Linux.com

Everything Linux and Open Source

OpenStreetMap project imports US government maps

October 11, 2007 (9:00:00 PM) - 2 years, 9 months ago

By: **Nathan Willis**

OpenStreetMap (OSM) is a collaborative project in the process of building a free, Web-accessible, user-editable map of the world. So far, most of its map data has come through user-contributed GPS traces, but OSM has recently undertaken the bulk import of government-collected data covering the entire United States. The massive import will jump-start OSM's US map coverage, but its sheer size poses an interesting challenge to the project's resources.

The traditional method of importing data to OSM is **outlined** on the project's wiki: Individual users collect GPS traces, convert the tracks to OSM form, label any objects, then upload the result to the OSM database. Map images themselves are rendered separately, as needed.

Since OSM's inception in 2004, the majority of the contributed data has been located in Western Europe, as is evident from the **full globe** at OSM's map site.

Several factors have combined to keep the United States relatively unmapped. OSM's founders live in Europe, so initial participation was higher there. Also, the considerably higher population density in countries like the United Kingdom meant that far more area could be mapped there with the same manpower as in less dense nations. Finally, a high-quality, comprehensive set of public domain maps already exists for the United States, making the GPS-trace method unappealing to volunteers.

That set of maps is the US Census Bureau's **Topologically Integrated Geographic Encoding and Referencing** (TIGER) system. The Census Bureau maintains the TIGER database to assist in its various mandated programs, including the decennial US census. Because the TIGER database is built with public funding, it is by law in the public domain.

TIGER contains the locations of nearly every street, highway, railroad, body of water, and legal boundary in the US. It is built from a combination of original US Geological Survey and Census Bureau maps, updated with data collected by Census Bureau staff while in the field. Though not perfect, it covers the entire US in detail.

Can't get there from here

The TIGER data has long been a tempting target for OSM. Being in the public domain, its use is not restricted by licensing as are most commercially available maps. And it is available electronically, in vector format, so if a suitable conversion utility were available, it could be converted automatically to OSM's format.

A bulk import of TIGER data was **attempted** in 2005, but the initial trials **failed to produce quality results** and the work was abandoned. In the spring of 2007, though, Brandon Martin-Anderson and Dave Hansen undertook a **brand new effort**, hunting down bugs in the previous conversion and import code, and starting fresh.

Martin-Anderson had written TIGER parsing code before, and with some help from other OSM developers worked it into a TIGER-to-OSM conversion script. The time-consuming part, he says, was mapping attributes from one form to the other. "People have a lot to say about how various TIGER tags are converted to OSM tags -- whether an A-class TIGER road is 'residential' or 'unclassified,' et cetera. I spent a great deal more time working out the tag conversion with other members of the community (a completely non-technical task) than writing software."

Once all involved were happy with the **attribute mapping**, Hansen downloaded the entire TIGER data set from the Census Bureau Web site in county-sized chunks, and ran the conversion script on his home computer. The resulting OSM-compatible data set consisted of 379,836,373 objects -- nodes, segments of streets, and so on.

Converting the data took several days of constant work, but it still needed to be uploaded to the live OSM server. In a postmortem of the attempted 2005 import, Hansen discovered that some of that effort's problems were the result of trying to import the converted data directly into the database. "There were very few controls on what exactly was uploaded, and I believe that some problems cropped up with its integrity. Parts were uploaded twice, or a point was mistakenly missed in the upload, then later referenced by a segment."

For reliability, Hansen initially began uploading the newly converted data through **JOSM**, the client-side application typically used for annotating and uploading GPS traces. This ensured that the TIGER data went through the same API as any other OSM input, averting the breakage associated with the 2005 attempt.

Mark your calendars: May '08

Although this import method was accurate and safer than dumping data directly into the database, it was agonizingly slow. Given that the complete TIGER data set was 20 times larger than the entire non-TIGER OSM collection, Hansen **calculated** that it would take between five and 10 years to complete the upload at the speeds he was getting through JOSM.

Eventually the OSM team devised a better plan. Hansen transferred the already converted data files to a development machine on the same rack as the OSM map server, and admin Tom Hughes dedicated three of the map server's 12 import daemons solely to the bulk upload. The improved data import started in early September.

Running night and day, seven days a week, the TIGER import should be completed in May or June of 2008. A **public Web page** keeps track of the stats, including the current throughput, percentage of the TIGER data imported, and a list of the completed counties. The curious can monitor the import's progress and compare it to the **combined stats** for the entire OSM database.

Although the improved TIGER import is far faster than the old, there is still a long time to wait before it finishes. Hansen came up with a way to make the wait less painful. He began by sorting the counties in the upload queue by population, so the most populated areas go first. Plus, he takes requests. If you want your county or counties imported next, **email Hansen** and he will bump them up in the **queue**.

A lesson in scalability

But time is not the only important factor. Shortly after starting the TIGER import, it became clear that the database machine itself would run out of disk space within a matter of weeks.

Hughes added additional storage capacity on September 27, and says he is prepared for more complications to crop up along the way. "Of course we've only loaded 10% of the Tiger data so far, and who knows what further challenges we will encounter as more of the data is loaded and as more of the rest of the world is mapped."

One such example purely on the software end is the size of the database index. The team had known for a long time that the existing index was inefficient, Hughes says. The database indexed all of its entries by their latitude and longitude, requiring the lookup of thousands of double-precision floating point values for any given geographic area.

Once the TIGER import began, though, the number of indices shot up dramatically, and the once-theoretical inefficiency of the system rapidly became a concrete problem. Luckily a solution was already in the works, and last week the database switched over to a **new index** dividing the globe into discrete tiles and putting far less strain on the server.

The chicken and the egg and the tiger

Although bulk-importing a data set like TIGER represents a big departure from the methodology used by OSM in the past, Hansen and Hughes both believe it is in the project's best interests. Hansen says, "The reality is that people have been told for years not to map too much in the US because "the TIGER upload will obviate the need for your work." That has kept mappers away. So, first of all, I think we have a duty to actually go do TIGER data after we've been turning some mappers away for quite some time."

The presence of the TIGER data, he continues, does not obviate the need for volunteer help. "What I find myself doing with the TIGER data is taking my traces and dragging the existing TIGER streets around to match up with my traces. This saves me the time of taking notes when I drive around, and a *ton* of time finding the street names for everything in my GPS traces. My workflow basically becomes matching up shapes in TIGER with shapes from my GPS. TIGER is a skeleton on which we can build some much better maps, just like GPS traces are a skeleton on which most of OSM has been built up to now."

Hughes describes the bulk import process as presenting an interesting contradiction to OSM volunteers. "Most of us probably like the idea of bringing in lots of data that gives us nice-looking maps, but equally we like getting out and exploring places and gathering data.... That said, most sources of data have some limitations, so getting out and doing a survey on the ground can help improve the quality of the results, even if you're starting from some sort of baseline data from somewhere else."

How to handle large-scale bulk imports is not merely an academic question for OSM. In July, commercial mapping company **AND Automotive Navigation Data** (AND) **donated** a large set of street maps to the project, covering the entire nation of the Netherlands and the major highway networks of India and China.

There are other **map sources** with potential to serve as bulk imports for OSM. Some are in the public domain, others available under licenses compatible with OSM's **policy** of publishing its maps under the Creative Commons Attribution Share Alike license.

And the TIGER data itself includes information that does not currently map into OSM's data model, such as street numbers. Hansen preserved this extra data when converting TIGER, in case it becomes useful at a later point due to expansion of OSM's model. As Hughes observed, properly matching third-party data to OSM's data model is second only to licensing questions as a topic for discussion when examining new potential sources.

Even when completed, the TIGER import will account for only a small percentage of the world's land surface. But more importantly, it will demonstrate the viability of OSM as a free, large-scale mapping solution, and that will benefit the project itself and independent map users alike.

Read in the original layout at: **http://www.linux.com/archive/feature/119493**