April 21, 2011

Enhance Catalog Searching with Geospatial Technology

Proposal to Library Lab – v 2
David Siegel
Geospatial Data and Information Software Engineer
Office for Information Systems
dave_siegel@harvard.edu

HOW THIS PROPOSAL DIFFERS FROM THE PREVIOUS VERSION

My previous proposal called for a two-phased approach to leveraging geocoded[1] catalog data to enhance searching library holdings. The proposal is now broken into three phases, with the first phase addressing questions from the review committee regarding who would use this, how they would use it and what they would get from it. The proposal now includes the request to use a pre-packaged set of data from Library Cloud for the initial phase. Depending on the results, the project would move into the next phase.

CONSIDERATIONS

The key idea of the original proposal was to demonstrate how to geocode existing catalog data in an automated manner without altering or augmenting existing catalog records in any way. Since editing all existing catalog records to hold geospatial information is cost prohibitive it's important to test the feasibility of automating the geocoding process. To a limited extent, we can determine if this is possible by working with pre-packaged data and we can demonstrate numerous use case scenarios. However, assuming that the geocoding process is deemed useful, determining if automated geocoding is feasible requires hitting live data and a broader effort. So, the proposal's original path to determining that is now captured in Phase II.

OVERVIEW

Increasingly, the way in which we go about our daily work is becoming more influenced by spatially-enabling technologies. Tools that facilitate georeferencing have gone mainstream and are no longer limited to registering scanned maps to ground truth, or having our cell phones broadcast locations so that we can view these data on Google Maps. Even though spatial technology is very widely used today, the ability to locate library resources and visualize this data in a geospatial context is lacking, especially in libraries. Technologies that scrub textual data and make associations by location are maturing, which means that the ways we incorporate the aspect of "place" into our catalog searches and the relevancy

we associate to it is becoming more important.

Transferring what we typically see as text to digital maps is a good start in visualizing spatial relationships. However, exploiting the full potential of such visualization cannot be realized until you apply more complex aspects of geospatial analysis to the data. Working with advanced techniques under the hood to match resources based on location context could launch the potential for data discovery beyond what we conceive as a catalog today.

Linking resources that are referenced in catalog data to accurate locations on maps is a difficult process, especially when using data that is **not** pre-encoded with such spatial tags as latitude/longitude or place names. Place names are routinely matched to gazetteer entries which make it possible to associate locations with objects. Applying tools for geotagging more abstract references by parsing text and matching spatial meaning to actual location is more difficult. However, outlining such a process not only provides a better understanding of the nature of your geospatial data but makes integration of library catalog data with spatial searching and pairing possible.

In the simplest cases, geoparsing matches the spatial attributes in catalog metadata (such as new location fields in VIA) to coordinates and places the results on a map. More robust scenarios involve uncovering opportunities for matching metadata titles (or other free text fields) to coordinates for placement on maps. These processes produce interesting results. However, the possibilities of applying such technology to catalog data can go much further.

In the most advanced scenario, complex geospatial analysis is applied to catalog data that has been enhanced with map coordinates to uncover new methods of resource discovery. This happens outside of a map display and pairs search criteria with the geospatial attributes of catalog data from disparate resources that might otherwise be obscure.

The project proposed here does not involve scrubbing text from book contents and representing the locations identified on a map. Rather, the intent of this proposal is to first apply geotagging concepts to a metadata-based mapping of catalog data, and then to incorporate geospatial analysis to aid in the discovery of resources.

PROPOSAL

This proposal is broken into three phases.

Phase I -- Geocode pre-packaged catalog data and expose use case scenarios

Phase I is an effort to demonstrate proof of concept. We will take pre-packaged data from Library Cloud and a GIS Specialist will geocode it. We will take those results, map them and demonstrate use case scenarios (via an interactive map and with mock-ups of screen shots) showing how the results of

geocoding the catalog data would enhance a user's catalog search experience. We would also determine the feasibility of automating the geocoding process for advancement to Phase II.

In this phase, a fixed dump of records matching the query "civil war" would be obtained from Library Cloud and a GIS Specialist would georeference the results. The georeferencing process would resolve the geographic meaning of the queries' metadata in such a way as to support rendering the objects on a map referenced by latitude and longitude. A user interface component will be constructed that would show a map (using the Google Maps API or OpenLayers) with icons depicting the geographic location of the geocoded catalog data. It would also allow the user to casually browse the catalog data within the context of a map, where resources are displayed as icons. Clicking the icons would display additional information specific to that resource.

Visualizing catalog data by placing it on a map is only one scenario of use where in most cases, users can immediately determine if the results are coincident with their area of interest. It is anticipated that with live data, users will be presented with catalog resources for areas they did not anticipate. Whether that smaller component of the overall proposal is helpful to users can only be determined by testing. At this point a user might want to use the map screen to restrict their search to a specific geographic extent. A new search could be driven by simply dragging a box around the area of interest and searching again. The map could be used as a base to refined search criteria. That capability however, would not be demonstrated until phase II where the Library Cloud's REST API would connect the user to real time queries that text filtering could be applied to.

It's envisioned that the mapping component of Phase I would be hosted at the Berkman Center. In this capacity it only serves as a test platform. However, the potential use of this could be a map embedded in almost any catalog thus making it accessible to a large user base. The data is accessed by the catalogs through API's supported by the content provider. In Phase I and Phase II the value added component is primarily a map to display location characteristics of catalog data. However, with Phase III additional contextual relationships within the catalogs are exposed using advanced geospatial analysis capabilities.

Phase I Deliverables:

1. Mock-ups of use cases scenarios including how maps might appear in existing catalogs to augment the discovery process. Also, how the results might appear in ShelfLife
2. A report containing a) Explanation of geocoding process and success rate, b) Determine if the catalog data is suitable for automated geocoding and c) suggestions for ways to increase successful match percentages
3. Recommendations for an automated process to geocode catalog data from Library Cloud
4. Cataloging recommendations for making records more spatially aware
5. Vetting process; Review findings with users and obtain additional ideas on how to interface the georeferenced data with catalog searches
6. A more definitive explanation of what's realistic for Phase II and how additional information

from the catalog is obtained (outlining a process to leverage Library Cloud's REST API for random text searches)

7. Library Cloud will receive a geocoded test set of records which they can expose to other developers and projects

Note on use case scenarios

Most use cases are not known because we have not geocoded and mapped catalog data and presented it to users for testing. The process of capturing the spatial nature of catalog data would evolve into many use cases. Most likely, numerous opportunities that are not yet known to us would become evident with a proof of concept. Numerous cases will surface from Phase I, but the research aspect of Phase II would uncover more use case scenarios and the full potential of implementing spatial analysis across the catalog data would be obtained in Phase III.

Note on how the map created in Phase I is used

The map from Phase I serves a proof of concept but also as an idea sandbox, where users can visualize how catalog data appears on an interactive map and share their ideas on how this would help them in searching Harvard resources. Moving forward to Phase II we would take the map from Phase I and use the Library Cloud REST API to query data. Phase I will assist us in determining if geocoding can be done on-the-fly or if the data would need pre-processing.

Phase II -- Automate geocoding of catalog data and integrate with Library Cloud's APIs for full data set and random searching

Phase II builds from the outcome of Phase I to automate the geocoding process of data residing in Library Cloud where search results for any query are placed on a map. It would also build on the number of possible methods of geocoding the catalog data to support more robust queries. For example, we might choose to map search results based on the publication city, or the subject, or the title. This would provide users with the ability to visualize search results in a dynamic way, and perhaps locate items more relevant to their interest, or learn new information about their search results.

Phase II still only scratches the surface of the potential of geocoding the catalog. Phase III would leverage the work from Phase I and Phase II to enable new types of resource discovery by, combining the geocoded catalog data with search queries to locate resources that meet specific geographic criteria. For example, provide all resources that are *spatially linked in the catalog* to the query; "Civil War".

In Phase II we integrate with Library Cloud's APIs full data set and random searching with automated geocoding. Phase II is partly a research activity that seeks to automate the process of geocoding data

from Library Cloud instead of using pre-packaged data.

In this phase new software would be written or existing applications (used in PhaseI) to parse metadata from Library Cloud (using its REST API) through an entity resolver and extract geographic locations. This would resolve the geographic meaning of the metadata in such a way as to support rendering the objects on a map referenced by latitude and longitude. The user interface component created in Phase I would be enhanced. The map will display icons depicting a resource description's probable geographic location. It would also allow the user to casually browse (not specifying any search criteria) catalog data in Library Cloud using a map, where resources are displayed as icons. Clicking the icons would display additional information specific to that resource from the metadata. Applying the geotagging techniques to the entire catalog and mapping the results would provide a high level view of the catalog contents in a spatial context. The most likely technology for communicating search results for the mapping user interface is GeoRSS[2].

Phase II Example;
A user could search the catalog for "Civil War". The search would hit the geospatial index, and extract the location reference from the metadata. It would augment the search results with a map display. The map would include markers on all locations where the metadata had a matchable place name. For example (based on queries already taken from Library Cloud), we would show an icon on Lebanon, and clicking the icon would reference the book "Lebanon in crisis, Lebanon History Civil War" or a marker in North Carolina would present; "Stand, Watie and the agony of the Cherokee Nation, United States History Civil War", or another marker in Nigeria; "The making of a nation, Nigeria History Civil War". The interface would also support filtering of catalog records based on location (from the map view) and keywords entered by the user and found in the metadata.

In Phase II, as another example of the possibilities of mashing up text with geospatial data we would also prototype embedding geospatial data within the ShelfLife UI's book page.

Phase II would also provide the foundation for applying more advanced geospatial operations to general catalog searches that would pair resources in ways not yet visualized. In addition to placing the results from geotagging on a map, the data gathered during the matching process itself would be stored in a database. In order to take the integration of geospatial attributes of catalog data to the next level, we would need to evaluate the spatial characteristics of such a database, which leads us to Phase III.

Phase III -- Leverage the use of GIS analysis methods for exploring georeferenced catalog data

In Phase III we would explore the opportunities for in-depth geospatial pairing of catalog records to support resource discovery by exploiting the spatial relationships of catalog records that we can uncover by georeferencing the data and applying geospatial analysis and filters to those records.

For Phase III we would develop a new catalog searching component implementing geospatial analysis. This would provide a platform for evaluating how library catalog searches are enhanced using geospatial technology. This would include exploring methods such as proximity analysis and spatial overlap. Phase III would also include outlining a process of spatially encoding catalog data to support browsing of spatially-tied data. The goal would be to present researchers with new ways of discovering resources.

Phase III Example;

Similar to Phase II, browsing a map (say, of Gettysburg, PA) would show results that match that location (books, artifacts, art and historic maps) such as titles by Chamberlain. However, the geographic analysis would also pair resources that matched aspects of that location and match to other records, such as Chamberlain at Little Round Top, and then cross reference it to other related locations, such as Appomattox or to related works from Chamberlain's time at Bowdoin.

The searching would now be able to present the user with a list of all catalog resources that in some way overlap with that location or are within a particular proximity. If placed on a map, filter options in the user interface would toggle icon displays based on user-selected criteria. If displayed as a list, results with highest scores would appear first. Also, using colors to indicate density (heat map) users would be able to see areas on a map that the catalog has the most information about.

While this would be a very limited amount of data to begin with, over time as the usefulness of such pairings became evident, more data would be tailored for such processes allowing for greater participation and advancements.

METRICS

Measuring the success of Phase I would require viewing the results of the geocoded catalog data on a map and determining the percentage of results that display "correctly" as compared with the total number of records from the catalog that did not provide a match. From there, overall success would be somewhat subjective. If what you saw was relevant and helpful when searching the catalog then you would move forward to Phase II. This is analogous to the research approach of ShelfLife. Additionally, the mock-ups of use scenarios will clearly outline benefits from geocoding the data.

Since I have not been able to determine if anything like Phase III has yet been done, success would probably be measured by whether or not the addition of geospatial analysis into the search engine would help researchers. Then it would be a matter of determining how to improve both the number and accuracy of the results, and soliciting new ideas for displaying the data.

COSTS -- **Phase I Only**

1 FTE GIS Specialist for 6 weeks

½ FTE software developer from Library Cloud for for 2 weeks
⅓ FTE Metadata Cataloger
Intermittent support from a Metadata Analyst
Miscellaneous consultation time from Systems Librarians
Acquisition of tools to perform geocoding. Not all are open source. An example is MetaCarta's Geo Referencing Engine. They might be willing to provide the software for research purposes at little or no cost.

[1] Relating information to geographic location
[2] GeoRSS, http://www.georss.org/Main_Page