White Paper

Report ID: 110701

Application Number: HK-50091-13

Project Director: Peter K. Bol

Institution: Harvard University

Reporting Period: 9/1/2013-8/31/2016

Report Due: 11/30/2016

Date Submitted: 11/30/2016

Cover Page

NEH White Paper 9/1/2013 – 8/31/2016
Due 11/30/2016

Type of report

White Paper for the period 9/1/2013 – 8/31/2016

Grant number

HK-50091-13

Title of project

Extending WorldMap to Make It Easier for Humanists and Others to Find, Use, and Publish Geospatial Information

Name of project director(s)

Dr. Peter Bol and Dr. Suzanne Blier

Name of grantee institution

Harvard University

Date white paper is submitted

11/30/2016

Overview

This white paper describes a grant (HK-50091-13), which is focused on spatio-temporal platform development in two main areas: 1) Making it easier for researchers across disciplines to discover and use geospatial information, specifically a key class of distributed information known as "web map services" and 2) enabling researchers to create and share information relating to changing place names over time using spatio-temporal gazetteers, or place name databases. For both areas, "depict date" or date which the geodata describes, is critical information to be brought forward within new search and visualization tools, together with geographic representations and other metadata.

Below we describe the project and how it unfolded over the three years (two years, plus one-year extension) we have been working on it. We believe we have developed all of the systems and capabilities promised, and in some fields we have proceeded beyond what was proposed.

Toward the beginning of this project we were buffeted by a number of challenges. First, the software platform we proposed to build upon, Esri Geoportal Server, failed to support the basic service harvesting and monitoring functions we needed. The system also proved to be not easily adapted. We tried another approach, that is, developing the required service harvesting and monitoring capabilities within GeoNode, (the platform WorldMap is based on). This approach turned out to be unscalable and also tied us too tightly to the development cycles of GeoNode.

This led to a challenge which ironically arose because of the success of the GeoNode platform, a platform which was extended very rapidly in 2013 and 2014 as large government agencies and organizations began to adopt and build on it. It was heartening to learn of the vibrant community developing around a platform which is important to us, yet it was also frustrating because it was complicated to plan enhancements against such a rapidly moving target.

Then as a final blow, we lost our lead developer to the private sector. As we started to fill the position, we re-evaluated our technical approach. We decided that whatever direction we chose, it should be one that is loosely coupled with GeoNode rather than tightly integrated with it. This would allow us the freedom to build what is needed without being whipsawed by developments in the broader community. And it would make any solution we would eventually develop more generally applicable, with the ability to run with any system, not just WorldMap/GeoNode. If what we did proved popular and useful, we would know that the GeoNode community would help us ensure that our project/projects would remain aligned with the GeoNode platform.

Once we were not tied to a specific platform, we were able to think afresh about the optimum starting point for the new search capability. In examining non-geospatial systems that focused on search, we realized that numerous very advanced search capabilities had already been built into open source systems like Apache Lucene and Solr, platforms which were being used in many large scale, (non-geospatial) systems. Despite this, powerful search capabilities such as faceting had not made their way into geospatial platforms, especially those dealing with historic information. Faceted search enables users to navigate a multi-dimensional information space by combining text search with a progressive narrowing of choices in each dimension (Wikipedia).

We realized it made more sense to add geospatial capabilities to an existing search platform rather than adding search to a geospatial platform. Regular topic faceting would be useful for geospatial metadata, but what if in addition it were possible to facet on other dimensions such as

space and time? As far we knew, faceting on space had not been done; however, if it were possible, we envisioned it being implemented as a dynamic spatial surface (heatmap). It would describe the geographic distribution of all overlapping map footprints in a search result.

If there are 100,000 maps in a collection and a query for "Roads" returns 5,000 layers, the user is able to see the distribution of layers for the query at a glance. If the time range desired is set to be from 1000BC to 1000AD, and the user needs data covering Kazakhstan, and the heatmap shows there is no data in that area, the user has quickly learned something useful about the collection (no data covering area of interest) and moves on. Conversely, if the user determines there is much data for a given time range and geographic area of interest she knows it makes sense to spend more time exploring the collection. If spatial extent is a critical dimension, as it is for spatial data, it is a missed opportunity to not use it. Visual feedback is very useful in search tools but especially critical in map service registries such as ours, which may be light on other metadata.

We determined that adding 2D faceting (returning a heatmap of feature counts) would be possible. At this step, we were fortunate to be able to work with David Smiley, one of the world's top Lucene/Solr committers. Early tests determined that 2D faceting in Lucene would be highly scalable and could pave the way for big data visualization as well.

We determined the approach would scale to support interactive visualization of the possibly millions of overlapping map layer footprints that we were confident of eventually being able to handle – and that we would be required to handle. Therefore, we added 2D faceting (geospatial heatmaps) to Lucene/Solr, and wrote a heatmap rendering library for OpenLayers. We then set up a proof of concept application based on OpenGeoPortal (http://opengeoportal.org), and it worked well. We now had a prototype for an innovative, effective means of searching against large collections of geospatial content. As Solr provided a robust environment for developing APIs, the approach also addressed the need to expose the service registry via a public API. Another benefit of spatial faceting is that it enables development of a very responsive client. For example, the approach supports a map tooltip that continuously displays layer count (or other information) as the user moves the mouse over the map. A fundamental benefit of Lucene is that is also supports faceting by time, making it possible to implement a time bar on the user interface which displays the distribution of results using a temporal histogram.

By building on Lucene, a number of new capabilities became possible within a system that scales to virtually any number of records using standard hardware. That is, the system is not only useful for map layer discovery but also can handle the 20 million geospatial features in WorldMap in support feature level search, another requirement of this grant. The decision to develop the search independent of WorldMap, and to build it around a search engine, enhancing it as needed, was a turning point in the project. From that point forward the project moved more rapidly.

HHypermap Core, the name of the search oriented improvements developed for this grant, including 2D faceting that the Center for Geographic Analysis added to Lucene, fulfilled the need for a very fast search database with a strong API development environment. Nevertheless, for this project we still needed a map service harvester and monitor. We wanted to design this so it could provide new search capabilities to WorldMap, but we did not want it to run it inside WorldMap/GeoNode and be dependent on WorldMap because we also felt that would limit its utility and the developer community for it. We discovered a good contender, GeoHealthCheck which was written by a leading OGC software expert and decided to port it to Django to make it more compatible with WorldMap/GeoNode. We then added a number of features we needed and called

the new system Harvard Hypermap (HHypermap Registry). Some of the features we added were: 1. support for scalable (asynchronous) harvesting and monitoring of OGC and Esri REST services; 2. gathering of usage statistics to use to improve search; 3. caching and re-projection of map services to support instant previews, speed-up of map display, and reducing strain on remote servers; 4. time-miner to better support search by depict date; 5. feature extraction to improve layer search.

We were on our way to the harvesting and monitoring system we needed. But how to make the registry comprehensive, and how would we know when it was complete? One part of the solution would be to crawl the web for endpoints. The problem with this approach is that it is likely that many endpoints do not end up in web pages, since they are not very interesting for humans to talk about. A very important source of endpoints will always be those which are contributed to the system by users, and we have made it easy for users to contribute them and have them harvested.

While it is possible to create an index of web pages by following hyperlinks (this is what Google does to create its index), it is not possible to create an index of web map services in the same way, as map services do not point to one another the way web pages do.

However, web pages are places to start. We wrote an application for crawling the web which runs in parallel on Hadoop and rapidly searches the web as represented by the Common Crawl (http://commoncrawl.org), looking for and recording OGC and Esri service endpoints. Here is a summary of the endpoints we discovered:

File/Service Type	URL Count
OGC servers (getcapabilities, getmap)	4,950
OGC WMTS tile server	4,690
ESRI Rest servers (arcgis/rest/services)	27,816
Total	**37,456**

(In our experience so far, harvesting server endpoints, we have found that the average number of layers on a server is approximately 30. If we extrapolate based just on these very incomplete numbers, we would have over a million layers.)

Since the crawler would not add much overhead, we decided to look for other kinds of geospatial materials that could be included in the registry sometime in the future. Here is a list of what we found:

File/Service Type	URL Count
kml/kmz	1,309,660
tfw	26,316
gpx	26,761
geojson	6,401
gdb	2,042
ecw	1,172
sid	169,500
jp2	153,563
zipped shape	62,221
thredds	411,111
Total	**2,168,747**

Another major source of endpoints will be the many catalogues on the web that store map service information. These include: CSW, WMS, Esri Open Data, Esri REST, Map Warper, David Rumsey, Open Geo Portal, GeoBlackLight, Thredds, and others. Each catalogue type requires a connector to harvest layer endpoints and metadata from them. We have created connectors in HHypermap for CSW, Map Warper, WMS, Esri REST, and WorldMap. Additionally, we have started a connector for the David Rumsey map collection.

Another way to make the registry more comprehensive is through the sharing of information between HHypermap instances. If instances were to share, each registry would get better. One of the benefits of this being an open source project (one which is being taken up by other organizations) is that useful features get added without our necessarily developing them. An example is the addition of the OGC standard CSW-T or transactional catalogue service capability which makes it possible for trusted CSW-T catalogue servers to share content with HHypermap and for HHypermap to share content with other CSW-T catalogues. This and other important enhancements (including addition of Elastic Search) have been recently sponsored by the company Boundless.

We are currently focusing on handling relatively simple map service types, i.e., OGC WMS and Esri REST Map, but there are other powerful OGC and Esri map service types which could be harvested and made more discoverable. So far we have had requests for OGC WFS (web feature services) and WPS (web processing services). Though not a requirement, we also included support for two flavors of tile service, those following the OGC WMTS standard, and the Esri tile format which is closely related to WMTS.

The core technology we developed has already proved itself useful outside of WorldMap. Soon after we had a proof of concept working, another project at Harvard, the Japan Data Archive, chose to use HHypermap Core to solve a problem displaying large numbers of query results on a map based on a database which contains more than a million features. A little later we were asked to participate in a proposal to Sloan to create a big data visualization capability for the Dataverse (http://dataverse.org) system, a social science data archiving system. The proposal was to add a big spatial data access capability to Dataverse. We proposed scaling up HHypermap Core and creating a system to provide interactive query access to the billion most recent geolocated tweets. (We had been harvesting geo-tweets since 2012 and had several billion stored on disk.) The Dataverse/CGA

team won the grant, and the Center for Geographic Analysis is currently building the system, called the "Billion Object Platform," or BOP, and in the process refining and hardening the HHypermap Core technology developed for this NEH grant. This lead to a collaboration with Massachusetts Open Cloud, a public cloud which has offered to host the BOP.

The Center for Geographic Analysis has created a client plugin for WorldMap so that WorldMap users can search the HHypermap registry, which contains all WorldMap layers and all harvestable layers for all known (at least at the present moment) remote OGC and ESRI map servers, and start using them. This client supports search by time, space, keyword, source and type, and has a variety of search features which are described below. An advanced search client being developed for the BOP supports temporal histograms and topic clouds, technology which is a good starting point for more advanced HHypermap Registry search.

Once we had harvested a few thousand remote services, it became clear many services did not contain information for depict date, including many layers within WorldMap. Layers missing formal date metadata did however often include date or date range references within the layer title or description. Though not specified in our proposal, we decided to improve temporal search for service layers, by writing a "Time Miner" which extracts date values from name and description fields and adds them to the date field used in search. This is a work in progress but it does promise to add considerable value. We have roughly doubled the total layers with dates in the system without adding in many spurious values. The system makes it clear which layers have dates ,which are automatically detected, and which are user defined.

Upon having a development version of the new registry search system working, we recognized that we were filling a niche. A federal agency started using it to build a registry of map services in order to make data more available within their large organization. Then the commercial mapping company Boundless incorporated HHypermap into their Boundless Exchange platform.

An important capability we developed which was not originally not part of our proposal -- but which we believe is critical for creating a truly useful navigation and display interface for very large numbers of map layers -- is layer caching using an application called MapProxy. This approach has several important benefits: 1. It enables instant previews such that mousing over a layer in the search client displays the layer instantly on the map; 2. Layers which are not in the spherical Mercator projection (used by the map client) are re-projected and displayed; 3. by caching on demand, no layer has to be requested from the server more than once, thus reducing demand on remote servers while greatly increasing display speed for the end user. HHypermap is now being developed by groups outside Harvard. For example, the enhancements to MapProxy to support Esri services was contributed by the U.S. State Department.

The above describes work related to the HHypermap layer registry. However the technology we developed is also useful for the spatio-temporal gazetteer, the development of which represents the other significant segment of this project.

We have developed a gazetteer capability which allows any user to contribute the features from any layer they own to their own custom gazetteer. The capability allows users to upload geospatial layers to the system and turn them into spatio-temporal gazetteers which can then be exposed within WorldMap and/or within other systems via a flexible gazetteer API. This RESTful, JSON API supports a variety of ways of for querying and aggregating gazetteer materials, including by date or date range, by gazetteer name (a user can add as many layers as she likes to a gazetteer name), map (a user can query by the layers within a particular WorldMap map), layer (a user can

query by layer name). Any layer which is contributed to the gazetteer is also contributed to the WorldMap or "house" crowdsourced gazetteer. Gazetteers will be searchable from the gazetteer tool on the map page.

We worked with temporal gazetteer authority Lex Berman on the design of the gazetteer, and are now working with the spatial historian and University of Maine professor Anne Knowles to refine the design of the gazetteer to support her requirements.

As already pointed out, a benefit of using Lucene/Solr is scalability. This allows us to support robustly the feature-level search requirement for this proposal which includes 20 million features. Feature level search allows a user to discover WorldMap layers by searching against their feature level content in addition to searching against layer metadata. This capability is supported using a separate Solr feature registry which includes all features in all WorldMap layers and makes them searchable by time, space, and keyword.

The depict date information provided to the feature registry for gazetteer layers will be the same as is defined for the gazetteer. Features from layers which have dates at the layer level, either user defined or automatically detected by the Time Miner, will be assigned those layer level dates. The house gazetteer will therefore contain a subset of a larger collection of temporally defined features. This larger collection likely lacks sufficient vetting to be considered a gazetteer; yet it could serve as an on-ramp to the gazetteer, providing visibility to valuable materials which could be contributed. This approach means gazetteer features will be accessible through the new advanced Solr API, which supports advanced clients that make use of spatial and temporal faceting, as well as being accessible via the more traditional API describe above.

Evaluation

The project was not independently evaluated. We did however evaluate the project through a public workshop and survey, "Workshop and Live Demo of WorldMap 2.0 with a User Survey" at the Harvard Center for Geographic Analysis, Conference on Space, Place, and Geographic Thinking in the Humanities in April of 2016.

At this workshop, at a hands-on workshop using an Alpha version of the new NEH funded search capabilities in WorldMap, participants were given a survey. Users were first provided an overview of WorldMap and the enhancements funded by this grant. Then the new search capabilities were demonstrated and attendees were asked to try out the application themselves on their laptops, and then fill out the survey.

In 2014 the CGA initiated a study of registered WorldMap users and 290 people responded. The study, "Understanding Today's Online GIS User Through the Lens of a WorldMap Survey" was published by Wiley. An important finding in the conclusion of this study was that WorldMap users are a diverse group in terms of profession, discipline, and technical background.

> …WorldMap users make up a diverse group by age, gender, profession, discipline, and nationality. More than half are not regular GIS users, and over a quarter are first-time GIS users. This survey provides quantitative evidence that a significant portion of successful WorldMap users are indeed people without any previous GIS experience. The system serves a broad range of disciplines and professions, as it was originally intended to do by the project team from the beginning. In this way,

WorldMap appears to be lowering technical barriers to making use of spatial information, and collaborating with such material.

The audience for this grant work is also WorldMap users but with a greater focus on researchers in the humanities and social sciences, especially those who care about spatial information, from the recent and distant past. The system is designed to make it easier for anyone to find and start using historic georeferenced data wherever it may be published on the web.

Unfortunately, the new search capabilities did not get moved into the production version of WorldMap until November 20th of 2016; thus, there has not been time to gauge the audience response from the broader spectrum of WorldMap users, or to do a follow-up study to determine whether the changes we have made have expanded or changed the audience. However, it is to point out that the technology we have developed is in no way dependent on WorldMap. This means that evaluating the impact on the audience must take into account use of the technology by other organizations running other systems, which use the technology we have developed.

We find it encouraging that two large organizations with diverse user bases (U.S. State Department and Boundless), have adopted HHypermap technology within their core platforms, and we are encouraged at the response to our April workshop and survey (see Section 4 above), as to the many presentations we have made inside and outside Harvard.

We plan to continue hosting the WorldMap platform and continue building on it and improving it. WorldMap with the NEH enhancements of this project is a main platform for enabling one of our core missions for our researchers, that is, to lower the barrier to the access of geospatial data and technology.

Presentations and Papers

Paper: "Implementing an open source spatio-temporal search platform for Spatial Data Infrastructures." Paolo Corti, Benjamin Lewis, Tom Kralidis, Ntabathia Jude Mwenda. Open Source Geospatial Research and Education Symposium. PeerJ, 2016. https://peerj.com/preprints/2238.

Presentation: "Implementing an open source spatio-temporal search platform for Spatial Data Infrastructures." Paolo Corti, Benjamin Lewis. Open Source Geospatial Research and Education Symposium. Perugia, Italy. 2016. https://docs.google.com/presentation/d/1gp5S4bxygAfOUkGUBg-_ZJouzjVLJDTVGBYpL6cIlEk/edit?usp=sharing.

"HHypermap: Heatmap Analytics of a Billion Tweets." David Smiley. Lucene Revolution Conference. Boston. 2016. https://lsr2016.sched.org/event/7amM/hhypermap-heatmap-analytics-of-a-billion-tweets.

"HHypermap: A Platform for Building and Maintaining a Global Geoservices Registry." Benjamin Lewis, Paolo Corti. CyberGIS'16. Urbana, Illinois. 2016. http://cybergis.illinois.edu/events/cybergis16/program.

Lecture and Workshop on WorldMap and HHypermap to Chinese Government Officials. Benjamin Lewis. George Mason University, 2016.

"Building an Open Source, Real-Time Updated, Billion Object Spatio-Temporal Search Platform. International Workshop on Cloud Computing and Big Data (IWCCBD)." Paolo Corti, Ben Lewis. Fairfax, VA. 2016. http://cga-download.hmdc.harvard.edu/publish_web/website_files/PDF_MISC/2016_IWCCBD_Paolo_Corti.pdf.

"Intro to Global Spatio-Temporal Search: NEH Funded Enhancements to WorldMap." Benjamin Lewis. 2016. Harvard Center for Geographic Analysis, Conference on Space, Place, and Geographic Thinking in the Humanities. Benjamin Lewis. 2016. https://vimeo.com/174556000.

"HHypermap: A Platform to Enable Geospatial Search." Benjamin Lewis. FOSS4G Conference, Raleigh, North Carolina. 2016. https://2016.foss4g-na.org/session/hhypermap-platform-enable-geospatial-search.

"WorldMap: Using GeoNode to Organize Distributed Geospatial Data." Benjamin Lewis. Secondary Cities Symposium Sponsored by the U.S. State Department. Harvard University. 2016. http://secondarycities.state.gov/events/dynamic-map-symp-page-5.html.

"Evaluating the Current State of Geospatial Software as a Service Platforms: A Comparison Study". Chapter in forthcoming book: Citizen Empowered Mapping. Benjamin Lewis, Weihe Wendy Guan, Alenka Poplin. Springer. 2016.

"WorldMap: Open Source Geospatial Infrastructure for Collaboration." Benjamin Lewis. Tongji Delegation. Harvard University. 2016.

WorldMap and HHypermap demonstration for Professor Henry Louis Gates. Harvard University. Suzanne Blier, Benjamin Lewis. 2016.

"The BOP (Billion Object Platform) and WorldMap / Dataverse Integration." Benjamin Lewis. Dataverse Community Meeting. 2016.

"Mining Time Signatures from Spatial Data in Support of Humanities Research." Jude Mwenda Ntabathia, Benjamin Lewis, Paolo Corti, Wendy Guan. American Association of Geographers. San Francisco, CA. 2016.

"Visualizing 10 Million GeoNames with Leaflet and Solr Heatmap Facets." Jack Reed, (co-founder GeoBlacklight). 2015. http://www.jack-reed.com/2015/06/29/visualizing-10-million-geonames-with-leaflet-solr-heatmap-facets.html

"Investigating Hadoop for Large Spatio-temporal Processing Tasks." David Strohschein, Stephen Mcdonald, Benjamin Lewis, Weihe Wendy Guan. American Association of Geographers. Chicago, IL. 2015.

"Understanding today's online GIS user through the lens of a WorldMap survey." Weihe (Wendy) Guan, Alenka Poplin and Benjamin G. Lewis, 2015. Transactions in GIS, x(y), John Wiley & Sons Ltd. DOI: http://onlinelibrary.wiley.com/doi/10.1111/tgis.12150/abstract.

"Investigating Hadoop for Large Spatiotemporal Processing Tasks". American Association of Geographers. David Strohschein, Stephen McDonald, Benjamin Lewis, Weihe Wendy Guan. Chicago, IL. 2015.

"Searching all the Web's Spatial Data." Steve McDonald. A talk at a Harvard tech organization which described NEH work on web crawling and harvesting. 2015. http://www.gis.harvard.edu/events/seminar-series/searching-all-webs-spatial-data.

"WorldMap: Platform to Support Collaboration." (A talk on the NEH work in the context of building an infrastructure to support the documentation of pre-removal Native American place name information.) Conference on Digital Mapping & the Post-Removal Indigenous Midwest. Benjamin Lewis. University of Iowa. 2015. https://obermann.uiowa.edu/events/digital-mapping-post-removal-indigenous-midwest.

Keynote address "Building a Public Infrastructure to Improve Geospatial Collaboration." Benjamin Lewis. 2015 Pennsylvania GIS Conference. State College, Pennsylvania. http://www.pagisconference.org/Documents/2015/2015GISProgramBook.pdf.

"WorldMap: Open Infrastructure to Support Geospatial Knowledge Building." Benjamin Lewis. Workshop for Chinese Government Officials. George Mason University. 2015.

"Addressing the Problem of Geospatial Search." Benjamin Lewis. National Science Foundation (NSF) Industry/University Cooperative Research Center (I/UCRC), Spatiotemporal Innovation Center, George Mason University. 2015.

"WorldMap: A Spatial Platform to Support Research and Collaboration." Benjamin Lewis. MIT Program on Information Science, Cambridge, Massachusetts. 2015.

"Extending WorldMap to Make It Easier for Humanists and Others to Find, Use, and Publish Geospatial Information." Peter K. Bol, Benjamin G. Lewis, Weihe Wendy Guan. CyberGIS2014 Conference. Redlands, California. 2014. https://www.youtube.com/watch?v=sS6pD_wqG8A.

"Expanding WorldMap." Stephen McDonald, Benjamin Lewis. Spatial Search Specialist Meeting. University of Santa Barbara. 2014. http://spatial.ucsb.edu/wp-content/uploads/smss2014-All_Position_Papers.pdf.

WorldMap workshop which focused on the utility of map services and this NEH funded approach for making them easier to discover and use. Ben Lewis. University Consortium for GIS (UCGIS) Symposium. Pasadena, California. 2014.

Presentation at the offices of the United States Geological Survey in Portland OR about on the NEH project to build a global registry of web map services. Ben Lewis. 2014. http://or.water.usgs.gov/brownbag/seminar_history/spring_summer2014seminars.html

Online talk to the Mid-Atlantic Geospatial Transportation Users Group on open source technologies for using map services to share transportation information. Ben Lewis. 2014. https://magtug.wordpress.com/2014/09/09/magtug-webinar-10714/

"Collaboration and Sharing is the Promise of the Web." Peter K. Bol. Shanghai Forum. 2013. http://www.shanghaiforum.fudan.edu.cn/en/index.php?c=news&a=detail&aid=1153.

Source code developed for this project

WorldMap: https://github.com/cga-harvard/cga-worldmap
HHypermap Registry: https://github.com/cga-harvard/HHypermap
HHypermap Core Solr: https://issues.apache.org/jira/browse/SOLR-7005
HHypermap Core Lucene: https://issues.apache.org/jira/browse/LUCENE-6191
Time Miner: https://github.com/cga-harvard/HHypermap/tree/master/hypermap/dynasty

Hosted implementations of the system developed for this project

WorldMap instance: http://worldmap.harvard.edu
HHypermap Search on Development server: http://hypersearch.cga.terranodo.io/maps/new (click "Add Layers")
HHypermap backend: http://hh.worldmap.harvard.edu