November, 2013

Enhance Library Catalog Searching with Geospatial Technology, Final Report

David Siegel
dave_siegel@harvard.edu

Marc McGee
mmcgee@fas.harvard.edu

Bonnie Burns
bburns@fas.harvard.edu

Summary of Accomplishments

Uncovering spatial relationships between catalog resources is an exciting effect of geocoding catalog records and exposes relationships within data that would not otherwise be obvious using "traditional" search methods that present information as long lists of text that the user must sort through. In Phase II we demonstrated how an interactive map interface can be used alongside of or even separate from traditional web catalog search tools that shows results in one-dimensional, paginated lists.

In the final phase of this project we completed several milestones to extend the work completed in Phase II. Probably the most significant of which was streamlining the geotagging process. That allowed us to geotag 1.1 million HOLLIS records and places them into three different open source-driven search indices. We created a solid foundation, infrastructure and model that are necessary to support large geotagging efforts with a goal towards shifting to open source geotagging tools once they mature. Also, adding so many new records provides a more representative data cross-section to demonstrate the benefits of geotagging catalogs.

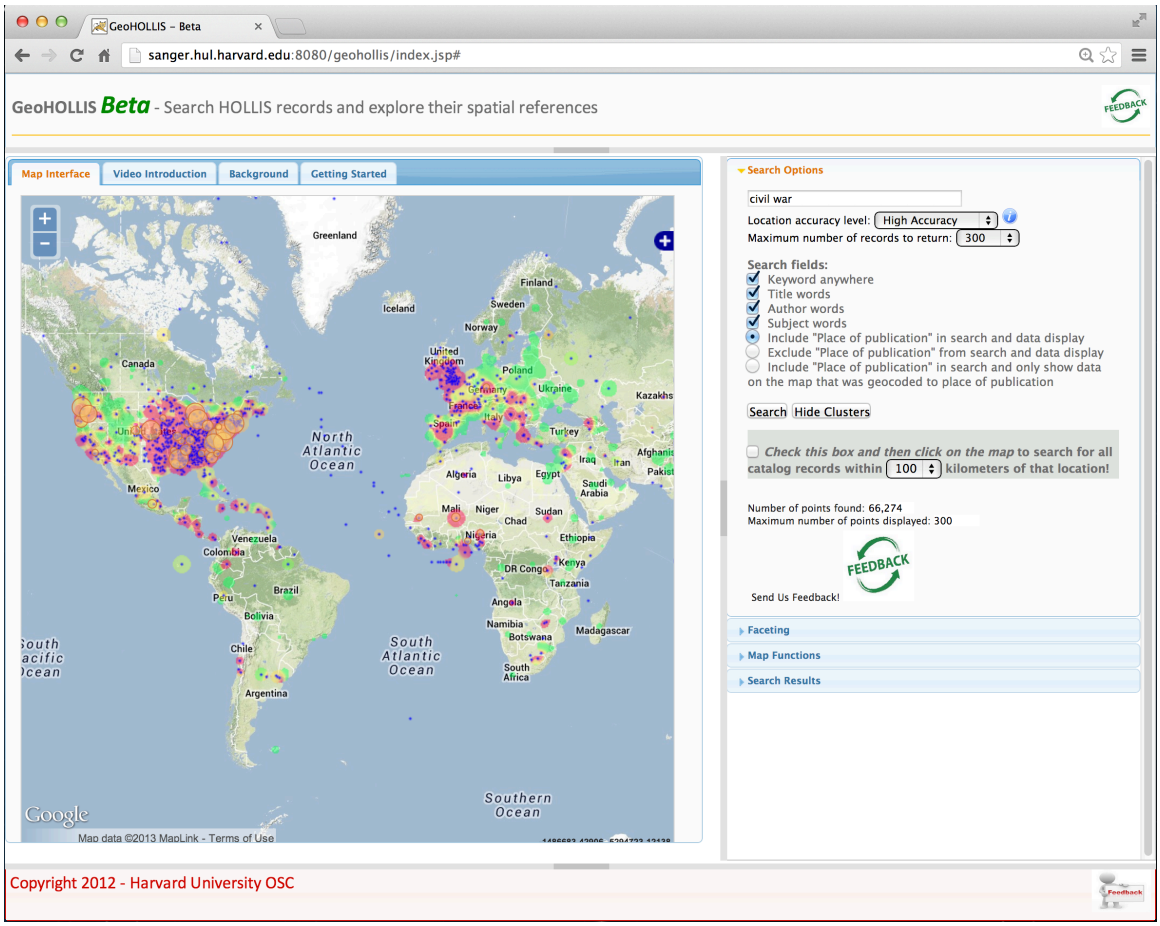
In the previous phase we created a visualization tool that presented one way that geospatial technology could be used to enhance the user experience of searching (HOLLIS) catalog data, called GeoHOLLIS Beta. Using a purely spatial search, we demonstrate how users can find catalog resources which are related spatially in ways that users didn't know existed. In this final phase, we increased the number of searchable records from 150,000 HOLLIS records to 1.1 million HOLLIS records. Our geotagging effort resulted in finding approximately 5 geospatial references per catalog record, resulting in approximately 5.8 million new records. The set of ~5.8 million records was added to the beta web map application GeoHOLLIS http://sanger.hul.harvard.edu/geohollis (hardware is not supported by Library infrastructure so the interface is not always available).

In this final phase we addressed the issue of server-side point clustering for presenting point data on maps. Server-side clustering is necessary in displaying large numbers of point records

because it groups the data and presents dots on the map that are sized based on the number of records they contain. This makes the map more readable because it is less cluttered. We developed two new models for delivering large volumes of point data to a map interface very quickly. We also included a third method that uses GPU processing to create map overlays. GPU's are graphics processors. Instead of having 12 CPU cores in a server the GPU gives us thousands, thus giving us the ability to process millions of data points and render maps for them in under a second. Server-side point clustering is an essential component in making effective web map displays. Before including GPUs to generate maps, GeoHOLLIS solely relied on client-side (browser-based) methods. Relying on client-side methods means that map display operations are dependent on the users' hardware and choice of browser and the limitations they impose (memory, network speed, CPU speed etc.). Included with these server-side methods is scale dependent point clustering. While there are projects that seek to display point data from catalogs on maps, they do not support a dynamic discovery platform like the one we've developed where clustered spatial results are presented in response to real-time queries. The closest we've seen is DPLA, but they have not geocoded all potential geographic references in the catalog data and the map presentation does not permit faceted browsing, clustering, linking, robust queries based on MARC field choice and the option for an open API to link disparate data sets.

Of the three methods mentioned, one Java solution was created and another that's specific to the data itself. In the third method this meant adding a geohash (http://geohash.org) to each coordinate pair where the software found text in a MARC record that had a possible spatial reference. This means that data grouping is done alongside query operations, thus meaning less data to transfer to the client for display on a map. The Java method capitalized on aggregation methods that are common in applications development, but were tuned to be spatially aware.

Heat maps, or density maps show areas where resources are available. Typically, colors are used to indicate hot spots where more resources exist. This is similar to a population density map except that it changes dynamically for each search performed. In the last phase we used MapD to provide this functionality for GeoHOLLIS (see screenshot below). MapD however, went from being a resource provided by an MIT developer to a private company. The set of records we coded in Phase II were supported in MapD and GeoHOLLIS for several months but the new set of ~5.8 million records were not ingested and provided as a service before the project ended. MapD is a partially open source product though and could prove extremely useful in future map visualization of catalog records. However, implementing the new MapD open source code was out of scope. The MapD software requires a server with at least 2 GPUs (graphics) cards. In the first phase a server and service were provided by MIT. In the future Harvard would need to acquire it's own hardware to run this. This is very cutting edge technology but it's something that we predict will be mainstream for GIS web applications in the near future.

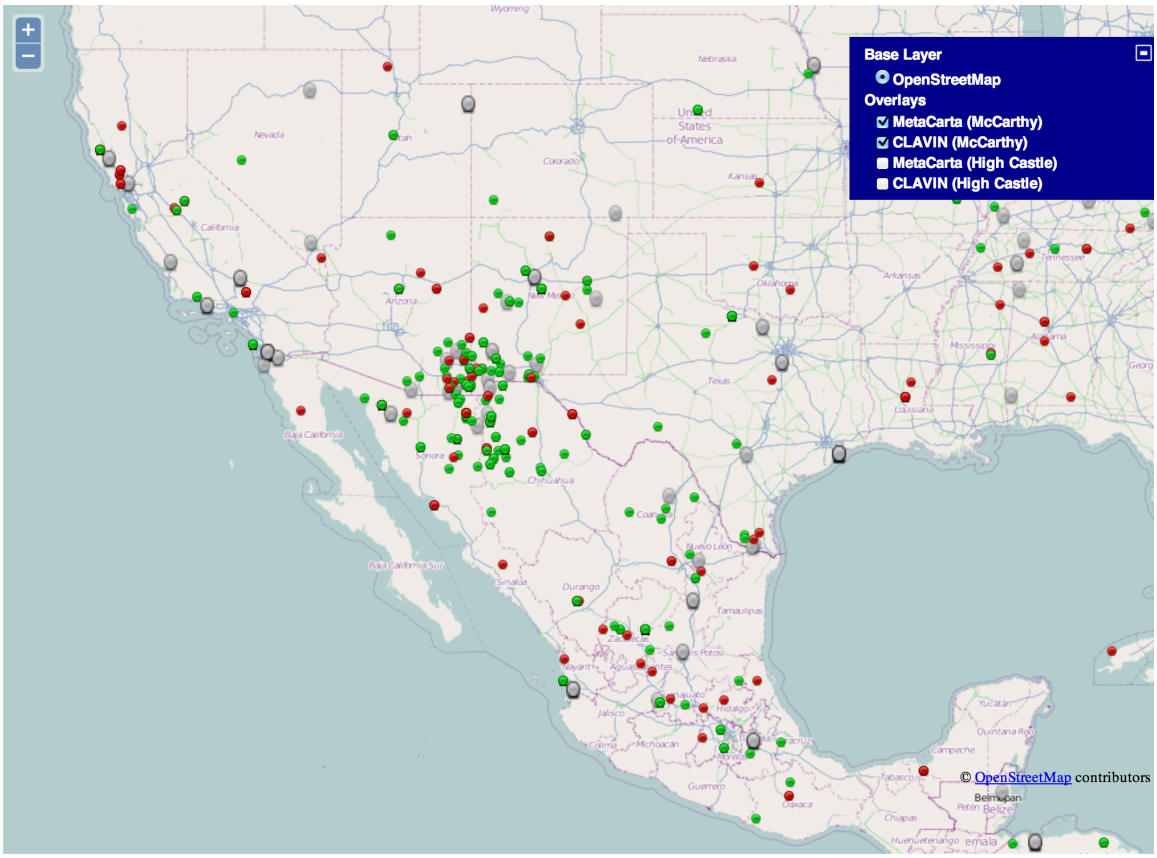


MapD in GeoHOLLIS with our clien-side point clustering as a map overlay

For the geotagging component of this project we used proprietary software from MetaCarta. There are open-source solutions evolving and one in particular, called CLAVIN shows the most promise to match the functionality of MetaCarta. We evaluated CLAVIN in this phase by running HOLLIS MARC records and free text from e-books through CLAVIN and MetaCarta and then plotting the results using map overlays as seen in the following image (grey dots represent CLAVIN's ability to pull geographic references, red and green are for the same text from MetaCarta).

When displayed together, the differences are quite clear, showing that MetaCarta located more geographic references from text. It also did so more accurately and with confidence levels that were more representative of actual results. We believe systems like CLAVIN will evolve rapidly and be ready for production-level applications within the next 3 to 5 years. Their main disadvantage is a smaller gazetteer (~8 million place names) than MetaCarta (~28 million place names). There are also issues with historic names that CLAVIN has not yet addressed.

CLAVIN also needs a more robust natural language processing engine. Textual references such as "15 miles east of Tucson" were easily plucked out of textual data by MetaCarta, but not by CLAVIN.



CLAVIN versus MetaCarta

In the last phase of the project we fine-tuned the applications schema that we used to store the georeferenced catalog records. We also altered our code base to support the new schema.

In the last weeks of the project we shared our methodology with other institutions that are interested in geotagging their catalog data and creating applications around the results. This includes, Stanford, Tufts, MIT and UNM.

Future Plans

1. We'll continue to coordinate with Library Cloud so that we can link the results of the geotagging to their API, making it more widely available.
2. We are preparing a demo for the Discovery Committee in hopes that they will recognize the value of geospatial search as a way to augment current text-based search applications.
3. We're hoping that the MapD group will ingest the newest set of geotagged HOLLIS records in their own interface. Their interface provides features that GeoHOLLIS does not, including a time slider to filter date fields and create time-based animations of catalog data.
4. Organizations such as DPLA are starting to create similar applications with geotagged catalog data. We want to see Harvard leverage its investment and finish the geotagging effort that was started. The set of geotagged Harvard data is more complete than the georeferenced DPLA data.