

Automating Tabular Data in a Geographic Context While Utilizing the GeoCommon's Platform

FGDC NSDI CAP Grant

Award Number G09AC00107, Category 4: Enabling Use of Government Tabular Data in a Geographic Context

Interim Report – November 2009

Executive Summary

FortiusOne has been extending the public GeoCommons geospatial data sharing and visualization portal to enable users to easily upload and georeference tabular data. This includes both geocoding of street-level address data as well as referencing with datasets by a common foreign key. In the first phase, FortiusOne developed and open-sourced a geocoding engine that is currently in use on GeoCommons public website. The next phases will continue to develop additional join capabilities across arbitrary data attributes as well as add industry standard service interfaces to data and maps. To validate and demonstrate the georeferencing capabilities, FortiusOne will georeference datasets from Grants.gov, USASpending, and FCC.

Project Narrative

There exist large amounts of data in tabular formats that contain a geographic reference but do not include geospatial coordinates. The goal of this project is to build and provide an automated and simple interface for users of the GeoCommons platform to georeference and share geographic data. The project was separated into two primary phases, geocoding address data, and geojoining against datasets based on a foreign key.

In the initial phase of the CAP grant, FortiusOne developed a modular geocoding engine that can accept either structured queries or free string entries and provide a list of street level, or less accurate, geographic location matches and estimates of quality. This engine was open-sourced with the purpose to more broadly promote the ability of tools to provide georeferencing.

The code for this project can be found at <http://github.com/geocommons/geocoder>.

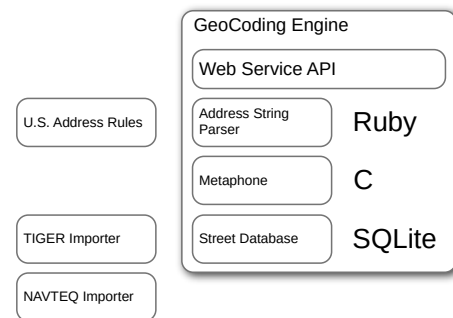


Figure 1: FortiusOne GeoCoder Modular Architecture

The first version of the geocoder only supports US street address parsing and uses the US TIGER/Line road data. It currently only supports typical geographic entities such as countries, cities, address ranges. In developing the geocoder, there are many special cases in addressing schemes for US data. In addition, there is not a publicly available, global address database to use for world wide geocoding. One potential solution is the growing OpenStreetMap project that is actively adding addressing in many countries.

The FortiusOne geocoder was integrated into the free, public GeoCommons website so that anyone can upload and georeference tabular data. As part of this work, interaction design and usability testing developed a step-by-step walkthrough of uploading data from spreadsheets, syndicated feeds, or other formats, specifying geographic identifier attributes, and reviewing of the georeferenced data. All uploaded data is made available through the GeoCommons Finder! application and can be visualized and analyzed with other datasets using GeoCommons Maker!

A GeoCommons portable server was also deployed to Jalabad, Afghanistan to support the data sharing and collaboration between multiple NGO's, US Agencies, and other organizations. The FortiusOne Geocoder was loaded with 30,000 Afghanistan village names and used to geocode tribal information, citizen information, and situation reports leading up to and during the elections in August 2009.

USAID's Global Development Commons, JICA, UNHCR, UNDP, and Ushahidi also collaborated to share information. All data that was shareable was dynamically federated daily from the Jalabad server to the public GeoCommons data portal and multiple visual analyses were shared at <http://news.geocommons.com/afghanistanelection09> and featured by BBC, New York Times, Washington Post and other international media outlets.

There is a growing community of developers that are beginning to build parsers for other addressing schemes and languages for the geocoder, such as the GeoNames database as well as the OpenStreetMap project.

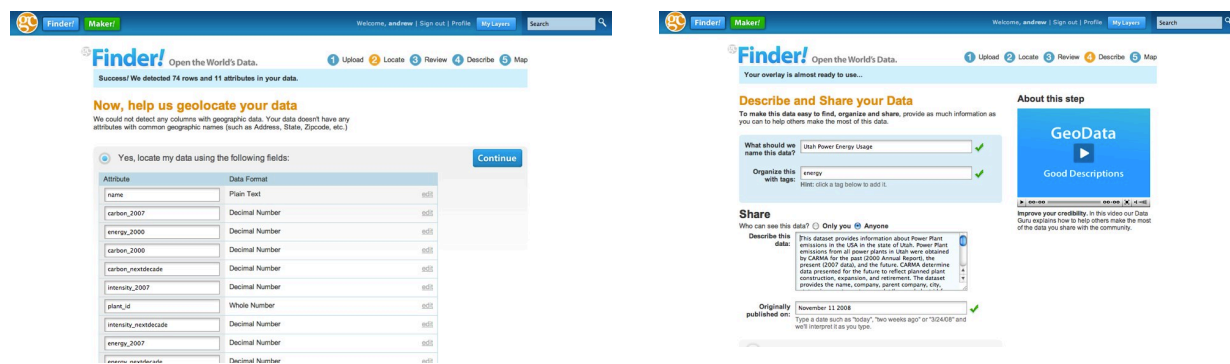


Figure 2: Geocoding Upload Steps

Initial development work has been performed on the join functionality. This includes an interface for choosing searching through existing datasets in Finder and common column identifiers. The datasets are then joined and the georeferenced dataset is then stored and shared publicly.

Next Steps

With Phase 1 complete in geocoding data, the next steps include finishing the join functionality for public GeoCommons and building into the FortiusOne platform. Part of this work will include adding 'fuzzy' matching of columns in order to handle slight differences in columns such as shortened names and misspellings (e.g. "Yosemite" and "Yosemite National Park").

The next phase will also include building Open Geospatial Consortium (OGC) specification interfaces for search and

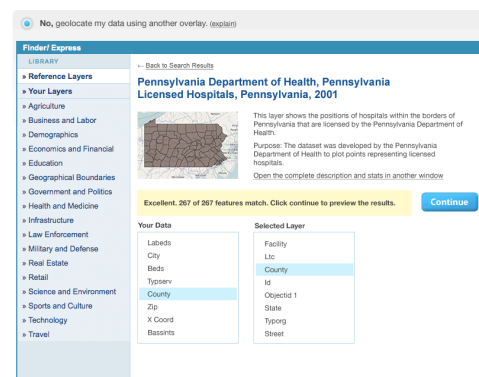


Figure 3: Geographic Join Interface

federation. This includes Catalog Services for the Web (CSW) and Web Feature Services (WFS). Combined with the existing OpenSearch-Geo and Atom syndication, these services will allow for easy federation between multiple GeoCommons servers and other data repositories such as Data.gov and Geospatial One Stop (GOS).

FortiusOne is currently seeking assistance in obtaining datasets and feeds to test and demonstrate georeferencing data. These example datasets will illustrate the potential applications and benefits of geospatial government information. In addition, FortiusOne is discussing how the technology built in this CAP grant can support providing data to Data.gov and Geospatial One Stop in easy to use, open data formats.