FINAL REPORT

Creating a Map-Based Search Interface for Electronic Theses and Dissertations

Katherine H. Weimer
Texas A&M University Libraries

Amigos Fellowship
2010-2012

August 25, 2012

Background

A considerable number of theses and dissertations contain information about a location. The location of study may be a research site for anthropology, an analysis of the economics of a particular country, the sociology of an immigrant community in an urban area, or a range of topics in geology, biology or many other fields of study. The Texas Digital Library began collecting electronic theses and dissertations in 2004. Since then, the collection has grown to many thousands from the 18 TDL member libraries and is searchable using the standard DSpace interface.

If theses and dissertations with topical locations are brought together using a map based search interface, researchers from around the state would be able to see visually, others who are doing research concerning the same or nearby locations within the state and around the globe. This will increase potential for collaboration among researchers and graduate students as well as increase understanding of the locations being studied. GoogleMaps and open source mapping software provide a web based search interface that is relatively easy to geocode and program for a wide variety of uses. They do not require a plug in or special software for users. The goal of this project is the creation of an open source applications that will map and display electronic theses and dissertations and that can then be applied to TDL and ultimately larger collections.

Objectives

As stated in the project proposal, the objective of this project is to create a map based search interface for Texas A&M University theses and dissertations, with the long-term goal to share the code widely so that other states can also make use of the results.

GoogleMaps and other applications have increased the amount of information that is collocated geographically. We see this quite often with businesses (as in finding the nearest brand-name store) and Web 2.0 content including Flickr and Twitter. This geographic presentation of information has significant potential for organizing library content as well. Individuals tend to classify and sort visually, and maps provide an intuitive context for intellectual output of many types. To take this innovation to the next step, that is, to experiment using a unique collection of intellectual output, in this case theses and dissertations, has huge potential. These items are born-digital and are by-and-large free from copyright. As the information is found, a link can be provided to direct the searcher to the actual item of interest. Researchers can then browse a map of their area of interest and learn who is studying the location as well as other topics studied at that location.

An early stage design of the project was created by entering locations from a sample of theses and dissertations from Texas A&M University from 2005. The project was conducted entirely manually using human labor to read each abstract and to code each location in KML for GoogleMaps. Manually, the project is quite time intensive. The goal of the project is to automate the process, from searching for place names in the text to geocoding the places in KML or other encodings for display on a map.

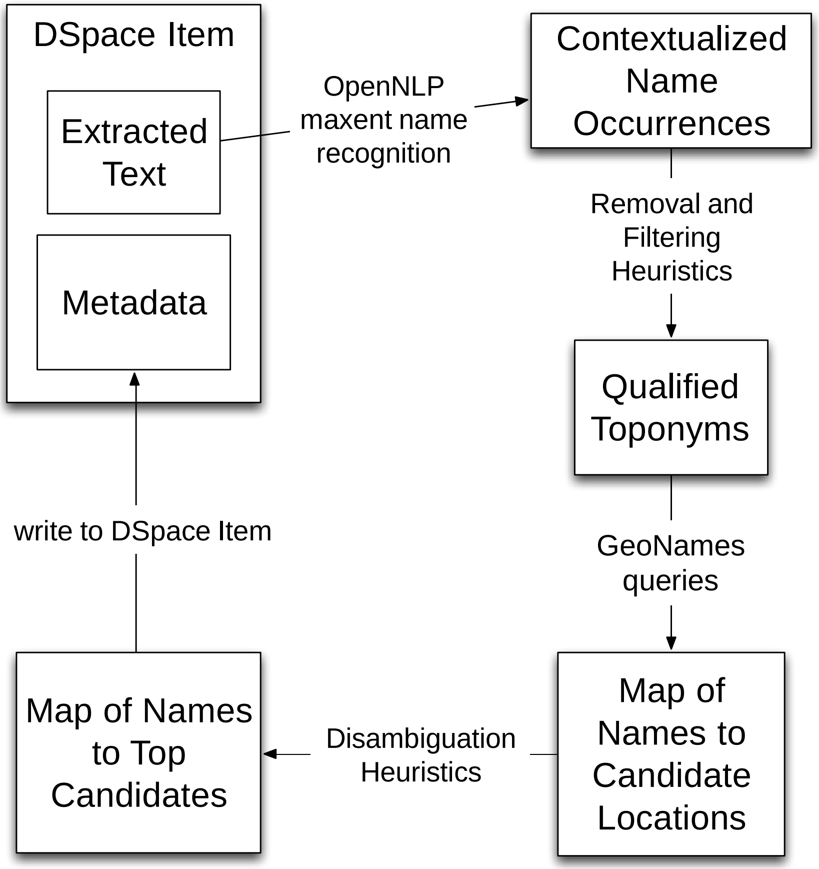
Early Stage of the Project

In the first year, the project progressed slowly. One student worker with computer programming skills was hired in the fall of 2010, however, unfortunately, the student made only moderate progress, then decided after just a few months to discontinue the project. In the meantime however, the researcher was pursuing the literature review and building collaborations and broader support within the library that proved to be quite fruitful in the second year of the grant. During the summer of 2011, a new direction began for the project due to the support of the library's Digital Initiatives manager's allotting a portion of a one programmer's time on this project. With the addition of James Creel, Sr. Software Developer's technical support, the project then got off the ground.

Project Design and Major Functions

The first major step was geoparsing the text of each thesis and dissertation using Java as a curation task in DSpace. This allows for recognition of text as possible toponyms (geographic names) with OpenNLP maxent name recognition. Certain parts of the text were ignored, such as the bibliographies, acknowledgements, appendices, etc. in order to avoid displaying in the final map locations that were not meaningful to the content. The toponyms were then filtered and sent as queries through the Geonames digital gazetteer. The Geonames server results are then disambiguated using a number of geoparsing heuristics, and the top-scored results along with the geographic coordinates furnished through Geonames are used to create geospatial metadata (Dublin Core element *dc.coverage.spatial*). The metadata are then output to a KML file for simple map creation. Three maps viewers were employed: GoogleMaps, OpenLayers and OpenWMS. Each map interface has pros and cons and further evaluation is ongoing. Upon viewing the map and zooming in to a point of interest, the user can click on the balloon which represents a text, then can click on the title, linked to the url of the document, and be routed to the actual pdf of the text. A diagram of the workflow is seen in Figure 1. The term 'map' used in the Figure 1 refers to a data structure map, not a geospatial visualization.

Figure 1. Geoparsing Workflow
(image credit: James Creel)


DSpace Item
Extracted Text
Metadata
OpenNLP maxent name recognition
Contextualized Name Occurrences
Removal and Filtering Heuristics
Qualified Toponyms
GeoNames queries
Map of Names to Candidate Locations
Disambiguation Heuristics
Map of Names to Top Candidates
write to DSpace Item

Discussion

Much thought and evaluation was put into the selection of the digital gazetteer. The researcher was familiar with the federal database, GeoNames, published by the U.S. Board on Geographic Names. The BGN maintains uniform geographic names used throughout the federal government and is a reliable and authoritative source. The National Geospatial-Intelligence Agency provides authoritative names for foreign place names. The researcher initially pursued these resources as the base gazetteer to work with. However, upon further reading and exploration, GeoNames.org was found to be a superior resource due to its inclusion of the two databases just mentioned, in addition to official gazetteers from countries around the world, and is easily and freely downloadable.

The metadata created through the KML curation task are used not only to create the points of interest for the map viewer, but to also serve as the database for the search filters. As preliminary map viewers were created and reviewed, more robust search features were seen as advantageous; thus, metadata fields were expanded in the KML to include not only the official place name and geographic coordinate for each location found in the text, but also the work's author, title, advisor, url, data and academic department. The latest version of the map includes a time slider search for date of publication, a keyword search for author or academic advisor name, and check boxes for academic college and / or department.

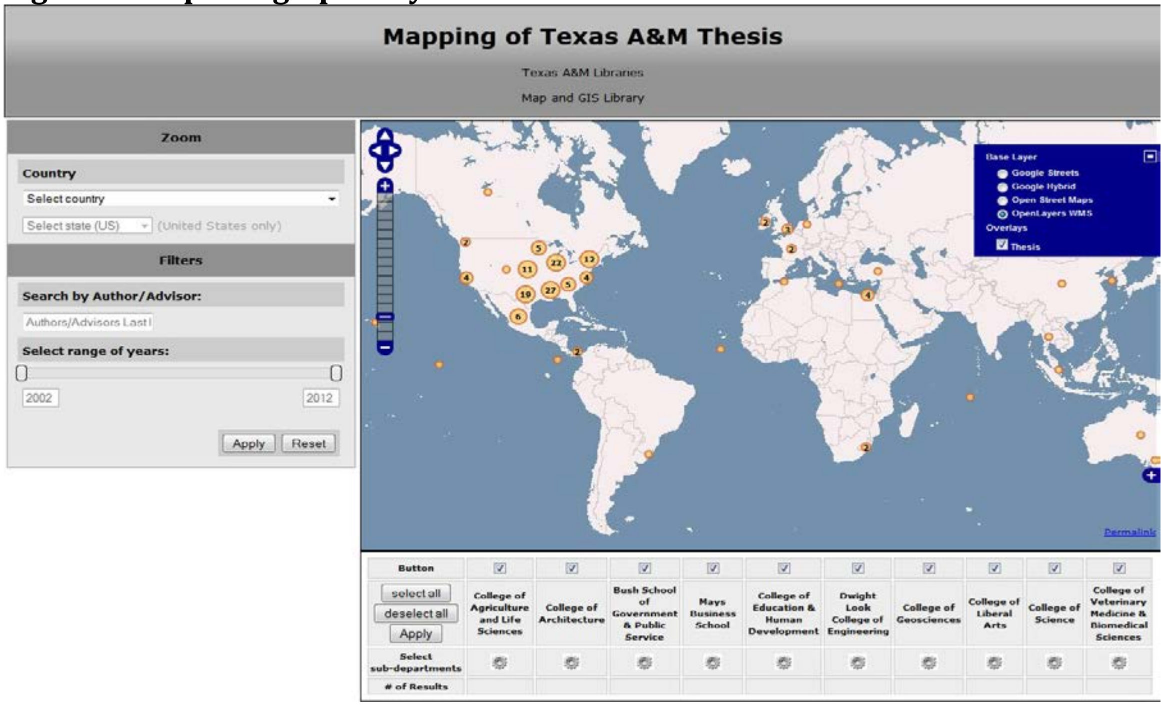
Place name disambiguation is the heaviest part of the project. While text recognition software can determine that the place named is "San Antonio," understanding that precise location is difficult since there are 477 results in the GeoNames (BGN) database alone which point to schools, churches, cemeteries, parks and assorted buildings across the US, as well as hundreds other places named "San Antonio" around the world. The heuristics involved in the place name disambiguation are quite complex, and involve statistical routines to predict relevance of the results. These tasks are still being refined.

The project was conducted using a sample of 29 documents from a variety of years and academic majors. The sample gave the researcher enough preliminary results which led to adjustments in the map design, metadata curation and map search filters. Due to the high volume of web traffic required by the queries to Geonames.org, the metadata curation task was time consuming, so, a larger sample was not readily available during the time frame of this research project. The expansion of the number and size of collection to be mapped will be addressed in future developments.

The Map Viewer

Grant funds were used to hire student workers to create the map interface. The initial map background was OpenLayers WMS. It provides a simple and easy to use interface and is open source, but did not provide great detail when zoomed in. (See Figure 2.) Other map backgrounds being evaluated are GoogleMaps and Open Street Maps. Open Street maps is open source, but, includes names in the native language of the country, so is not completely user friendly. Google is not open source, but has a nicely displayed product. In our current version, the user has the choice to select which map background they would like displayed, by clicking on a button at the upper right hand portion of the map. As the project progresses, the maps will be further evaluated and user tested.

Figure 2. Map using OpenLayers WMS Interface



Mapping of Texas A&M Thesis

Texas A&M Libraries

Map and GIS Library

Zoom

Country

Select country

Select state (US) (United States only)

Filters

Search by Author/Advisor:

Authors/Advisors Last I

Select range of years:

2002 2012

Apply Reset

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										

The map display shows a world map with all locations represented by a button, with clustering when locations are repeated. Once zoomed in, the clustered results divide into smaller groupings. (See Figures 3A and 3B)

Figure 3A. Zoomed in View showing Cluster Areas Broken Down

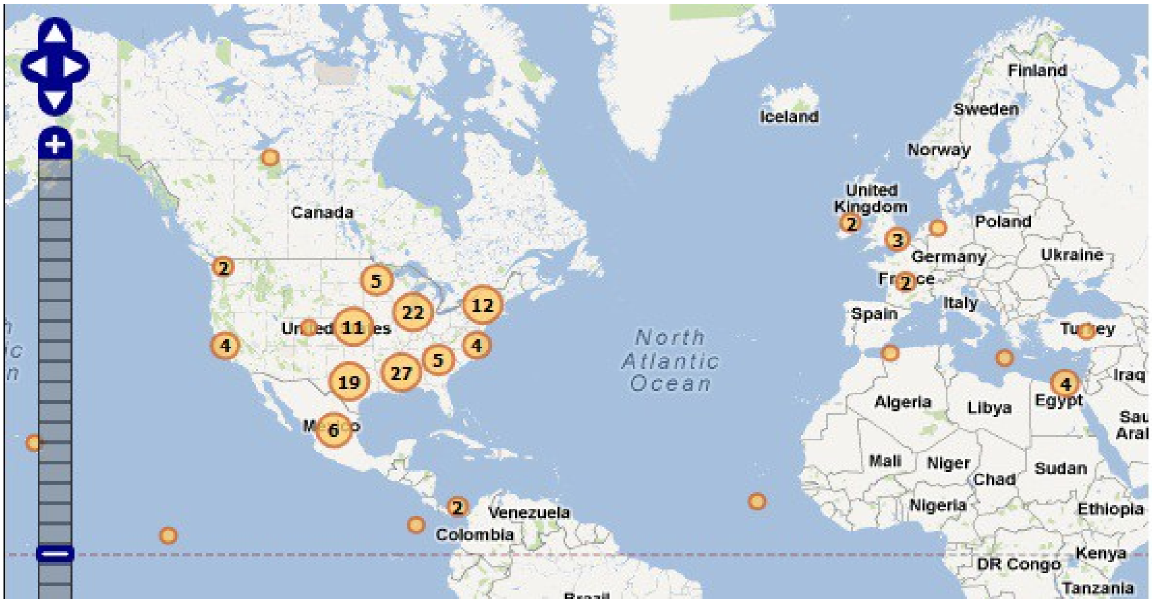
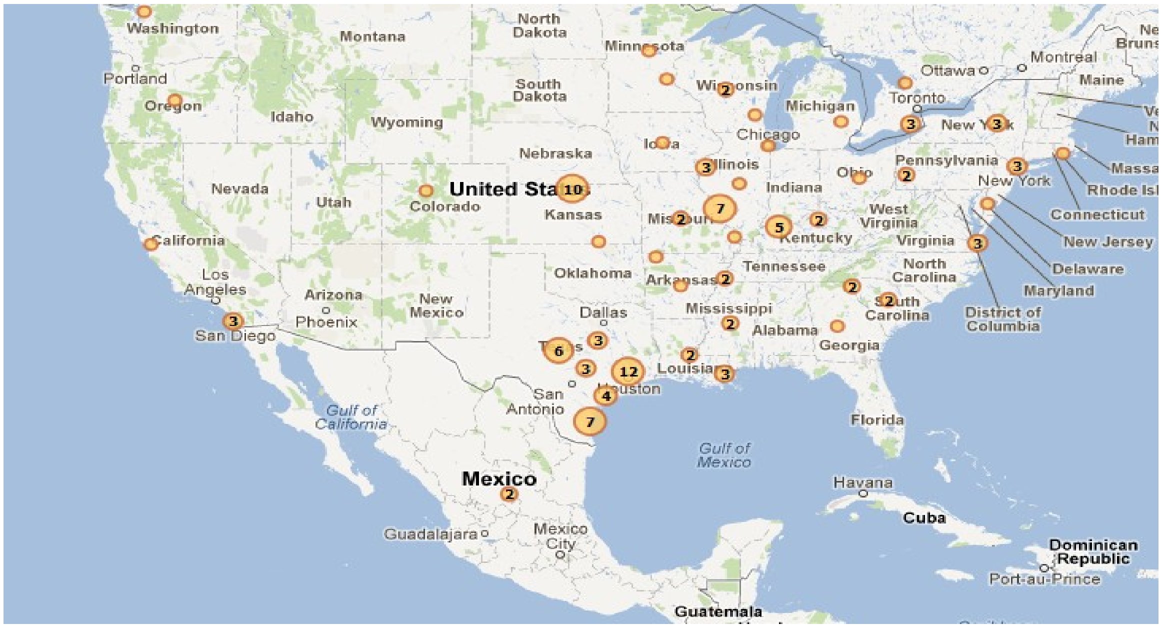
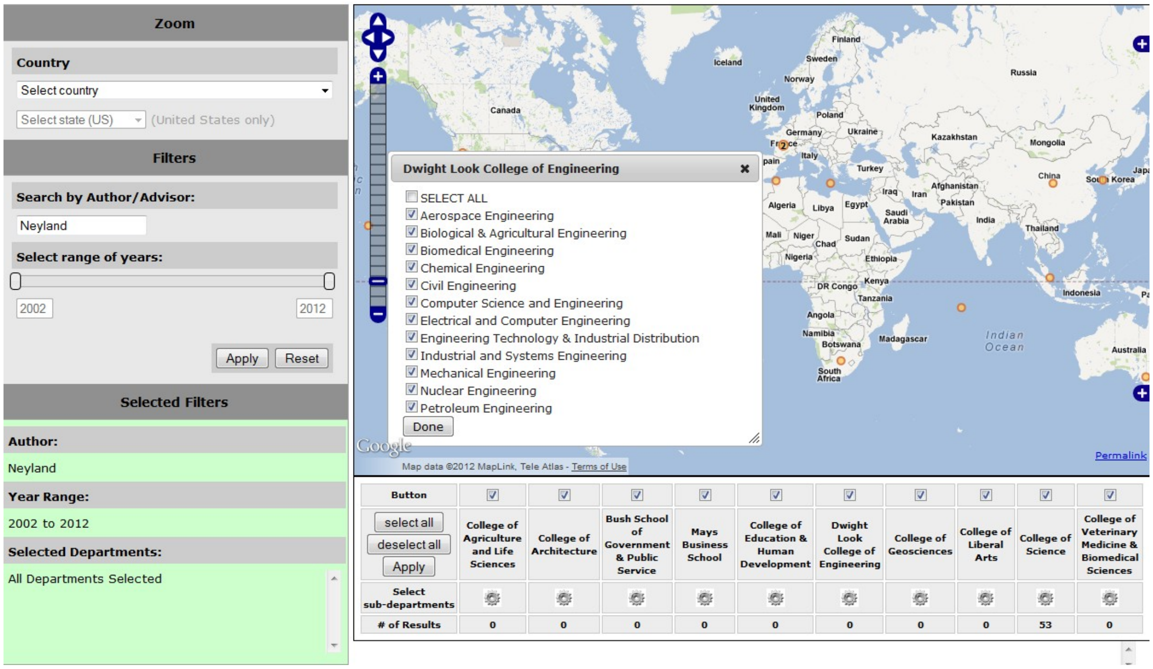


Figure 3B. The Same Map Zoomed in Further to Show Expansion of Cluster Area



Users may search for items, such as author name or date, using the fill box and time slider on the left hand side. Additionally, Texas A&M University colleges and departments are searchable using a series of check boxes located just below the map. Users may select any or all colleges and departments. For example, a search of the author by the name 'Neyland' and time period between 2002 and 2012 with all departments selected, results in the map display in Figure 4. When using the search filters, the selected filters are shown in the lower left hand box. Figure 5 displays a title of interest with complete metadata, with the title hyperlinked to the document.

Figure 4. Map Showing Author Search for Neyland between 2002 and 2012, with Department Selection Options Expanded



Zoom

Country

Select country

Select state (US) (United States only)

Filters

Search by Author/Advisor:

Neyland

Select range of years:

2002 2012

Apply Reset

Selected Filters

Author:

Neyland

Year Range:

2002 to 2012

Selected Departments:

All Departments Selected

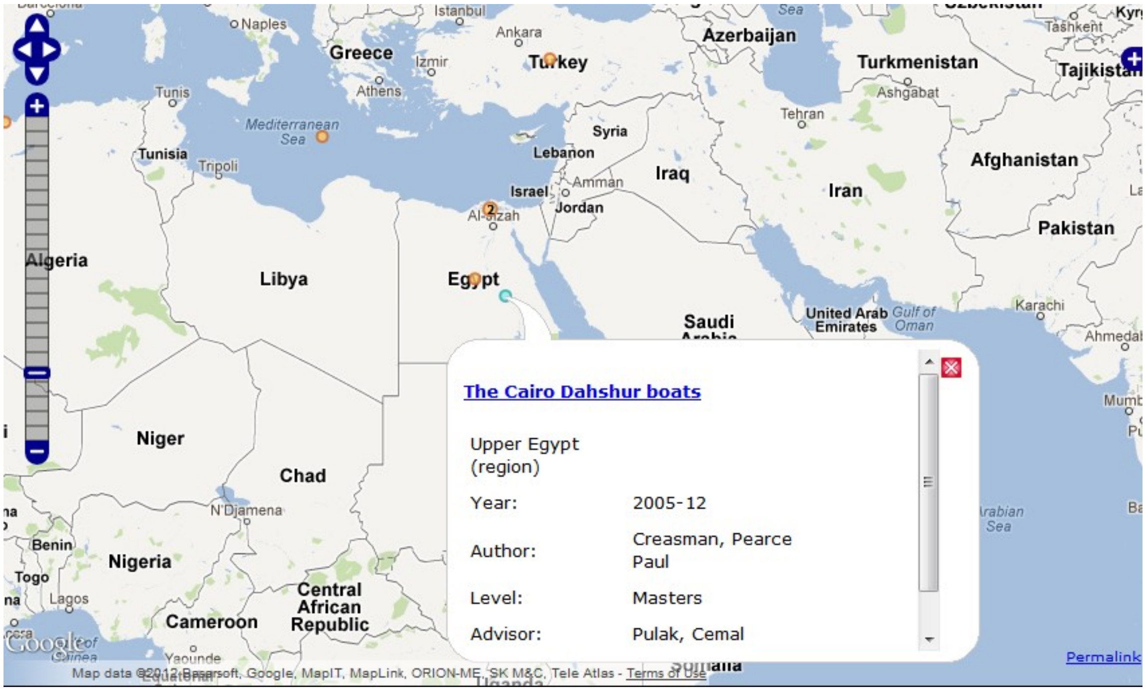
Dwight Look College of Engineering ×

- ☐ SELECT ALL
- ☑ Aerospace Engineering
- ☑ Biological & Agricultural Engineering
- ☑ Biomedical Engineering
- ☑ Chemical Engineering
- ☑ Civil Engineering
- ☑ Computer Science and Engineering
- ☑ Electrical and Computer Engineering
- ☑ Engineering Technology & Industrial Distribution
- ☑ Industrial and Systems Engineering
- ☑ Mechanical Engineering
- ☑ Nuclear Engineering
- ☑ Petroleum Engineering

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	53	0

Figure 5. Selected Item Showing Title and Metadata Linked to the Full Text



The Cairo Dahshur boats

Upper Egypt (region)

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

Dissemination of Information

The researcher gave a preliminary report to the library in February 2012. The project was well received locally. Later that month a formal presentation, supported by the Amigos funding award, was made at the Association of American Geographer's national conference in New York City. "Spatial Representation and Visualization of Theses and Dissertations: A Web-based Mapping Tool for Graduate Scholarship," was presented by Weimer, with Creel listed as co-presenter, as part of the AAG session, "Online geospatial visualization: concepts, developments, tools, frameworks and related topics." The venue was valuable due to the feedback given by attendees on the mapping aspects of the project.

The project was presented a second time over the summer, this time by the research collaborator, James Creel. While not funded by Amigos, it was credited for support of the project. Creel, with Weimer listed as co-presenter, gave a poster presentation at the Joint Conference on Digital Libraries, held in June 2012 in Washington, DC, on "Toponym Extraction and Resolution in a Digital Library." JCDL is a very competitive venue, and valuable feedback was gathered at this event also.

Quality, Evaluation and Future Plans

The major factor in the usefulness of this product is the quality of the content. To that end, the researcher has begun a quality control task to compare the automated output to the same sample done manually. A sample of theses was read manually and locations were recorded on a spreadsheet. Those locations were then looked up manually in Geonames.org and coordinates were recorded. This will allow for a comparison between a manual record and the automated output. The final steps in this assessment have not been completed yet, however, during the manual process a number questions were raised which have led to refinement of the disambiguation heuristics. Future work on quality control, particularly comparing a manual look up method to the automated results is desired and will be pursued in the future.

An additional factor in the quality of the map results have to do with the text selected to be mapped. While locations are mentioned through many texts, those mentioned only once or in an obscure way may be less relevant to be mapped than those locations frequently mentioned, or mentioned in a more meaningful portion of the text, such as in the abstract. Further analysis will be conducted to assess the parts of text that should be included in the map display.

Preliminary discussions have begun with staff at the Texas Digital Library and indicate an interest in sharing the project and implementing it statewide.