American Association of Geographers

Boston, Massachusetts

April, 2017

Harvard Hypermap:

An Open Source Framework for Making the World's Geospatial Information more Accessible

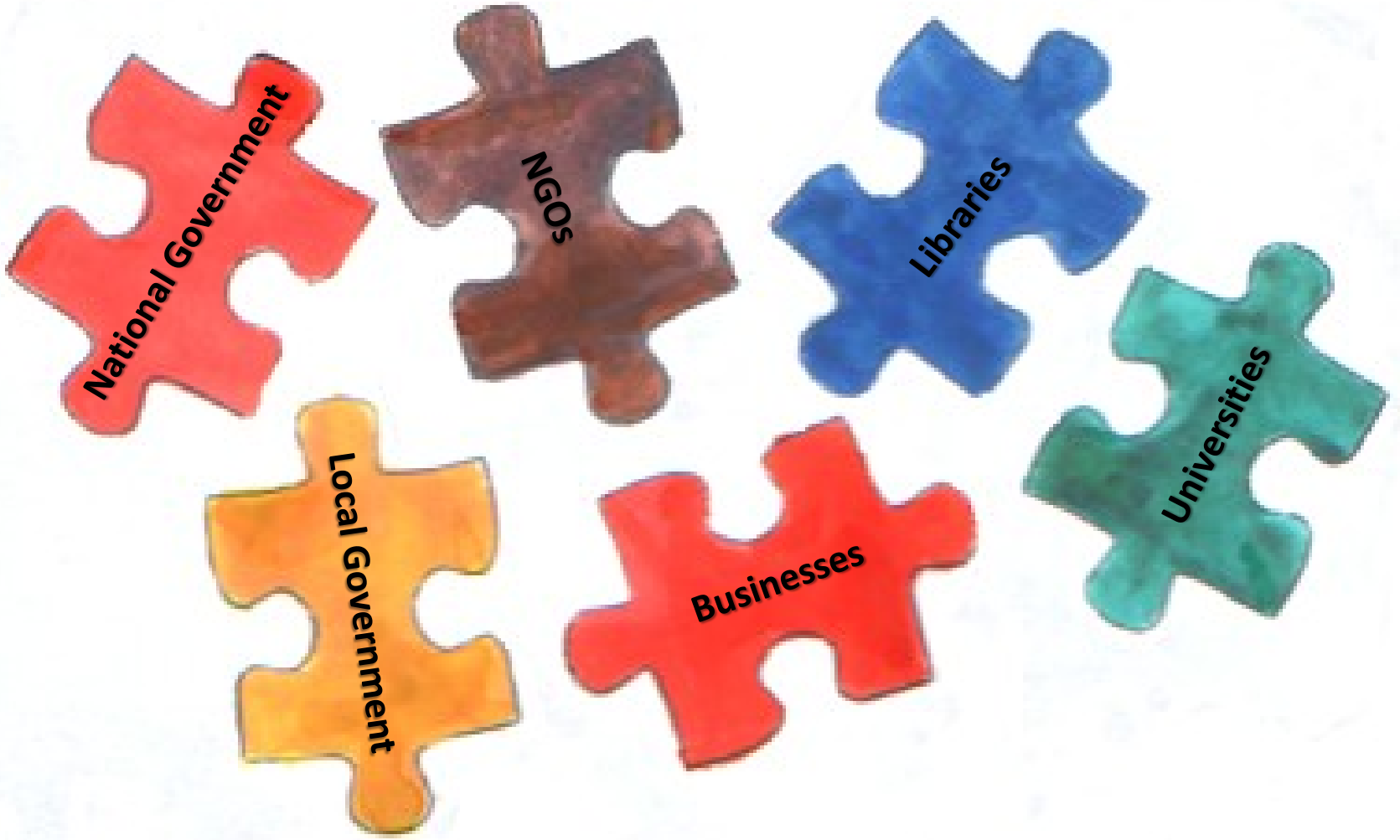
Benjamin Lewis, Paolo Corti, Wendy Guan

Center for
Geographic Analysis

Harvard University

http://worldmap.harvard.edu

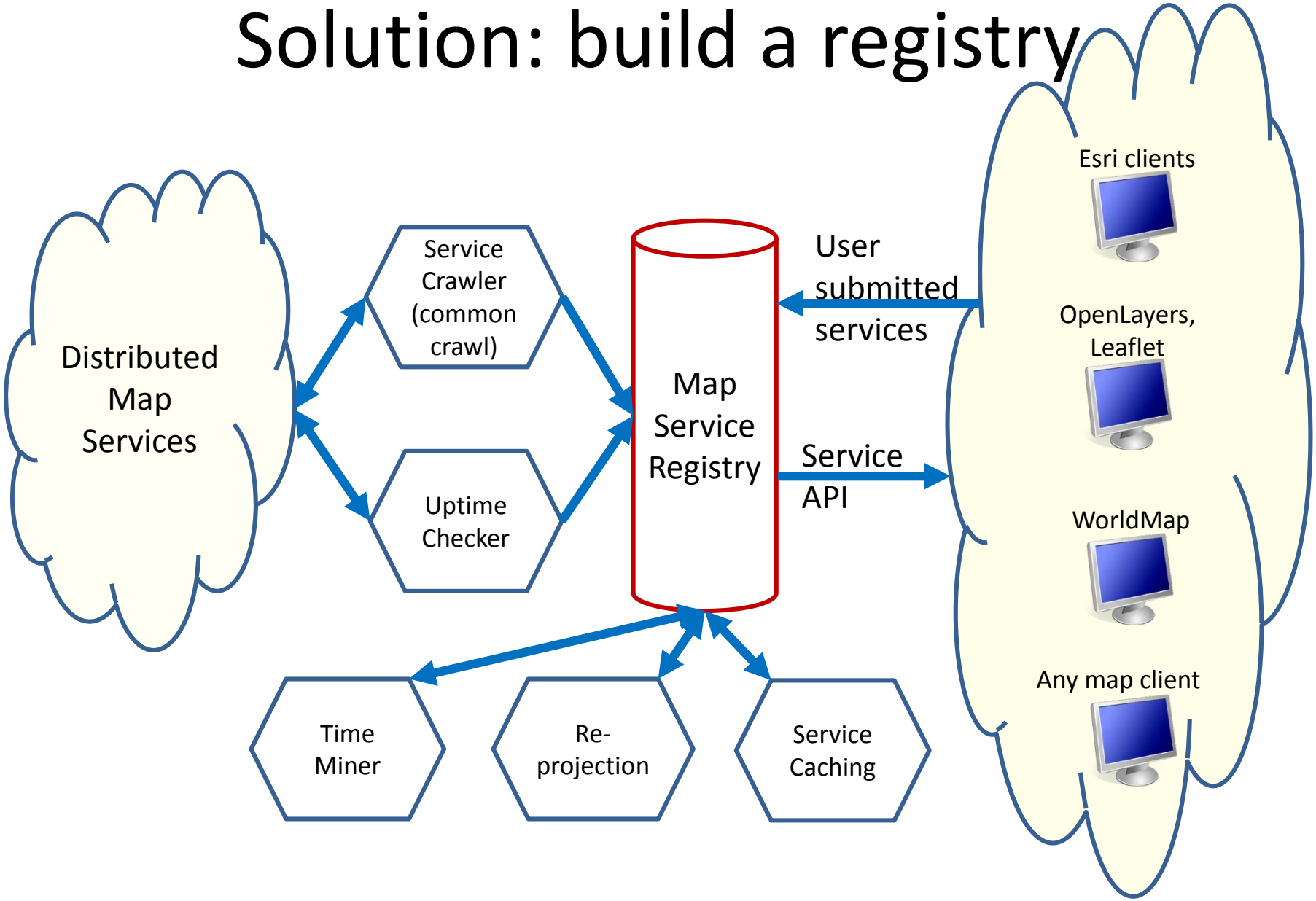
Where are the geo-data?



Web Map Services

- A powerful, modern way to provide geo-data via the Internet
- Distributed across thousands of servers globally
- No central directory
- Tremendous variety in publishing service formats (WMS, WFS, Esri REST, KML, etc.)
- Difficult to find, and difficult to use

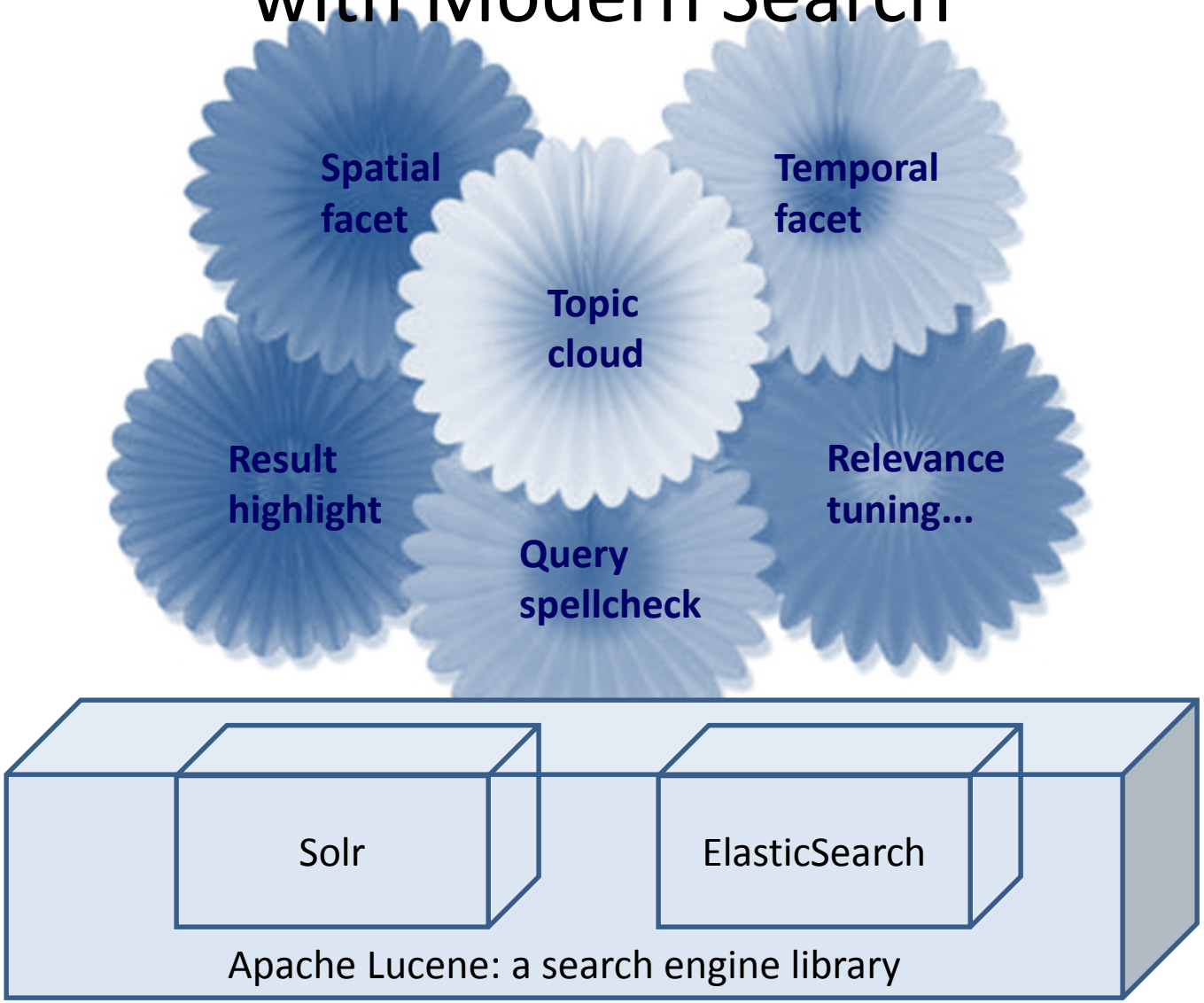
Solution: build a registry



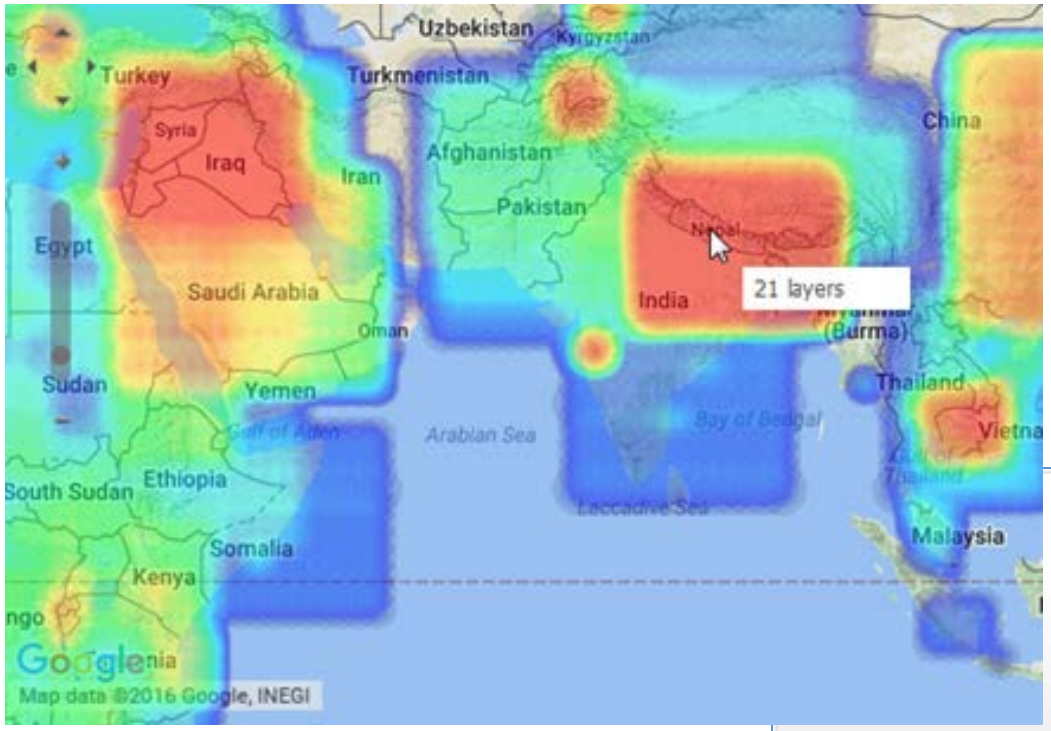
Harvard HHypermap Highlights

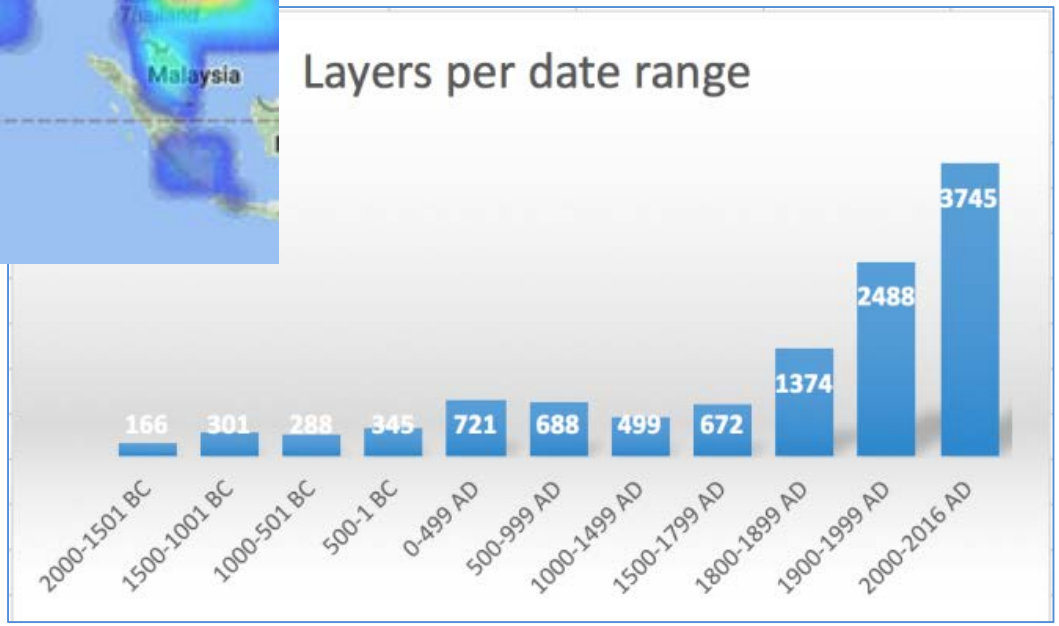
- Supports comprehensive search
  - Visualize distribution of results
  - Support space and time
  - Fast
- Facilitates sharing between registries
- Accepts user contributions
- Improves over time
- Software is open source

Spatial Data Infrastructure (SDI) with Modern Search



Implementation in WorldMap





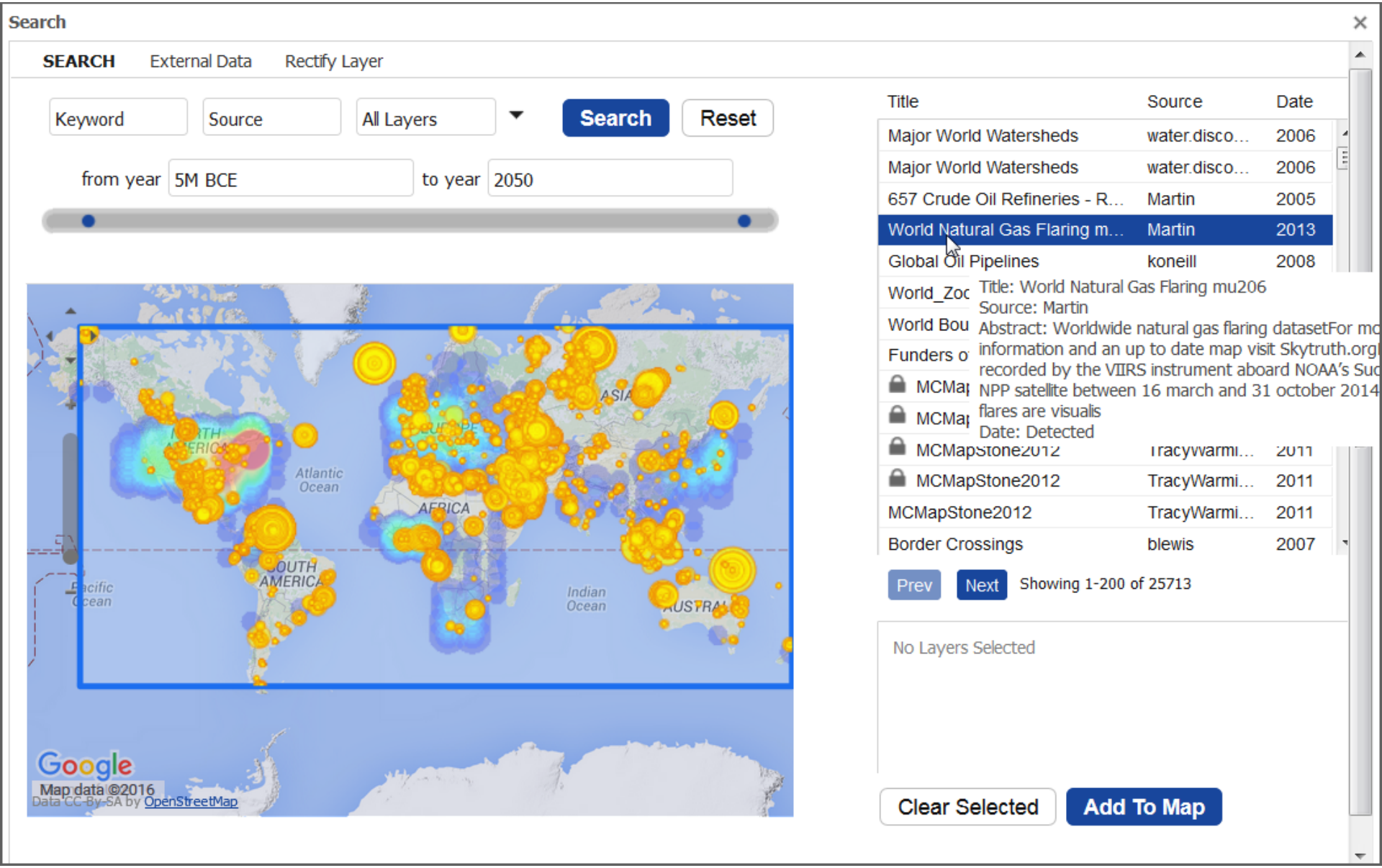
HHypermap as WorldMap's search engine

- Provides access to about 100,000 layers
- 25,000 are from WorldMap itself, the rest are on remote servers
- Accesses 13,000 remote map services

- A fraction of existing online map services
- More are being added

Current search on WorldMap

- Supports *faceted* search by
  - Geographic space (heatmap for any number of results)
  - Time (temporal histogram)
  - Subject matter (topic clouds)
  - Source (data owner, publisher)
- and potentially other dimensions

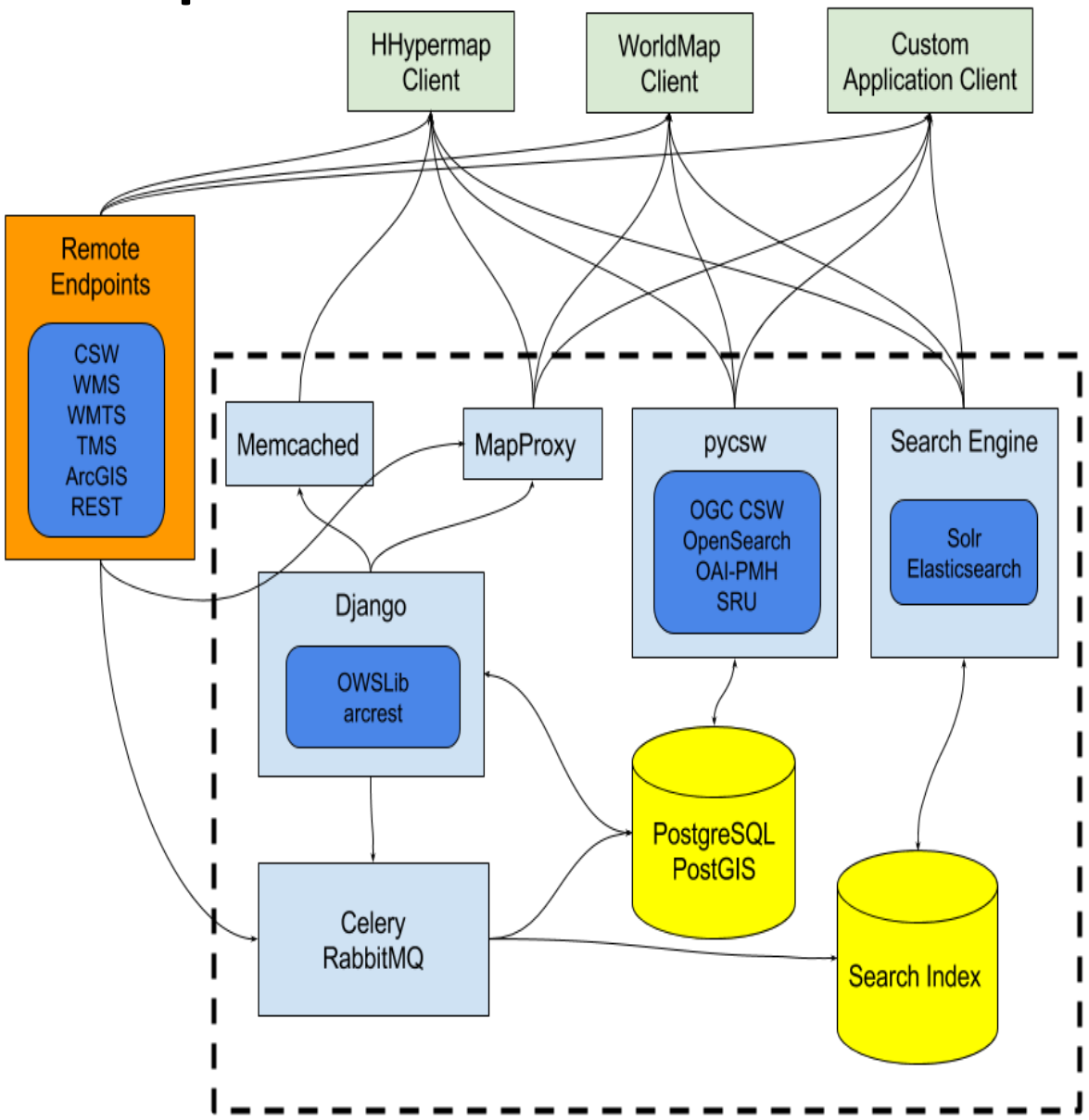


Video demo (3 minutes): https://vimeo.com/164167343 (narrated by Ben Lewis)

HHypermap Architecture

Built on open source software:

- Celery
- RabbitMQ
- Django
- Lucene
  - Solr
  - Elasticsearch
- MapProxy
- Memcached
- OWSLib
- PostgreSQL
- PostGIS
- pycsw



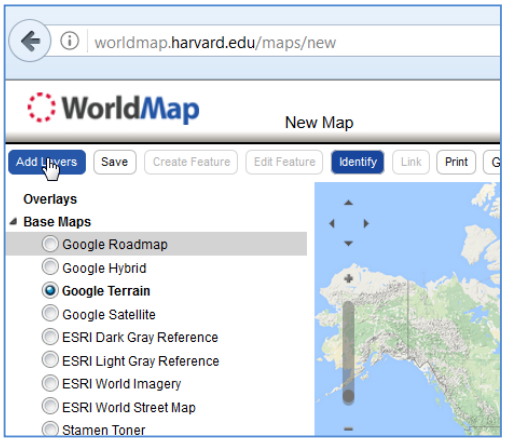
Automated Gathering of Map Service Endpoints to Harvest

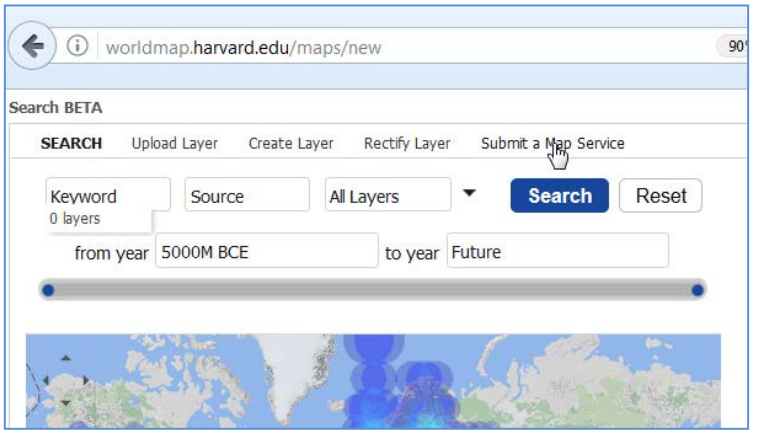
- Search web for signatures using the Common Crawl (CC) archive
- Store as compressed Web Archive (WARC) formatted files on Amazon S3.
- Employ multiple machines to process the data in parallel on Amazon EC2
- Use Hadoop/YARN framework, execute Map/Reduce functions to aggregate information about URLs to spatial assets
- Collect URLs for later processing and harvesting

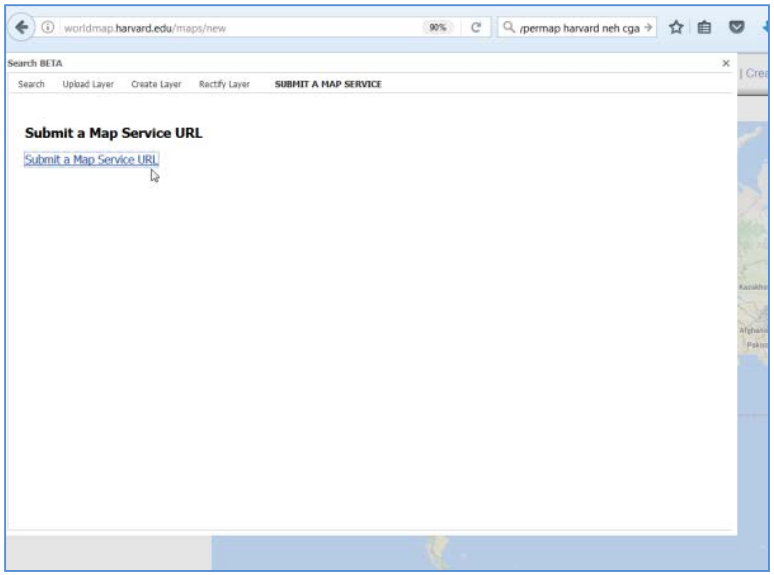
Signatures for Hadoop search

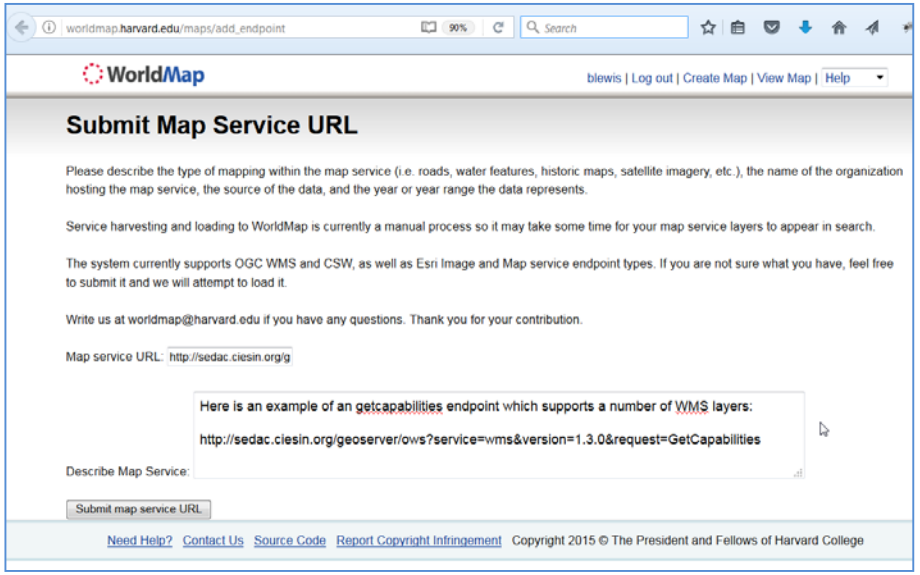
- OGC Services
  - Look for "?request getcapabilities" and not “test” in the href URL
- ESRI Rest Services
  - Look for “/arcgis/rest/services” in the target-DOMAIN-URI of the WARC Response Header text
- KML or KMZ files
  - Look for an href URL ending in .kml or .kmz files
- Compressed shapefiles
  - Look for “shape” or “shp” and string ending with “.zip” in the href URL
- Tile Servers
  - Look for “tile” or “tiles” and string ending with “.png” in the href URL

User Submit Service Endpoint to Harvest









Time Miner – to enrich metadata

- Temporal metadata for geospatial datasets is often weak.
- In a crowd-sourced data repository, data creators and contributors often do not create detailed metadata.
- Many data sets have temporal properties, but time is often ambiguously defined, mentioned as unstructured text in the title, abstract, and elsewhere.
- Time is often not referred to using a standard date/time format such as ISO 8601, but as descriptive text.

Time Miner Logic Implementation

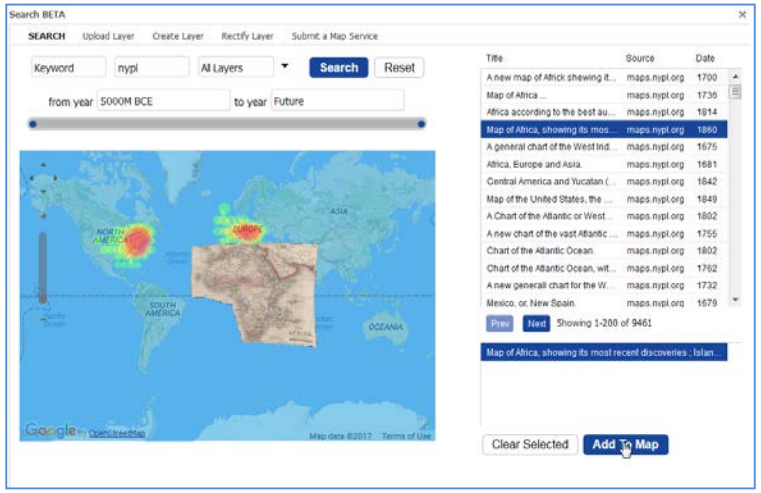
1. Look for date in the date range (lower) section of the metadata and choose the earlier date. (Date: from Metadata)
2. If there is no #1 above, look in top date in metadata but only use it if it is 2010 or earlier. (Date: from Metadata)
3. If there is no #2 above, look for 4 digit numbers in title first, then abstract, which are less than or equal to 2016 (present year) (Date: Detected)
4. If there IS a date in #3 above, check to see whether there is a CE or AD or BCE or BC after it and apply math accordingly (Date: Detected)
5. If there is no #3 above, look for 1, 2, or 3 digit numbers with associated CE, AD, BCE, BC, and apply math accordingly (Date: Detected)

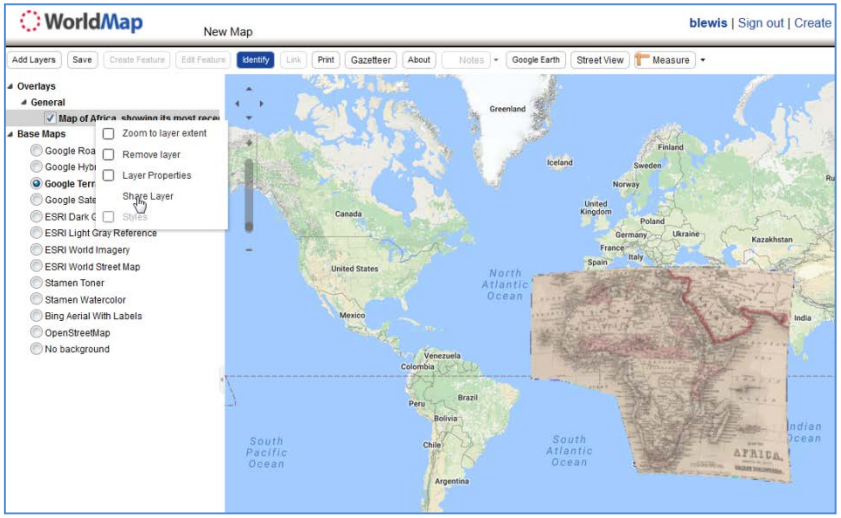
Time Miner: in addition to dates, recognize some periods

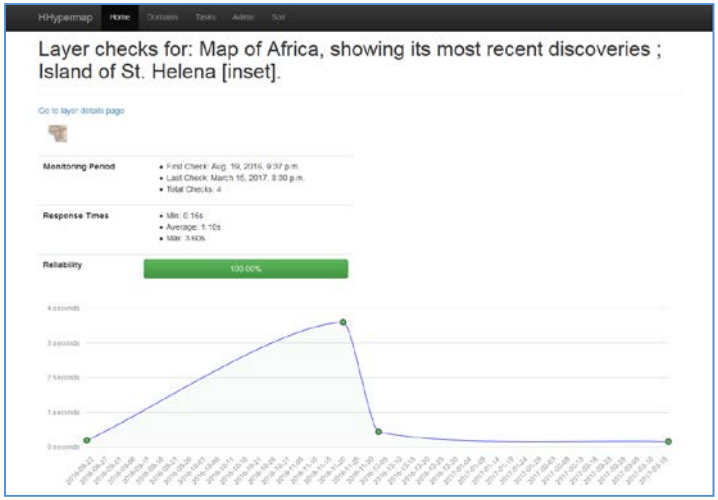
- ca. 2100-1600 BCE Xia, Hsia
- ca. 1600-1050 BCE Shang
- ca. 1046-256 BCE Zhou, Chou
- 221-206 BCE Qin, Ch'in
- 206 BCE-220 CE Han
- 581-618 CE Sui
- 618-906 Tang, T'ang
- 960-1279 Song, Sung
- 1279-1368 Yuan
- 1368-1644 Ming
- 1644-1912 Qing, Ch'ing

Source: http://afe.easia.columbia.edu/timelines/china_timeline.htm

Each layer monitored for uptime

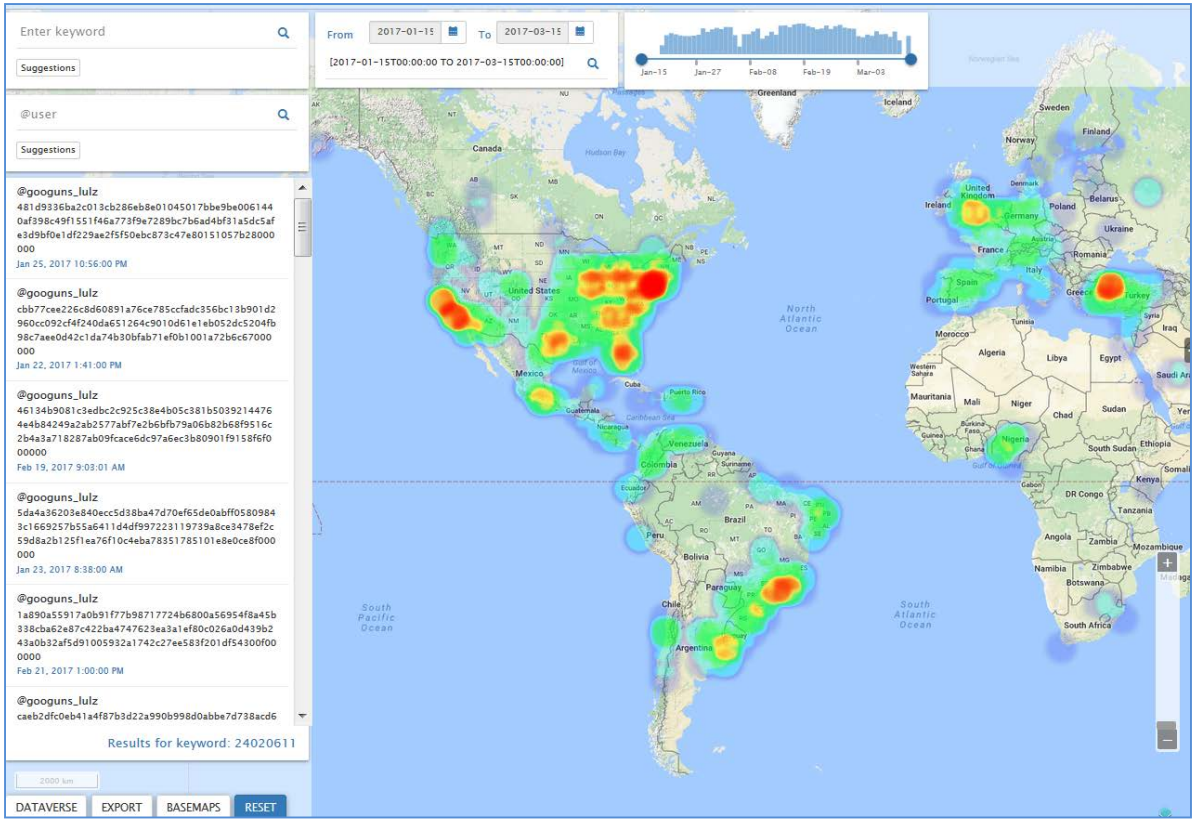






Scalable to support virtually any number of datasets

Another project the “Billion Object Platform” (BOP) demonstrates how robust this platform is!



http://terranodo.io/angular-search/#/search

Some examples of catalogues that could be brought together

- ArcGIS Open Data – Esri collection of 44,000 open datasets and growing.
- Geodata.gov, Geoplatform.gov – The U.S. Federal government has built a data sharing platform for U.S. data using CKAN software and ArcGIS Online.
- INSPIRE Geoportal – Spatial data portal of the European Commission.
- GEOSS registry – Group on Earth Observations registry of 850 map service collections.
- Geopole.org – CSW catalogue service providing access to 400,000 layers.
- Geoblacklight - Platform developed by Stanford and other universities to provide fast search access to geospatial library holdings.
- OpenGeoPortal – Platform developed by Tufts and other universities to provide fast search access to geospatial library holdings.
- Geonetwork – Geospatial catalogue maintained by the Food and Agriculture Organization of the United Nations.
- Spatineo.com – Commercial service which is currently monitoring 40,000 web services containing 899,000 layers.
- New York Public Library Collection
- David Rumsey Collection
- Many CKAN portals
- Many Thredds servers

Help us improve the system

- Try it out. If you can't find services you know are out there, submit them to us and we will add them.
- If you would like to harvest metadata from other systems, write a connector. We will provide guidance.
- If you would like to bring such search into other applications besides WorldMap, write a client. We will help.
- If you would like to set up your own HHypermap registry instance, the code is available.
- If you have other features you would like, let us know.

THANK YOU!
Questions?

Harvard Hypermap:
An Open Source Framework for Making the World's Geospatial Information more Accessible

Benjamin Lewis, Paolo Corti, Wendy Guan

Center for
Geographic Analysis

Harvard University

http://worldmap.harvard.edu