Determining and Mapping Locations of Study in Theses and Dissertations: A Visualization Tool for ETDs

Katherine H. Weimer
James Creel
TxETD Conference
February 28, 2013
ETDs

• Grey literature
• In recent years – ‘born digital’ = repository
• Represents cutting edge research across the university
• Value in showing subject matter
• Increase access → Increase citation rates
Information Retrieval Challenge

• How access locations represented in text?
  – 20% of web queries have a geographic relation (Ahlers)

• Gaps in keyword or subject search in traditional catalog

• Location information is increasingly in demand (Reid)
What about a Visual Search?

- ETDs searchable by map interface?
  - Visual representation of research output
  - Enhance possibilities for collaborations and networking
  - Enhance access to the collection
Sample = ETDs from 2005

- Prototype work began in 2008-2009
- Abstracts – reviewed for mention of locations
- 1049 total ETDs in 2005
- 300 included locations (< 30%)
- 130 contained international locations (> 10%)
Metadata for 2005 ETDs

• Pulled from repository:
  – Author name
  – Department
  – Major
  – Graduation date
  – Abstract title
  – Handle

Data provided by Laura Hammons, TAMU Thesis Office
Manual lookup in Official Gazetteer

• *US* = USGS/ Board on Geographic Names
  – [GNIS](http://www.geonames.org) Geographic Names Information System

• *Foreign* = National Geospatial-Intelligence Agency - [GNS GEOOnet names server](http://geonames.org)

• Includes:
  – official and variant names
  – coordinates (point not polygon)
  – elevation
  – administrative hierarchy
Early Map Prototype
Title: Cost of being a Mexican immigrant and being a Mexican non-citizen in California and Texas
Author: Takei, Isao
Major: Sociology
Location: California, Texas
See also Texas
2010-2012 AMIGOS Fellowship

• Fellowship and Opportunity Award Program ([www.amigos.org/node/497](http://www.amigos.org/node/497))
  – Two year grant
  – Student worker/programmer
  – *Data collection for manual lookup*
  – *Map search interface creation*
2011 – Geoparsing Work Begins

1. Overarching *goal is to automate geocoding*
2. Match toponym in text against gazetteer
3. Protocol for place name disambiguation
4. Obtain geographic coordinates from gazetter
5. Encode coordinates in item surrogates for map-based view
6. Create map with link to ETD in repository
Issues & Quality Control with Metadata Creation

- Disambiguation of names
- Proper functionality of the gazetteer
- Define portion of text to be searched
Approach to Interactive Map

- Use HTML to create the web interface
- Use KML to create the placemarks and to embed hyperlinks
- Use Javascript to create Google Map and to import the KML file
  - Geoxml3 is a KML processor for the Google Maps JavaScript API V3
Webpage Model

- EPA Greenhouse Gas Emissions

http://ghgdata.epa.gov/ghgp/main.do
Desired Map Functionality

1. Base map: use GoogleMaps and others
2. Marker clustering of placemarks
3. List of the displayed placemarks
4. Dropdown menu for countries and states in the US
5. Dropdown menu for departments grouped by college
   1. Selection of multiple departments in more than one college
   2. If selecting the college, then select all departments within the college
6. Search by author
7. Time range slider (by year)
8. Use the Web-friendly University Brand color palette
Web page Plan Part 1: Google map of the world
Web page Plan Part 2: Marker clusters of placemarks

- Placemarks will contain a brief description of the thesis or dissertation and a hyperlink to the document.
- Large numbers of placemarks could be represented by marker clusters:
  - In the EPA map, there are 50 place marks in the Austin and San Antonio area. When a user clicks on the cluster additional clusters or place marks are displayed.
Web page Plan
Part 3:
List of displayed placemarks

<table>
<thead>
<tr>
<th>Facility Name/Location</th>
<th>2010 Emissions (metric tons CO2e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>121 REGIONAL DISPOSAL FACILITY MELISSA, TX, 75454</td>
<td>114,408</td>
</tr>
<tr>
<td>23rd and 3rd BROOKLYN, NY, 11232</td>
<td>82,955</td>
</tr>
<tr>
<td>29-6 #2 Central Delivery Point Blanco, NM, 87412</td>
<td>25,175</td>
</tr>
<tr>
<td>30-5 Central Delivery Point Compressor Station Aztec, NM, 87410</td>
<td>38,881</td>
</tr>
<tr>
<td>31-6 Central Delivery Point BLOOMFIELD, NM, 87413</td>
<td>63,881</td>
</tr>
<tr>
<td>32-7 Central Delivery Point BLANCO, NM, 87412</td>
<td>56,476</td>
</tr>
<tr>
<td>32-8 NO2 CDP COMPRESSOR STATION BLANCO, NM, 87412</td>
<td>55,072</td>
</tr>
<tr>
<td>3M BROWNWOOD BROWNWOOD, TX, 76801</td>
<td>36,877</td>
</tr>
<tr>
<td>3M CO MAPLEWOOD, MN, 55144</td>
<td>75,800</td>
</tr>
<tr>
<td>3M CO - GUIN GUIN, AL, 35563</td>
<td>28,749</td>
</tr>
</tbody>
</table>
Web page Plan Parts 4 & 5: Dropdown menu and Checkboxes

• Dropdown menu can allow selection of countries and US states

• Checkboxes can allow for selection by College or Department.
Web page Plan Parts 6 & 7: Search options and Time slider

- Users can search by keywords, location, author and year using the search tools
- Time slider allows the users to refine the display by range of years
Data contained in the KML file

- The KML file with locations includes:
  - Author
  - Title
  - Academic department
  - Advisor
  - PhD or Master
  - Year
  - Place
  - Keywords
  - Url to document

- Info box displays:
  - Author
  - Title
  - Academic department
  - PhD or Master
  - Year
  - Place
  - Url to document
Beta Version of Map:  
Showing Google Street Maps
Search by Author or Range of Years
Limit by Department, Location and Year
Zooming in, Clusters Expand
Selection Options
Zooming Further, Clusters Expand, Refining Selections
Hover over Cluster, Table of Results by Title

<table>
<thead>
<tr>
<th>College of Agriculture and Life Sciences</th>
<th>College of Architecture</th>
<th>Bush School of Government &amp; Public Service</th>
<th>Mays Business School</th>
<th>College of Education &amp; Human Development</th>
<th>Dwight Look College of Engineering</th>
<th>College of Geosciences</th>
<th>College of Liberal Arts</th>
<th>College of Science</th>
<th>College of Veterinary Medicine &amp; Biomedical Sciences</th>
</tr>
</thead>
</table>

### In Ballast Water: The Shipping Industry's Contributions to the Transport and Distribution of Microbial Species in Texas

- Ray; Texas, Nueces County (bay)
- 2009-08
- Neyland, Elizabeth B.
- Masters
- Golden, Susan
- Brinkmeyer, Robin
- Biology

### Bacteria in Ballast Water: The Shipping Industry Transport and Distribution of Microbial Species

- Nueces River; Texas, Nueces County (stream)
- Year:
- Author:
- Level:
- Advisor:
- Department:
User clicks on Point of Interest → Title and Metadata Appear with Link to Text
Automated Process / Geoparser

- Review Comparable Models
- Setting
- Pre-processing
- Name Extraction
Geoparser: Comparable Models

• Edinburgh Geoparser
  – Grover, et. al. used OCR with historic records, provided GeoCrossWalk

• DIGMAP Geoparser
  – Martins, et al. used originally for DIGMAP digital library of historic maps
Geoparser: Setting

• Dspace 1.7 supports curation tasks
  – Custom Java programs
• ETD Geoparser project:
  – Suggest New Metadata
  – Generate KML
A Multi-step Process

- Geoparsing involves several stages addressing two main problems:
  1) Name extraction
  2) Name disambiguation
Geoparser Workflow
(‘map’ refers to data structure)
Geoparser: Pre-Processing

- DSpace filter-media script extracts plain-text from PDFs.
- Suggest New Metadata curation task
  - Partitions the document into sections using regular expressions tailored to our ETDs
- Geoparser logic
  - Excludes sections containing toponyms not part of the subject matter (author-affiliation locations, conference locations, etc.)
Name Extraction

- ‘Named Entity Recognition’ or NER
- Various robust and freely available tools:
  - Stanford NLP
  - OpenNLP
  - Mallet
Geoparser: Name Extraction

• Uses OpenNLP or Stanford NLP
• Classifies spans of text based on freely available training data
• Tokenization, Name-finder
• Toponym occurrences are counted in the document
Name Disambiguation

- Requires reliable data- or knowledge-base
- We employ the Geonames dataset
  - Conglomeration of International gazetteers
    - Includes GNIS (USGS)
- Several complimentary methods
  - Rule-based
  - Heuristic
  - Statistical
Heuristics: Overview

• Several heuristics help indicate the probable referent of a given toponym
• Some - based on context-clues in text
• Others - based on general observations about human discourse
Heuristics: Context-based

- One document, one sense
- Unambiguous extended names i.e. “Paris, France”
- Clustering of places (‘nearby locations’)
- Favor contained candidates
- Favor candidates of mentioned feature code
Heuristics: Generalized

• Favor:
  – Higher-level administrative units (countries, states, cities)
  – Locations of larger population

• Heuristics - grouped into refinement iterations and then applied sequentially
Ongoing Quality Assessment

- Comparison of human annotations to geoparser output
- Deeper analysis of accuracy
Future Work

• Explore statistical disambiguation
• Explore relevance of toponyms to the subject matter
• Expand to other ETD years
• Expand to all TDL ETDs
• Expand to other digital collections or collection types, even the library catalog?

There is still much work to be done!
References

- Ahlers & Boll, “Location Based Web Search” in *The Geospatial Web*
- DigMap. portal.digmap.edu/
- Reid, James. “GeoXwalk – A Gazetteer Server and Service for UK Academia” (ECDL 2003)
Contact:

– Kathy Weimer
k-weimer@library.tamu.edu

– James Creel
jcreel@library.tamu.edu