadata in KML file

- **Author**
- **Title**
- **Academic department**
- **Advisor**
- **PhD or Master**
- **Year**
- **Place** (*created via geoparsing*)
- **Keywords**
- **URL to document**

```plaintext
dc.creator
dc.title
thesis.degree.department
dc.contributor.advisor
thesis.degree.level
dc.date.submitted
dc.coverage.spatial
dc.subject
dc.identifier.uri
```
Zoom to location of interest
Geoparser

- Comparable Models
  - Edinburgh (Grover, et al.)
  - DIGMAP (Martins, et al.)

- Setting
  - DSpace 1.7 + supports curation tasks
  - Suggest New Metadata
Name Extraction & Disambiguation

- Name Extraction
  - ‘Named Entity Recognition’ or NER
  - OpenNLP, Stanford NLP, Mallet
  - Classifies spans of text based on freely available training data
  - Toponym occurrences are recorded in the document

- Disambiguation
  - Requires reliable knowledge base
  - Geonames.org
  - Methods: Rule-based, Heuristic, Statistical
Heuristics

**Context Based:**
- Unambiguous extended names i.e. “Paris, France”
- Favor candidates of mentioned feature type
- Clustering of places (‘nearby locations’)
- Favor contained candidates

**Generalized:**
- Favor higher-level administrative units (countries, states, cities)
- Favor locations of larger population
Evaluate Output

- Compare human annotations to automated output
- Examine precision & recall of name extraction
- Examine accuracy of name disambiguation
Lessons Learned

- **Geonames**
  - Web look up returns are unclear as to how results are prioritized
  - Web look up is done by name but returns places without the search term in their name – due to inclusion of the search term in the hierarchy
  - Suggested best practice – put geonames dataset into your own database

- **OpenNLP** - lots of false positives on short strings (eg. Ca, Me)

- Implementing name extraction is comparatively easier with Stanford NLP
Future Plans

- Use statistical techniques for name disambiguation
- Consider relevance of toponyms when performing name extraction
- Evaluate the tool on other digital collections
- Improve the scalability of the map on large data sets
- Integrate the tool into document submitter/curator workflow
Questions?

Kathy Weimer
k-weimer@library.tamu.edu

James Creel
jcreel@library.tamu.edu