November 16, 2012

Enhance Library Catalog Searching with Geospatial Technology, Phase II Report

David Siegel
dave_siegel@harvard.edu

Marc McGee
mmcgee@fas.harvard.edu

Bonnie Burns
bburns@fas.harvard.edu

Project Summary

The goal of Phase II was to build off the proof of concept done in our previous work and;

1. Create a functional discovery tool based on a map interface that allows dynamic search and visualization of catalog resources that were geocoded.
2. Start building an infrastructure necessary to support large geocoding efforts that would eventually capitalize on open source geocoding tools.

We created a visualization tool that presents one way that geospatial technology could be used to enhance the user experience of searching (HOLLIS) catalog data. Using a purely spatial search, we also demonstrate how users can find catalog resources which are spatially related in ways that they didn't know existed.

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.

The proof of concept phase demonstrated the potential uses of enhancing catalog searches by geocoding catalog records, and the research involved in Phase I included looking into possible workflows for the geocoding process. In Phase II we started building an infrastructure necessary to support a large geocoding effort so that when open source geocoding tools mature and make the technology available to the masses, entire catalogs could be ingested. Our goal was to begin building an infrastructure with a library focus, and to place the model, code and source data into the public domain along with a usable application.

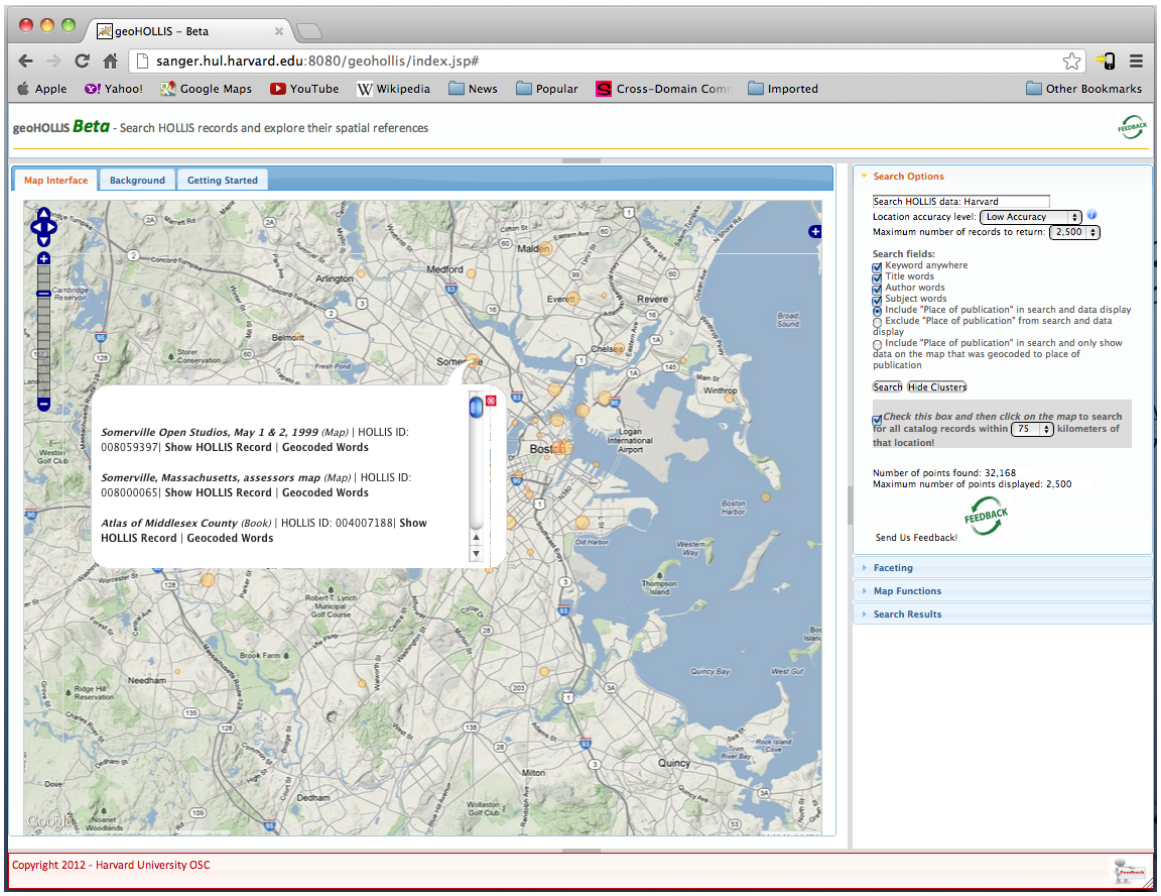
We have met these big-picture goals and while we are happy with the results we still hit unexpected technical and institutional challenges that altered our timeline significantly and kept us from creating a more polished and completed end product.

Accomplishments

We created a Beta user interface (http://sanger.hul.harvard.edu:8080/geohollis) where users can interract with a subset of geocoded HOLLIS catalog data in a mapping environment. We opened the Beta site to the entire web to allow anyone to try the system out and post feedback. * *The server holding the Beta site was hacked on Nov. 12 and LTS systems administrators are working to restore it. As of Nov. 16, the system is not restored and online yet.*

For this project, we define geocoding as the process of converting implicit location references in catalog data such as “Widener Library”, “Pavillion Beach” or "Parc de Versailles" into formal geographic locations expressed as longitude and latitude that we can use to plot locations on a map. We’ve taken catalog data, added these references to the records themselves and made them discoverable via a map interface using “traditional” search methods (textual search), spatial search (all data located near x,y) or a combination of both (“civil war” near Philadelphia). Like Bing, Yahoo or Google maps functionality that allows you to identify location based services by identifying a place of interest (tool rental near ZIP 02138) we have made catalog resources available for discovery based on location.

To create the user interface we geocoded approximately 150,000 HOLLIS records. From these records, we were able to extract approximately 790,000 geospatial references. Some HOLLIS records had no geographic references while others had several. These 790K geocoded references are available for search and display through the user interface. Several methods are provided for browsing the data.



http://sanger.hul.harvard.edu:8080/geohollis demonstrates a Beta interface into spatially browsing approximately 150,000 HOLLIS records that were geocoded.

Numerous tasks were involved in or to create the user interface. These included:

1. Developing a process to extract HOLLIS records from the Library Cloud API
2. Developing the methodology to take the data from Library Cloud and geocode the data
3. Create a Solr schema and index to hold these geocoded records
4. Tune a search engine to handle textual and geographic discovery. This also included creating the infrastructure necessary to ingest 790K records. This architecture is expandable to handle millions of records.
5. Create a user interface as a window to these data. This included developing spatial searching (records within a specified distance), faceted browsing and advanced point clustering to display thousands of points clearly on a map at one time.

Perhaps more important than placing a system online to spatially browse data, a more significant outcome was implementing tools in the web site for advanced discovery; An answer to the common problem of “You don’t know about what you do not know”. As we added more records to the system we started to realize goals we had outlined in our first proposal. Specifically, this is the spatial browse aspect of the user interface where users can click on the map and see all catalog results that have a spatial reference within a specific distance of their

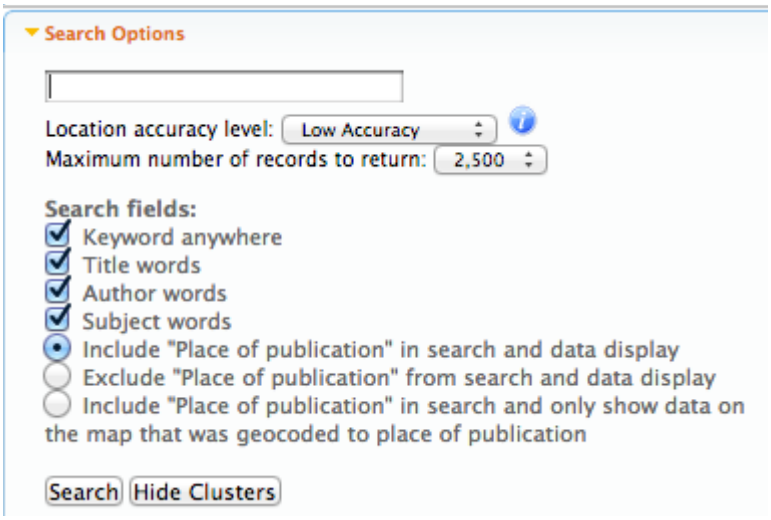
place of interest. Since most HOLLIS records have multiple spatial references, exposing data that's related by location becomes extremely easy.

As previously mentioned, the Beta implementation of GeoHOLLIS contains about 150,000 catalog records of the over 12 million in the complete HOLLIS catalog. Within those 150,000 records are about 790,000 place name references that have been geocoded and can be displayed on the map (in groups/clusters of < 3,500 at one time). The collections/themes extracted from Library Cloud include:

- Houghton Library's Thomas Hollis collection
- Harvard Map Collection records
- Yiddish language records
- Joshua Chamberlain and related records
- "civil war" – keyword anywhere
- "world's fair" – keyword anywhere
- "Harvard" in the title
- Falkland Islands War records

The site provides a search box for users to enter text such as "the Falklands Campaign", just like most catalog discovery tools but unlike traditional catalogs, search results display on a map that outlines the full spatial aspects of the campaign such as personal narratives from journalists publishing their work in Britain; Records that you must otherwise locate by scrolling through pages of records when viewing strictly results from typical text searches.

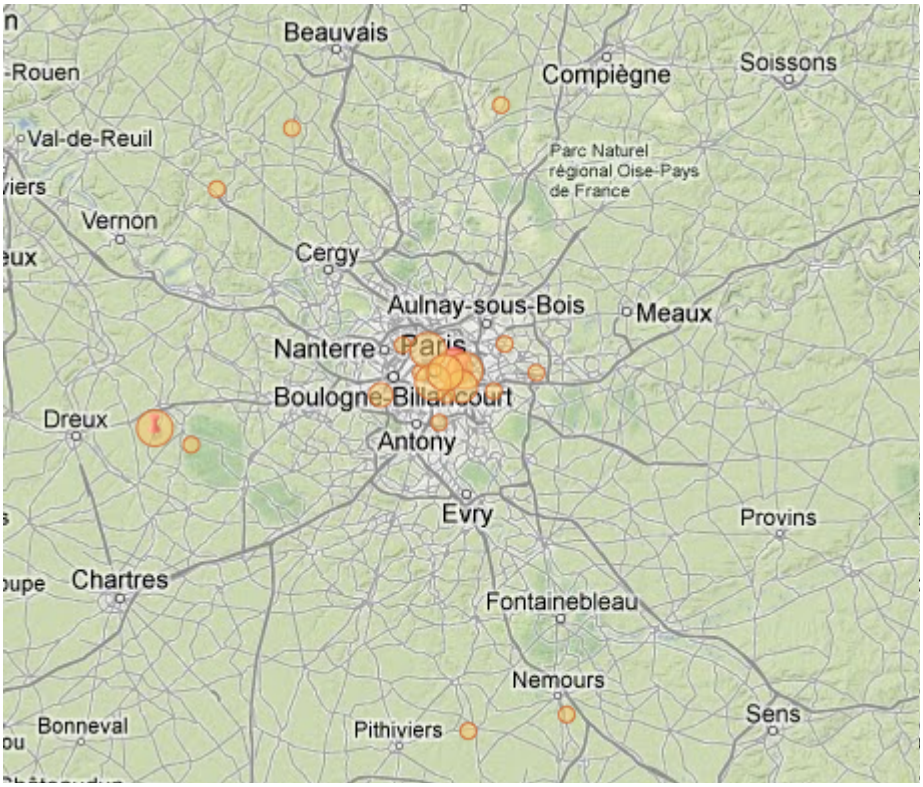
There are several options in the user interface including what fields to search or leave out. It should be noted that this is not a gazetteer search. While you can enter place names like "Cambridge" it is meant to provide a mechanism to search catalog text such as "Joshua Chamberlain".



While a text search exposes resources when you have a specific keywords in mind, a geographic search alone or combined with a text search can show users catalog resources that are related by place thereby allowing discovery of resources that users would not know are related. Or, by showing a broad geographic impact. This is better understood with an example.

Radius, or geographic search. This capability allows you to click on the map and ask the system to expose all catalog resources (of those geocoded) that reference that location.

Here I am interested on resources that reference Paris. I click the map over Paris and the system presents me with all records within 100 kilometers of where I clicked on the map.



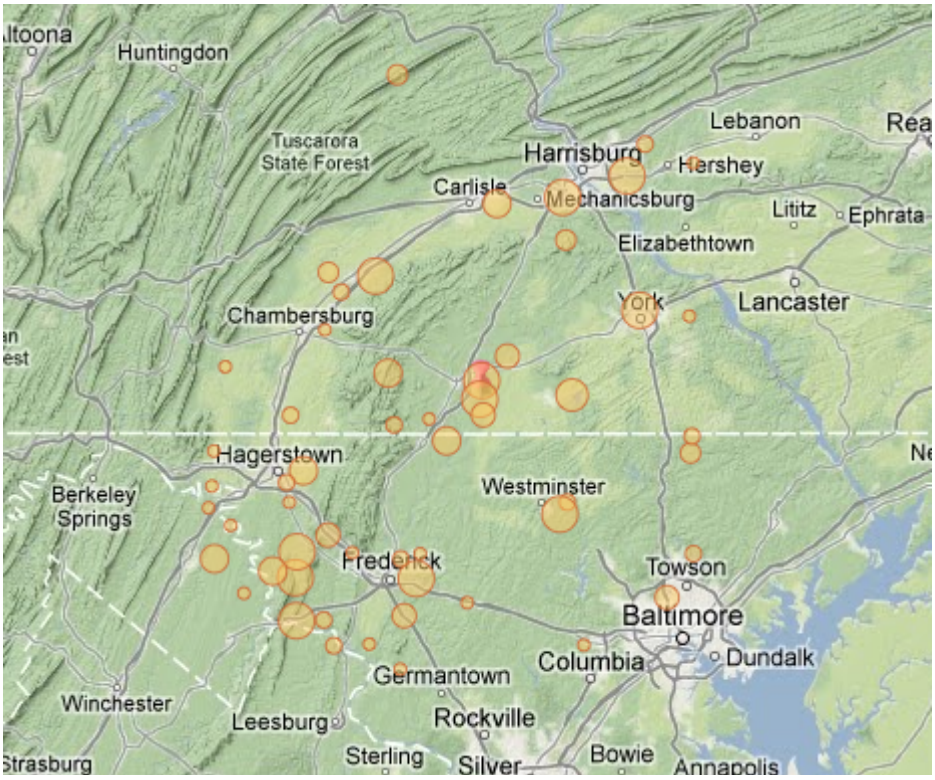
Using the mouse and clicking on a clusters (the size of the cluster is related to the number of records at that location) on the map shows a diverse set of data for this location;

From a book about the Eiffel Tower:
Eiffel's tower *(Book)* | HOLLIS ID: 011980801

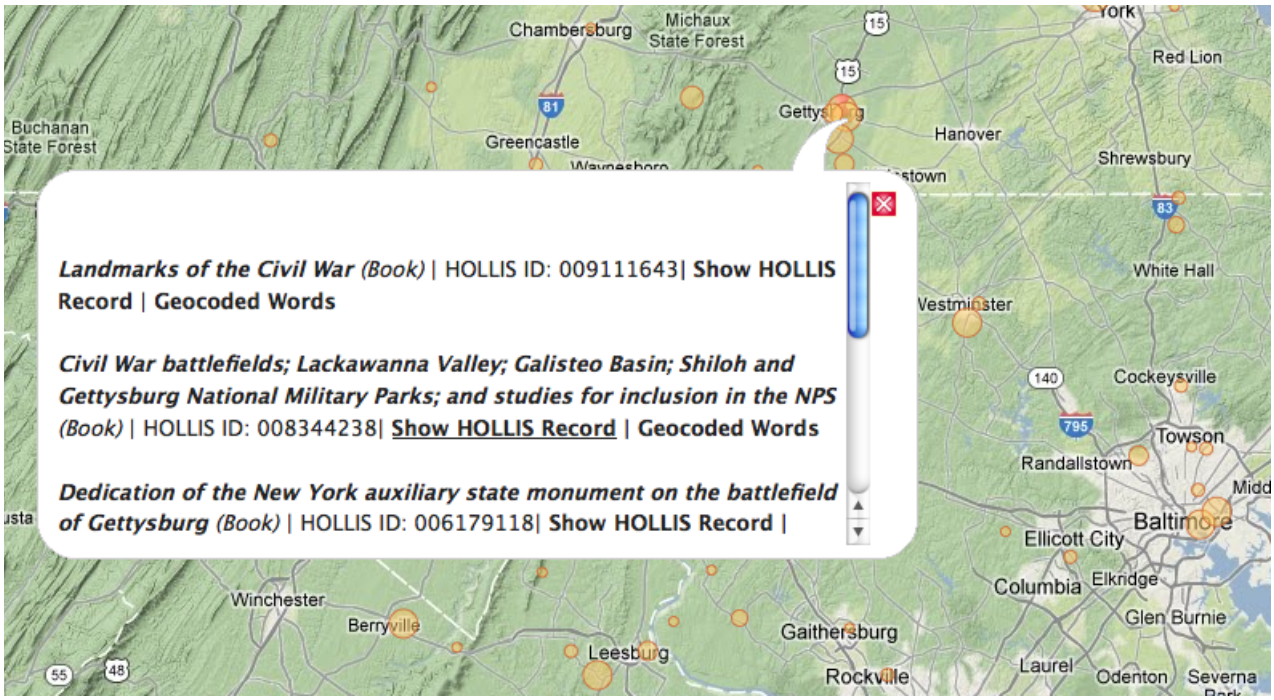
To a panoramic map of Versailles:
Panorama de Versailles *(Map)* | HOLLIS ID: 012265526

And a cultural sound recording:
Klezmer la russe *(Sound Recording)* | HOLLIS ID: 007527742

The spatial search functionality can be used alone as outlined above or used in conjunction with text searching; Next I enter “civil war” as my text search, request all data within 100 kilometers and then click near Gettysburg, PA. The system first filters all data that meet the textual search parameters and then uses the latitude and longitude assigned to each record to filter results within the specified distance. Results are ranked by spatial proximity and text match.

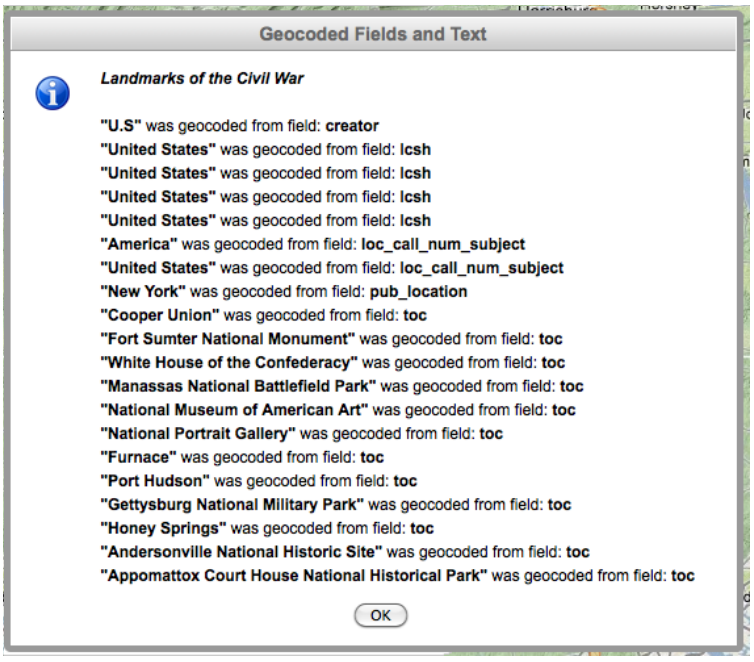


Clicking on a cluster shows HOLLIS records that have geographic references at that location.



From here, selecting a map cluster and then clicking on “Show HOLLIS Record” will open a new window displaying the HOLLIS bibliographic record. Using this technique users can see not only resource titles available to them but the spatial distribution of their search, making it a more effective solution for locating data.

Also, for all data displayed on the map, the UI shows the actual text from the HOLLIS catalog record that was geocoded when you click the “Geocoded Words” link:



For this HOLLIS record, most geographic references were found in the Table of Contents (toc), with some references in the Library of Congress Subject Heading (lcsh) and resource's author(s) (creator). This particular record would show up multiple times, or places where the geocodes were matched to places on the ground.

Other accomplishments include:

- Faceted browsing - the ability to limit a search based on: material type (e.g. books, maps, sound recordings, etc.), authors, Library of Congress Subject Headings, and holding library
- Custom map parameters - the ability to change some mapping options including the way in which the point data is grouped into clusters
- Search results - a more traditional textual display of the top search results for a given query with more extensive bibliographic information.
- Options to map data by specific catalog record field such as "only by place of publications" or "do not map place of publication". Search include/exclude Author, Title, Keyword, Subject.
- A video with an introductory tutorial of features and explanation of the system.

We are also making the geocoded data from Library Cloud available online for others to experiment with.

From this work we feel that there's a platform for users to interact with catalog data in a spatial environment to determine the viability of incorporating spatial methods in other Harvard catalogs.

Challenges <u>– please explain anything you couldn't do</u>

The challenges we encountered are broken into three categories.

Technical hurdles:

The technical task we did not address in this round was the implementation of a time slider in the map interface. This was primarily due to data acquisition difficulties and problems we had with Solr date field handling. There are several plug-ins that assist applications in mapping variables (place of publication and time for example) which could be used to make this happen. This technical hurdle could not be overcome because of time constraints.

Institutional Barriers

Metadata schema changes: We decided early in the project to use metadata records from the Harvard DPLA open metadata set. We had to wait until March before an open API was

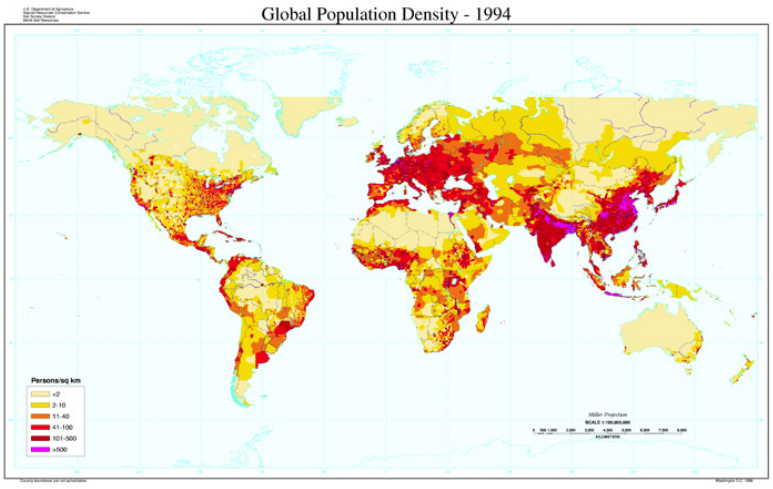
available to use to retrieve records. Then, in August, after starting our development, we learned that the DPLA API was no longer being developed at Harvard and the responsibility for the development of a new metadata schema and API was being transferred to a third party consultant. Since the new DPLA API is still under development we decided to use the Harvard LibraryCloud API, which was not available at the beginning of this project. This required us to start over from scratch: querying and geocoding records, writing a new schema, and adjusting the existing code base to account for the new schema. Having access to the Harvard metadata set through a stable and robust API with published documentation is an essential component of this project.

Deadline/Time Issues

The staff working on this project are also involved in OpenGeoPortal (http://opengeoportal.org/), the underlaying technology to to HGL (Harvard Geospatial Library). We had originally intended to integrate our work into the OGP interface. OGP is a collaborative effort between several institutions including Tufts, MIT, Princeton and Stanford (http://opengeoportal.org). OGP is in the middle of an upgrade which made it too difficult to integrate with our project. We did however, use many of the OGP Javascript libraries for the Beta interface. This will make integration with OGP much easier should we continue in that direction.

Like the following population density map, we originally sought to include a map overlay (using AcidMaps: http://ams.xoomcode.com/) to show data density of referenced catalog locations. Darker areas on the map would indicate increased density of catalog resources available. The deadline approached before we could integrate this tool. This also would make it possible to visualize where Harvard Library resource availability is concentrated.

Global Population Density Map



http://soils.usda.gov/use/worldsoils/mapindex/popden.html

Server-side point clustering for our map display; The Beta interface imposes a limit of how many points can display on the map at a time. This is because raw data from search results is sent to the browser for processing and rendering. This places an undue burden on the client's browser to account for having an underpowered server. This limitation would be greatly minimized if we were to process these data on the server, and send smaller volumes of data to the browser for display. The need for this tool will increase dramatically when more records are added to the system.

Next Steps

Currently there is a global trend towards spatially enabling data, especially catalogs and OCR text from books. This can be witnessed by looking at the current DPLA schema, or sites such as GapVis (http://gap.alexandriaarchive.org/gapvis/index.html#index), oldmapsonline, some components of metaLAB and the PELAGIOS Project . We believe there are several tasks necessary if the Library decides that pursuing this direction is worthwhile.

Probably the most important step is gathering user feedback. Getting users to look at the application and provide feedback is essential for planning further development. Now that the site is publicly available we are hoping that users will use the feedback tool to send us their ideas or comments. This information will need vetting too, preferably with the help of a user interface specialist and librarians. This feedback will help determine if the application is a good discovery tool. And, if so, should it remain a stand alone application or should it be merged with existing catalog systems.

Education is essential. Since geospatial concepts are not always intuitive or familiar, we need to make sure users understand how the application assists in discovery. Searching catalog data usually means browsing through pages and lists of results presented mostly as text. We need a specialist in data visualization and user interfaces to help bridge the gap between textual displays and map displays. We need to accurately gauge; "Do users understand the concept" and "Is this tool useful to them for their discovery needs?" Is it better as an API where the traditional catalog uses the geotagged data to pull in related records/inventory/resources? If so does that mean it is a real-time API or a pre-processed set of data? These are all questions that could be answered in the next step of information gathering from site users.

It is important to keep geocoding catalog records and ingesting them into the Beta site too. That also provides the raw data for others who can use it in their own applications or research. One possibility is placing the value added geocoded metadata in the Library Cloud API.

It is also important to align our efforts with similar initiatives and outline development and implementation tasks with institutions that have expressed interest in doing similar work. This requires testing the infrastructure we've created elsewhere and fine tuning it based on that experience. Increasing capacity and throughput will require securing a much more robust web server that's held by the institutional maintenance and support groups.

Presenting thousands of data points on a map may sound simple but it is not. Server side point clustering to display search results is needed to increase system speed and capacity. There are numerous research projects involving methods to display vast volumes of data on maps but there has to be a technically sound methodology to wean these methods into a production environment.

We need to automate the procedures we developed to geocode catalog data so that we can increase throughput and geocode all HOLLIS catalog records efficiently. Making the process more generic is important too so that other institutions can use what we've done instead of reinventing it. As predicted, open source geocoding tools are catching up and our schema needs to be flexible enough to handle subtle differences between geocoding solutions. We have a system in place that would allow us to geocode ~60 million records using our current software license from Metacarta. With the work done so far we have an excellent head start, but there is still a lot to do in order to provide an open environment for using these data in different ways.

Confirmation that you deposited code, if applicable

Budget spent – I'll provide you with official figures to include

We spent approximately 50% of our budget.

A list of any publicity you did, e.g., articles, blog posts, podcasts, etc.

A video tutorial/introduction was put together. http://youtu.be/ULkZE8TVP3I

A list of any presentations you gave that involved your project

OpenGeoportal Summit @ MIT - April 13, 2012

Library Lab Lightning Talks - July 26, 2012

Ignite Spatial Boston 4 - November 13, 2012