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

Kathy Weimer
Professor and Curator of Maps

Naga R. Modala
Research Assistant

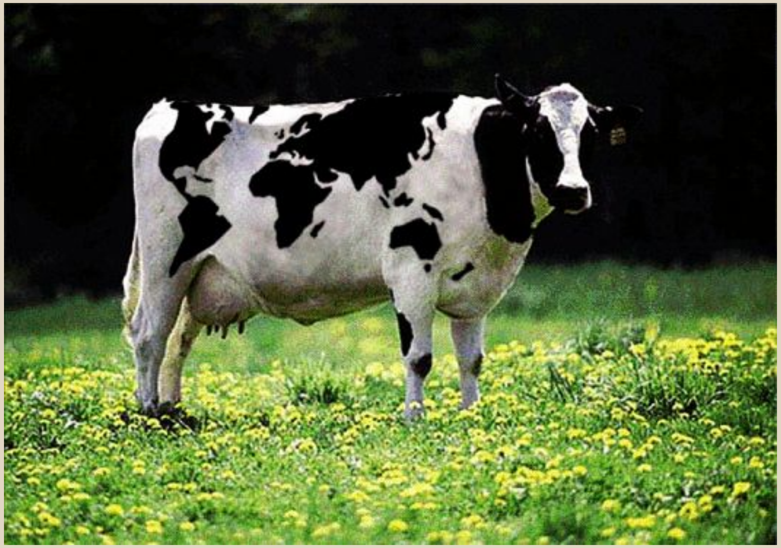
James Creel
Sr. Software Applications Developer

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***

rce: h[illegible]es.tamu.edu/tag/johnny-manzi

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

TEXAS A&M UNIVERSITY

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

LIBRARIES
TEXAS A&M UNIVERSITY
MAP AND GIS LIBRARY

Amigos Library Services
Resource Sharing Through Technology

ETD Policies

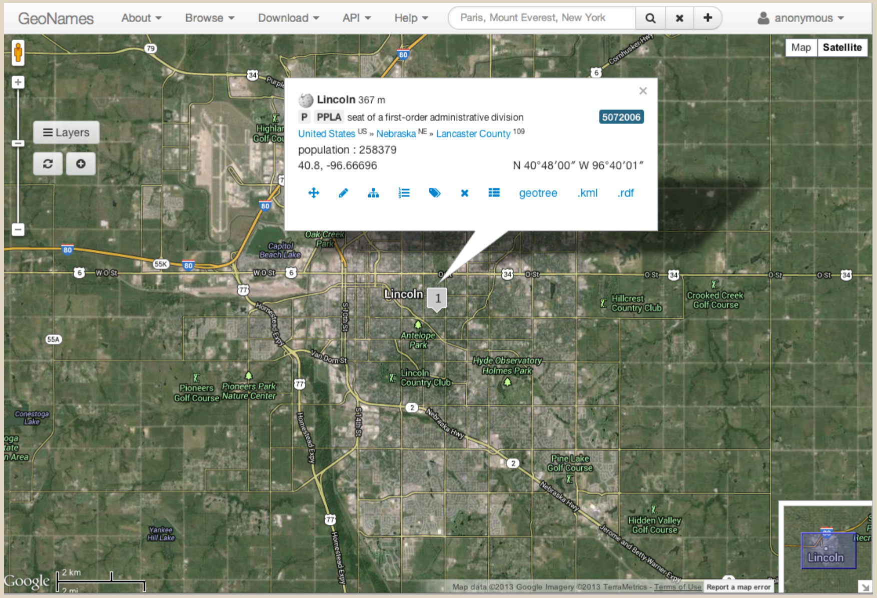
TDL.ORG
TEXAS DIGITAL LIBRARY

IDHMC
TEXAS A&M UNIVERSITY
Initiative for Digital Humanities, Media, and Culture

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



Lincoln 367 m

P PPLA seat of a first-order administrative division 5072006

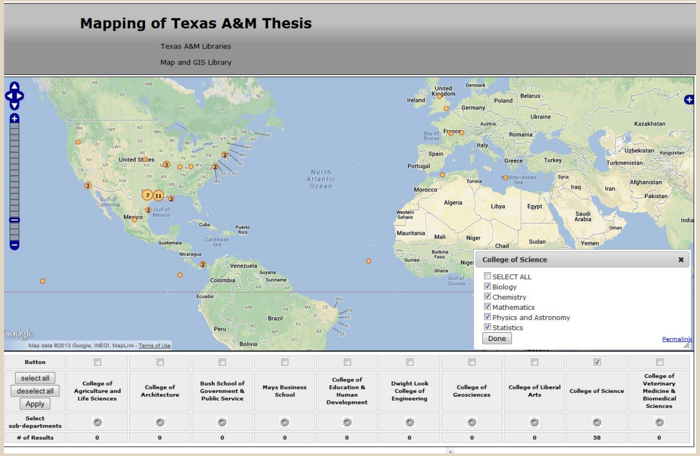
United States US » Nebraska NE » Lancaster County 109

population : 258379

40.8, -96.66696 N 40°48′00″ W 96°40′01″

Desired Map Functionality

- Read KML output from geoparser
- Base map: GoogleMaps, OpenLayers, Open StreetMaps
- 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



Mapping of Texas A&M Thesis

Texas A&M Libraries

Map and GIS Library

College of Science ✖

- ☐ SELECT ALL
- ☑ Biology
- ☑ Chemistry
- ☑ Mathematics
- ☑ Physics and Astronomy
- ☑ Statistics

Done

Permalink

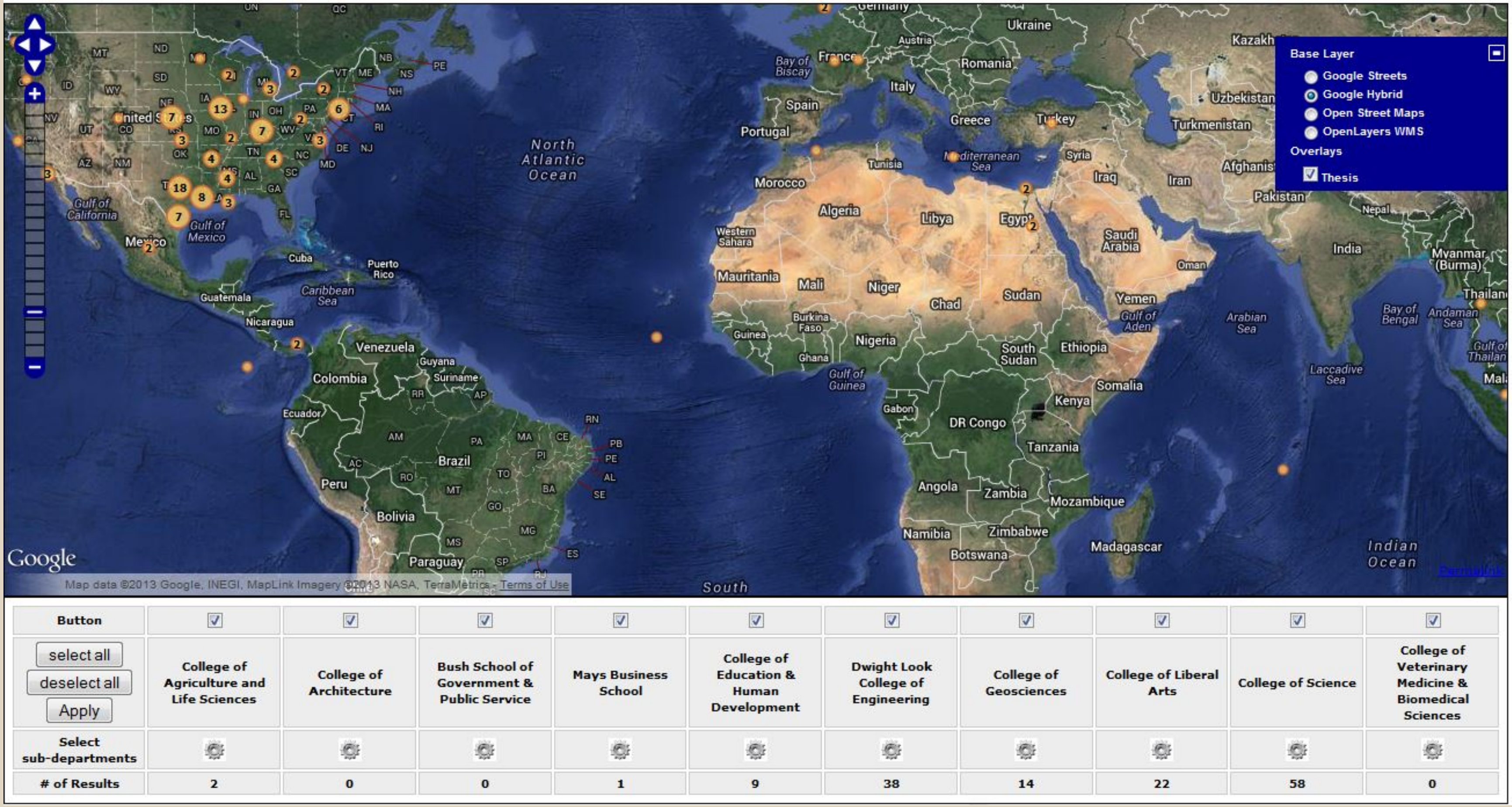
Map data ©2013 Google, INEGI, MapLink - Terms of Use

Button	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐
select all deselect all Apply	**College of Agriculture and Life Sciences**	**College of Architecture**	**Bush School of Government & Public Service**	**Mays Business School**	**College of Education & Human Development**	**Dwight Look College of Engineering**	**College of Geosciences**	**College of Liberal Arts**	**College of Science**	**College of Veterinary Medicine & Biomedical Sciences**
Select sub-departments	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙
# of Results	0	0	0	0	0	0	0	0	58	0

Metadata in KML file

• Author	*dc.creator*
• Title	*dc.title*
• Academic department	*thesis.degree.department*
• Advisor	*dc.contributor.advisor*
• PhD or Master	*thesis.degree.level*
• Year	*dc.date.submitted*
• Place *(created via geoparsing)*	*dc.coverage.spatial*
• Keywords	*dc.subject*
• URL to document	*dc.identifier.uri*

Map Prototype



Base Layer

○ Google Streets

● Google Hybrid

○ Open Street Maps

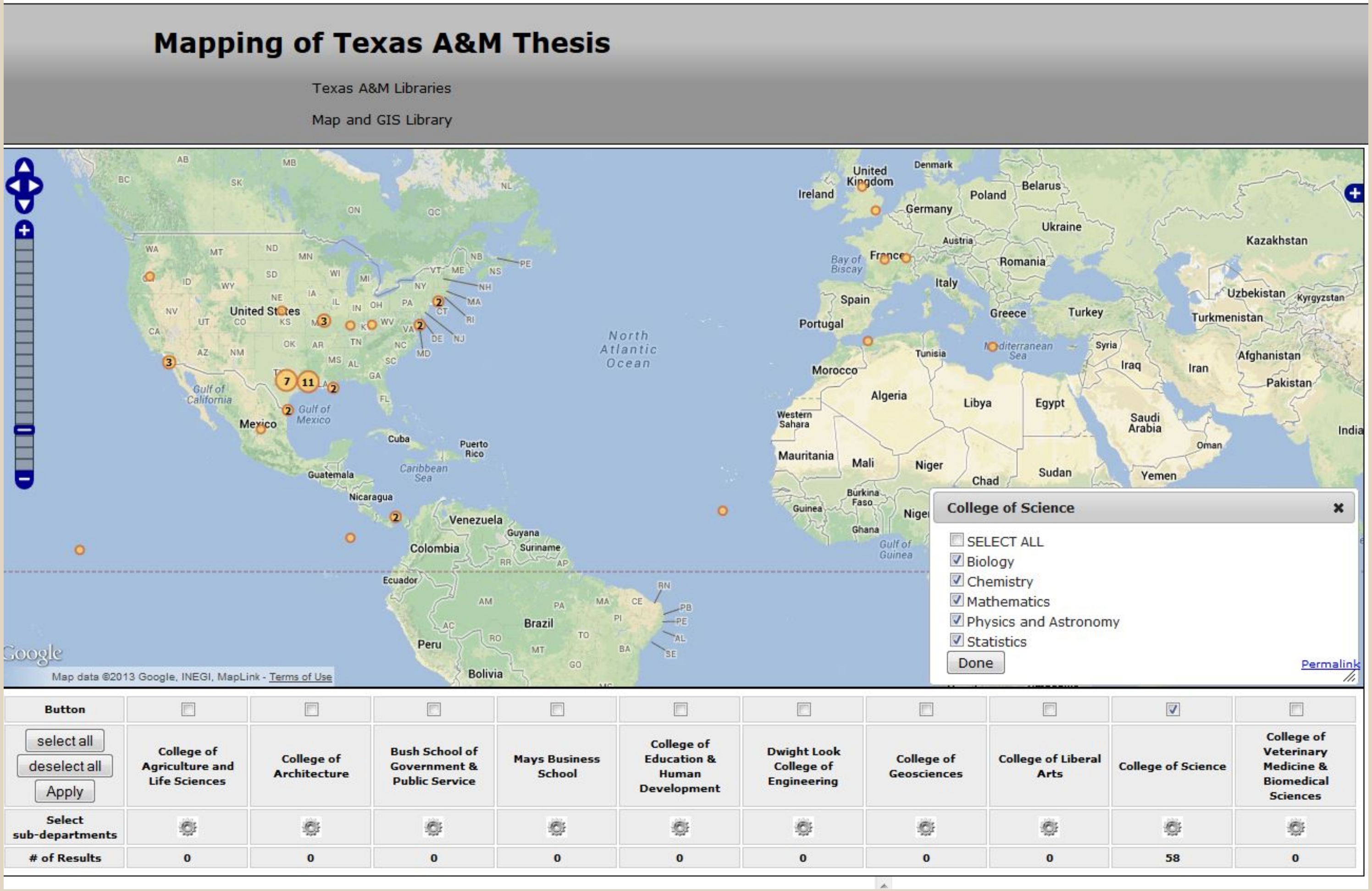
○ OpenLayers WMS

Overlays

☑ Thesis

Button	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
select all deselect all Apply	College of Agriculture and Life Sciences	College of Architecture	Bush School of Government & Public Service	Mays Business School	College of Education & Human Development	Dwight Look College of Engineering	College of Geosciences	College of Liberal Arts	College of Science	College of Veterinary Medicine & Biomedical Sciences
Select sub-departments	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙
# of Results	2	0	0	1	9	38	14	22	58	0

Map Prototype Department Filter



Mapping of Texas A&M Thesis

Texas A&M Libraries

Map and GIS Library

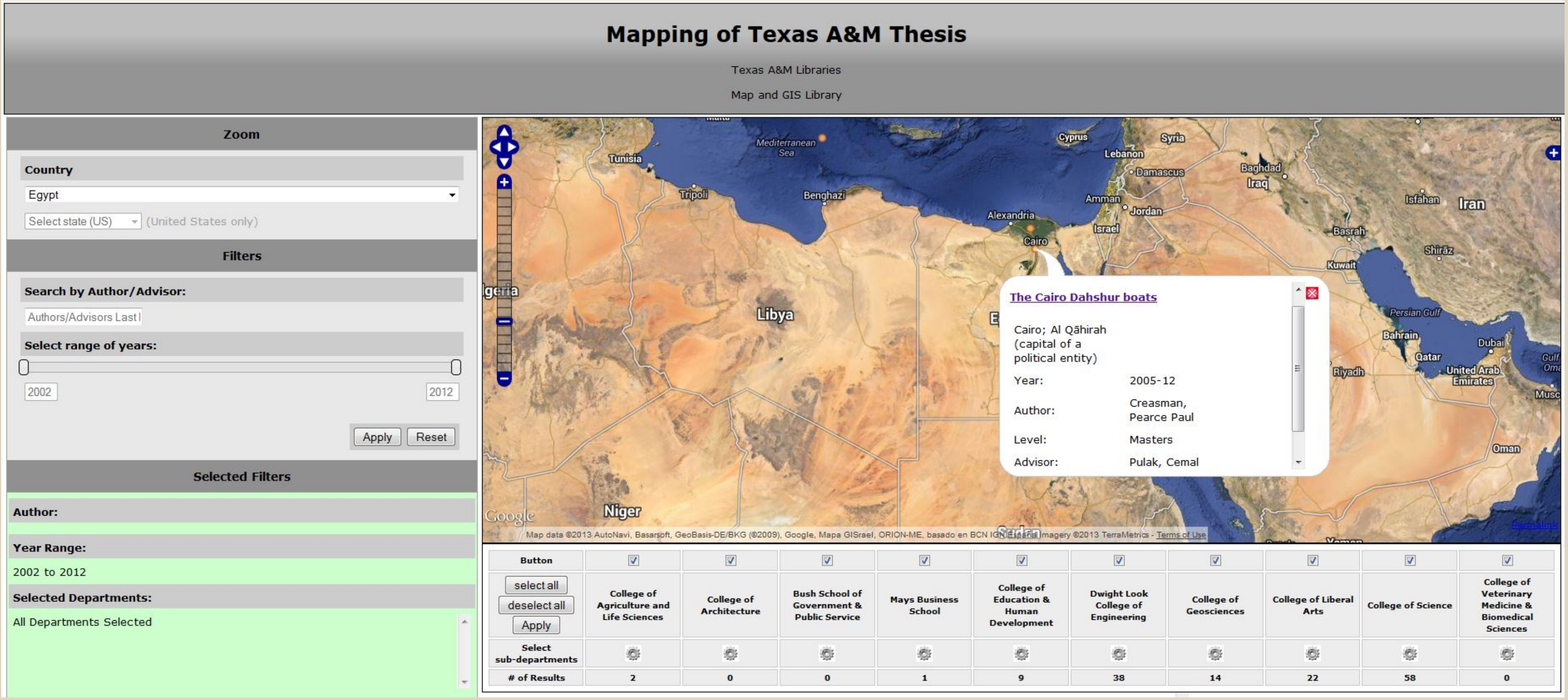
College of Science ✖

- ☐ SELECT ALL
- ☑ Biology
- ☑ Chemistry
- ☑ Mathematics
- ☑ Physics and Astronomy
- ☑ Statistics

Done

Button	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐
select all deselect all Apply	College of Agriculture and Life Sciences	College of Architecture	Bush School of Government & Public Service	Mays Business School	College of Education & Human Development	Dwight Look College of Engineering	College of Geosciences	College of Liberal Arts	College of Science	College of Veterinary Medicine & Biomedical Sciences
Select sub-departments	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙
# of Results	0	0	0	0	0	0	0	0	58	0

Map Prototype – Result Popup



Mapping of Texas A&M Thesis

Texas A&M Libraries

Map and GIS Library

Zoom

Country

Egypt

Select state (US) (United States only)

Filters

Search by Author/Advisor:

Authors/Advisors Last N

Select range of years:

2002 2012

Apply Reset

Selected Filters

Author:

Year Range:

2002 to 2012

Selected Departments:

All Departments Selected

The Cairo Dahshur boats

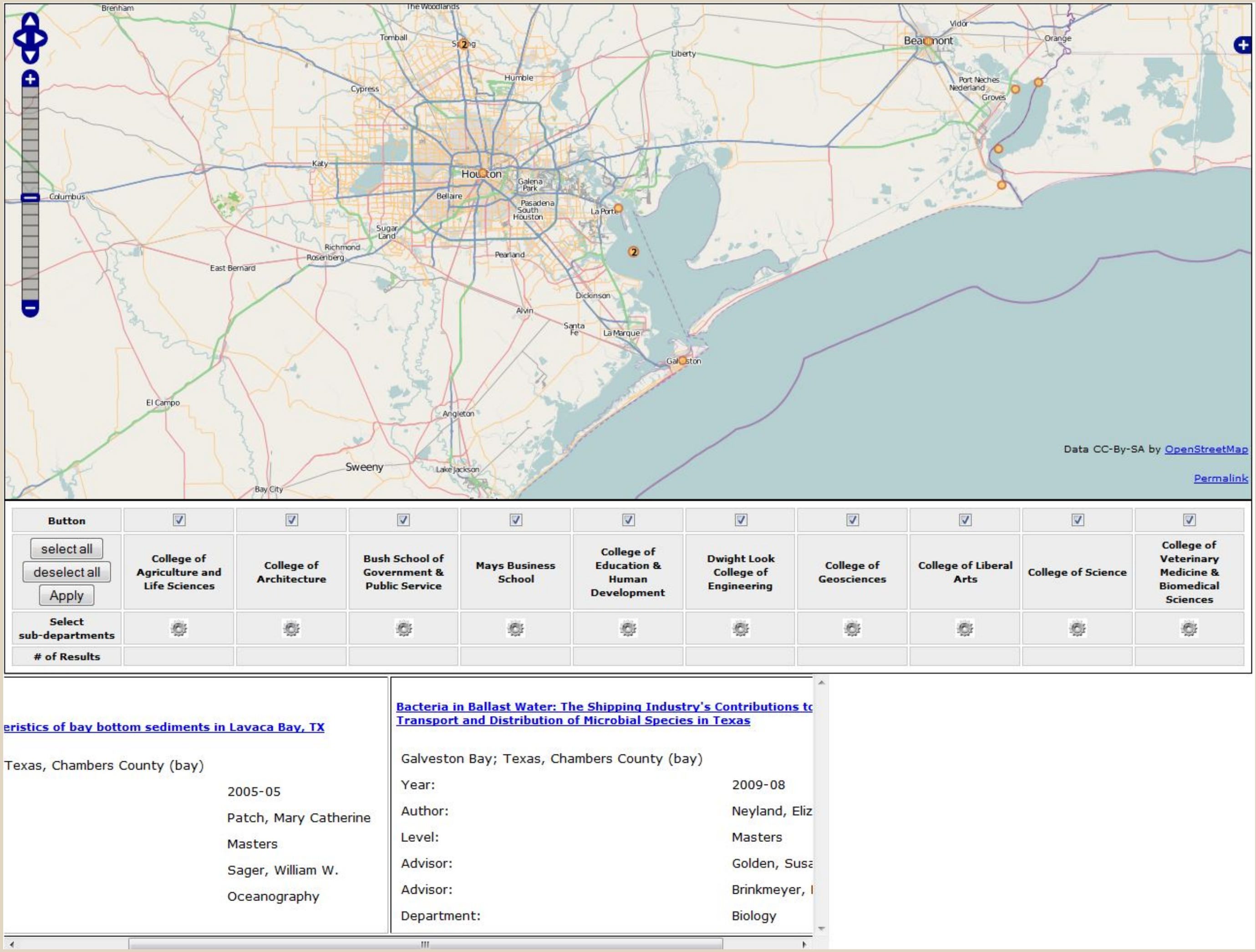
Cairo; Al Qāhirah (capital of a political entity)

Year:	2005-12
Author:	Creasman, Pearce Paul
Level:	Masters
Advisor:	Pulak, Cemal

Map data ©2013 AutoNavi, Basarsoft, GeoBasis-DE/BKG (©2009), Google, Mapa GISrael, ORION-ME, basado en BCN IGN España, Imagery ©2013 TerraMetrics - Terms of Use

Button	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
select all deselect all Apply	College of Agriculture and Life Sciences	College of Architecture	Bush School of Government & Public Service	Mays Business School	College of Education & Human Development	Dwight Look College of Engineering	College of Geosciences	College of Liberal Arts	College of Science	College of Veterinary Medicine & Biomedical Sciences
Select sub-departments	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙
# of Results	2	0	0	1	9	38	14	22	58	0

Zoom to location of interest



Button	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
select all deselect all Apply	College of Agriculture and Life Sciences	College of Architecture	Bush School of Government & Public Service	Mays Business School	College of Education & Human Development	Dwight Look College of Engineering	College of Geosciences	College of Liberal Arts	College of Science	College of Veterinary Medicine & Biomedical Sciences
Select sub-departments	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙	⚙
# of Results										

eristics of bay bottom sediments in Lavaca Bay, TX

Texas, Chambers County (bay)

	2005-05
	Patch, Mary Catherine
	Masters
	Sager, William W.
	Oceanography

Bacteria in Ballast Water: The Shipping Industry's Contributions tc Transport and Distribution of Microbial Species in Texas

Galveston Bay; Texas, Chambers County (bay)

Year:	2009-08
Author:	Neyland, Eliz
Level:	Masters
Advisor:	Golden, Susa
Advisor:	Brinkmeyer, I
Department:	Biology

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:

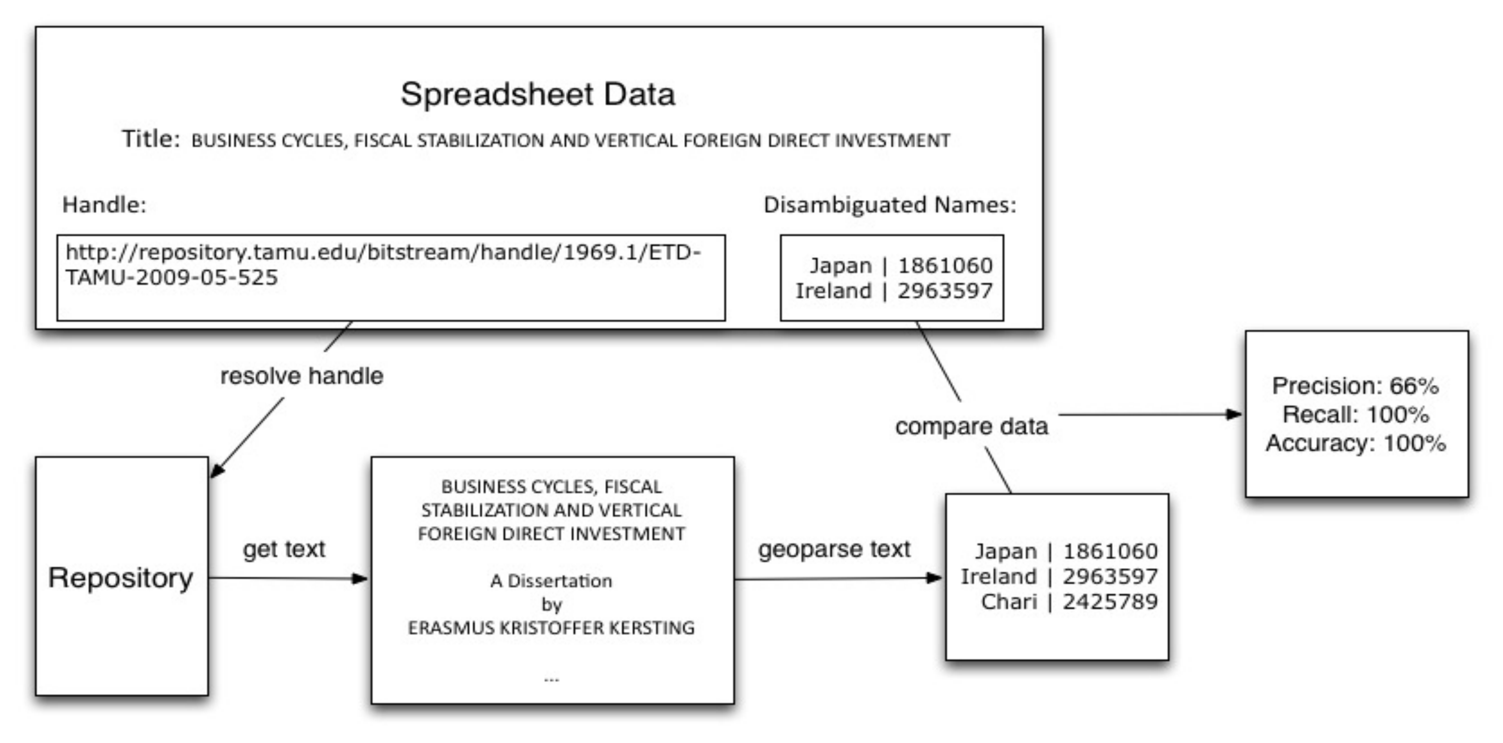
- Unambiguous extended names i.e. “Paris, France”
- Favor candidates of mentioned feature type
- Clustering of places (‘nearby locations’)
- Favor contained candidates

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



Spreadsheet Data
Title: BUSINESS CYCLES, FISCAL STABILIZATION AND VERTICAL FOREIGN DIRECT INVESTMENT
Handle:
http://repository.tamu.edu/bitstream/handle/1969.1/ETD-TAMU-2009-05-525
Disambiguated Names:
Japan | 1861060
Ireland | 2963597
resolve handle
Repository
get text
BUSINESS CYCLES, FISCAL STABILIZATION AND VERTICAL FOREIGN DIRECT INVESTMENT
A Dissertation
by
ERASMUS KRISTOFFER KERSTING
...
geoparse text
Japan | 1861060
Ireland | 2963597
Chari | 2425789
compare data
Precision: 66%
Recall: 100%
Accuracy: 100%

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 tem 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