Digital Humanities 2013
University of Nebraska-Lincoln, 16-19 July 2013

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

Katherine H. Weimer, Texas A&M University Libraries
James Creel, Texas A&M University Libraries
Naga Raghuveer Modala, Texas A&M University, Dept. of Biological and Agricultural Engineering
Rohit Garagate, Texas A&M University, Dept. of Computer Science

Topic(s):
- geospatial analysis
- interfaces and technology
- metadata
- natural language processing
- maps and mapping
- data mining/ text mining

Keyword(s):
- geoparsing
- mapping
- digital collections
- visual browsing

While texts contain extensive mentions of locations, traditional library catalogs are lacking when searching for geographic locations. Ahlers and Boll (2010) state that approximately twenty percent of web queries have a geographic relation. Buckland (2007), Hill (2006) and others have emphasized the need for collections to be searchable using geographic means. With the advent of visualization tools and web based mapping, there are numerous possibilities to gain insight into the geographic content of texts. Gregory, and Hardie (2011), Bodenhamer, Corrigan and Harris (2010), Dear, Ketchum, Luria and Richardson (2011) and others discuss the role of geographic information in the humanities. Texts, maps and photographs have been described using map interfaces; however, the grey literature of graduate scholarship output has not thus far been presented graphically. Researchers at Texas A&M University Libraries are taking theses and dissertations, geoparsing those texts, and creating a visual, map-based search interface in order to glean better understanding of the locations and topics presented in these scholarly works.

The ETDMap is a prototype which automatically discerns the places mentioned in digital documents (i.e. geoparsing) and through a series of automated steps creates a map to browse the collection. Researchers gained conceptually from work outlined by Grover, et al (2010) and Leidner (2007). This geoparser operates in the

context of a DSpace institutional repository. Development has focused on the electronic theses and dissertations collection, although the software is applicable to any textual content in the repository. This abstract will present the current status and overview of the geoparser and map search interface tool.

The geoparser is implemented as a curation task in the DSpace repository, using the Java programming language. The geoparser automatically parses the text document, dividing it into sections and identifying prospective toponyms. The geoparser excludes certain portions of the document, such as bibliographies and appendices, since the toponyms found in those sections are typically not directly related to the subject matter of the text. The toponyms are filtered according to several heuristic criteria, and then are used to construct queries through the GeoNames Web-based Java API. The GeoNames server returns a set of locations that match the queries. The geoparser then applies disambiguation heuristics to score the locations and determine the most likely referent of the toponym. Finally, the top-scored locations, along with geospatial metadata (including coordinates) are written to metadata fields on the item under consideration. These metadata are then output to a KML file for viewing in a variety of interfaces. The term 'map' used in Figure 1 refers to a data structure map, not a geospatial visualization.

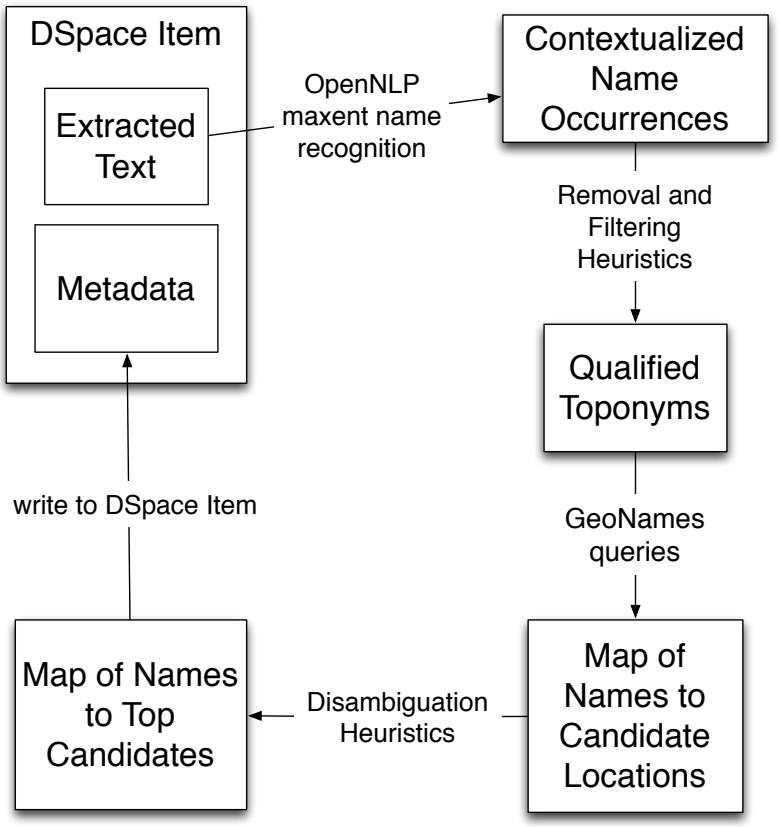

Figure 1.
Geoparsing Workflow [image credit: James Creel]

The geoparser uses regular expressions to partition the document text into sections. Theses and dissertations follow a predictable and regulated document format, which allows for clean results. Currently, the sections of interest are the abstract and main document body. These are processed by the geoparser, while the references, vita and appendices are ignored. The geoparser identifies toponyms in several stages using the OpenNLP maximum entropy software library[i] The ETDmap utilizes training data available on the OpenNLP Models page. [ii] The detected potential toponyms are stored along with contextual information, such as the number of occurrences and the locations of those occurrences, all of which is used in subsequent heuristic processing.

The locations referred to are discerned by a variety of heuristics. The primary Java entities used in the process are the CandidateMapper, the RefinementImpl (Implementation) and DisambiguationHeuristic. Pruning heuristics, which eliminate spurious prospective toponyms, are being implemented, and are under review and refinement. Those include heuristics that ignore short or common words, ignore singe occurrences, and require exact matches to records found in GeoNames. Additionally, scoring heuristics add points to scores associated with the possible toponyms. Populated places receive higher scores, as do those closer to other candidate locations. Once the stock of heuristics has been exhausted, the candidates with the top scores are selected as referents of the toponyms.

The Generate KML curation task reads the geospatial metadata thus generated (including geospatial coordinates provided by the GeoNames server) and encodes it in a KML file attached to the item in DSpace. This KML file includes placemarks for each of the mentioned places and includes description on each placemark with the title, author, advisor, url, date and department. At the collection level, the repository supplies a link to a KML file generated on request that consists of the aggregation of all the KML files generated for items in the collection.

GeoNames.org was selected as the gazetteer for the project due to its inclusion of numerous official gazetteers from countries around the world, and because it is easily and freely downloadable, so therefore practical for use in this case. In terms of visualization, 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 included 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.

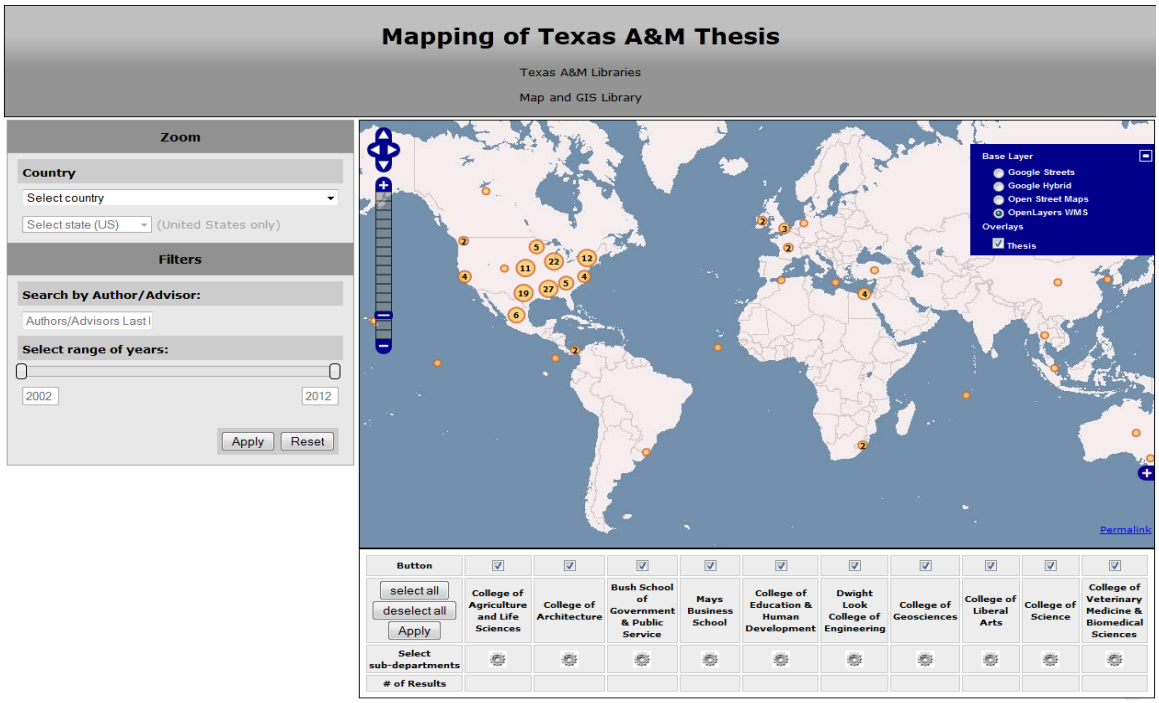

Figure 2.
Map Using OpenLayers WMS Interface

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. During the early development stage, metadata fields were expanded in the KML to include official place names and geographic coordinates for each location mentioned in the text as well as 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. Clustering of results enables the user, once zoomed into their areas of interest, to further refine the browse as the results are then broken into smaller groupings (Figure 3).

Figure 3.
Zoomed in View, Showing Expansion of Clusters

Once the user has pinpointed a location of interest, they may click on the pointer and be forwarded to the full text document located in the university's institutional repository, DSpace. (Figure 4)

Figure 4.
Selected Item Showing Title and Metadata Linked to the Full Text

Future Research

Research continues on refinement and development of the geoparser. Two short-term goals figure prominently in current efforts: an evaluation of the tool, and implementation of a statistical classifier as an augmentation to heuristic-based geoparsing.

Evaluation of the tool presents complications for the traditional precision/recall metrics of information retrieval. While these metrics are easily applicable to the disambiguation task, their application to the name extraction task is less straightforward. The mere occurrence of a toponym in a text does not indicate its relevance to the subject matter. We recognize and deal with certain negative cases by ignoring particular document sections like vitae, references, and appendices, but passing mentions of places occur in body text as well. We plan to implement a statistical comparator for target documents and a set of pre-selected documents known to refer meaningfully to particular places (encyclopedia articles, for instance). The statistics used for the comparator will include term-vectors or other textual derivatives. Techniques gleaned from this development will likely find application in the disambiguation task as well.

We have prepared a set of manually identified and disambiguated toponyms for approximately 100 theses as a basis for our pending evaluation of the toponym disambiguation task. Evaluation of the toponym extraction task will require more subtlety, perhaps including a user study to assess human understanding of the relevance of particular place mentions to document subject matter. Additional user studies will be conducted to enhance usability of the map interface. Finally, we plan to apply the mature application to collections beyond electronic theses and dissertations.

Funding

This research was supported in part by AMIGOS Library Services [Fellowship 2010-2012 to K. Weimer].

References

Ahlers, D. and S. Boll (2010) Location Based Web Search *in The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*, ed. by Scharl & Tochtermann (Springer).

Bodenhamer, D., J. Corriagn and T. Harris, eds. (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship* Bloomington: Indiana University Press.

Buckland, M., A. Chen, F. Gey, R. Larson, R. Mostern and V. Petras (2007). Geographic Search: Catalogs, Gazetteers, and Maps. *College & Research Libraries* 68:(5): 376-387.

Dear, M. J. Ketchum, S. Luria and D. Richardson, eds. (2011). *Geohumanities: Art, History, Text at the Edge of Place* New York: Routledge

Gregory, I. and A. Hardie (2011). Visual GISting: Bringing Together Corpus Liguistics and Geographical Information Systems *Literary & Linguistic Computing* 26(3):297-314.

Grover, C. et al. (2010) Use of the Edinburgh Geoparser for Georeferencing Digitized Historical Collections *Philosophical Transactions of the Royal Society A* 368:3875-3889.

Hill, L. L. (2006). *Georeferencing: The Geographic Associations of Information*. Cambridge, MA: MIT Press.

Leidner, J. (2007). *Toponym Resolution in Text* Doctoral Dissertation (University of Edinburgh) http://hdl.handle.net/1842/1849

Notes

[i] http://incubator.apache.org/opennlp/
[ii] http://opennlp.sourceforge.net/models-1.5