ReadWriteWeb

Using Public Data to Fight a War

By Pete Warden / February 7, 2011 4:30 PM / 2 Comments

How does a technology built for apartment-hunting end up being evaluated by the U.S. Army for use in Afghanistan? Cazoodle *(http://cazoodle.com)* is using public data sources like Flickr *(http://flickr.com/)* and OpenStreetMap *(http://openstreetmap.org/)* to build detailed guidebooks for American soldiers. Last week at Strata *(http://strataconf.com/strata2011)* I sat down with company CTO Govind Kabra to find out how they do it.

Its project for the Army is to build a detailed database of information about places in Afghanistan, using only public sources on the Web. The goal is to describe in detail the towns and cities including everything from names, locations and populations, as well as lists and coordinates for schools, mosques, banks and hotels.

The military already collects this sort of information, but using traditional offline sources through groups like the National Geospatial-Intelligence Agency *(https://www1.nga.mil/Pages/Default.aspx)* . It's a slow and dangerous process to send personnel door to door for research within war-torn countries, and though the agency's budget is classified, presumably very expensive. The hope is that by using online, crowdsourced data from sites like Wikipedia and Flickr, it will be possible to gather rich information without putting lives at risk, all at a fraction of the cost.

Origins

Cazoodle was started four years ago at the University of Illinois - Urbana Champaign. Kubra and his co-founders were graduate students, so naturally the top of their priority list was finding a cheap apartment.

As they trawled through Craiglist, following links to other sites, consulting maps and looking up details, they realized that what they really needed was an automated way of pulling the information they cared about from all these disparate sources, and putting it into a single spreadsheet they could use to make their decisions easier. They formed the company to build this system, and created an apartment search engine based on the technology *(http://www.cazoodle.com/apartment-search.php)* .

The founders knew there were lots of other problems that would also benefit from the same underlying technology, so they branched out into shopping *(http://www.cazoodle.com/shopping-search.php)* and vacations *(http://vacation.cazoodle.com/)* , and also started building custom search engines for enterprise customers.

That was when they spotted a Small Business Innovation Research grant opportunity *(https://www.fbo.gov/index?s=opportunity&mode=form&id=42545f1cf87af61b648c1c85a8c56303&tab=core&_cview=0)* from the U.S. Department of Defense. The task was to curate public information on the Web related to Afghanistan into a single database that Army personnel could use to guide their operations. Their technology already took a soup of unstructured Web pages related to locations and converted it into a spreadsheet of data, cleanly split into labeled columns, so it seemed like a natural fit for this problem.

Technology

To understand how it works, imagine trying to create a list of mosques in a small town in Afghanistan. There's no handy Yellow Pages you can refer to, and the maps don't have that much detail. However, if you go to Wikipedia you can pull out basic information about a town like Pul-i-Alam, and then look through the OpenStreetMap data for Afghanistan to spot locations that are tagged as religious buildings, eg:

```
<node id="282153330" lat="34.5154772" lon="69.1804459">
<tag k="amenity" v="place_of_worship"/>
<tag k="name" v="Puli Khishti Mosque"/>
<tag k="religion" v="muslim"/>
</node>
```

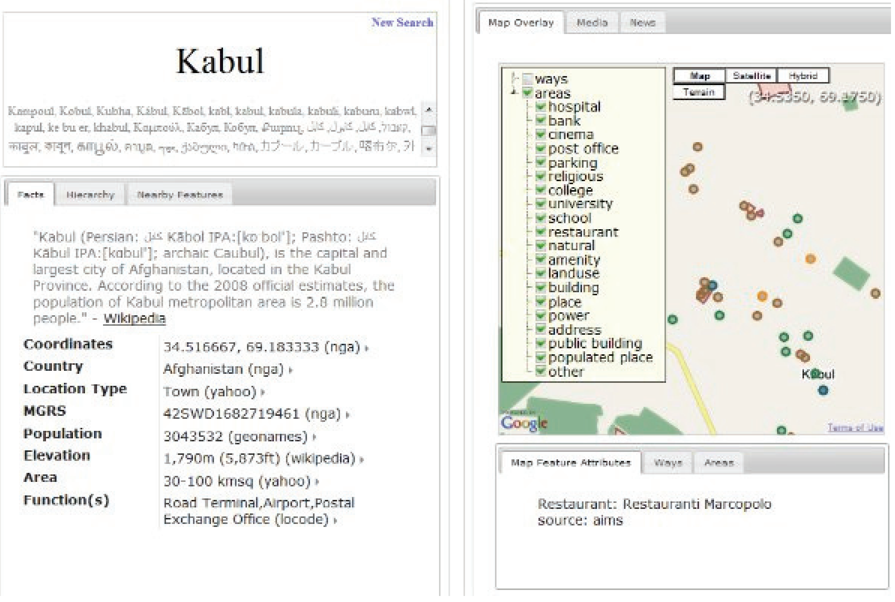
Dawlatabad
Maz
Sarif
Holm
Sulgara

(http://www.flickr.com/photos/tags/mosque/map?&fLat=35.4338&fLon=67.489&zl=10) That reveals the explicit information that people have entered, but what's particularly impressive about Cazoodle's work is that it also merges in implicit information from sources like Flickr. For example, running a search on the photo service *(http://www.flickr.com/photos/tags/mosque/map?&fLat=35.4338&fLon=67.489&zl=10)* shows hundreds of photos taken within Afghanistan mentioning "mosque" in their descriptions. The coordinates can be pulled out of the geotagged photos, and used as an input to the list of mosques for the town they were taken in.

Without realizing it, photographers are helping to build up a crowd-sourced map of everywhere they shoot. This isn't completely unprecedented; during World War Two the BBC appealed for holiday photos of the beaches of Normandy for an exhibition. In fact, the 9 million snaps received were used to research landing sites for the coming invasion.

Results

The end result of the gathering process is a console that Army personnel can use to pull up information on towns across the country, giving a detailed breakdown of all the data that's been gathered on the location:



Kabul

Kanpoul, Kobul, Kubha, Kâbul, Kâbol, kabl, kabul, kabula, kabuli, kabuni, kabwl, kapul, ke bu er, khabul, Καμπούλ, Кабул, Кобул, Քաբուլ, کابل, کابل, کابل, קאבול, काबुल, কাবুল, காபூல், കാബൂൾ, ..., ქაბული, ካቡል, カブール, カーブル, 喀布尔, 카

"Kabul (Persian: کابل Kābol IPA:[kɔ bol']; Pashto: کابل Kābul IPA:[kɑbul']; archaic Caubul), is the capital and largest city of Afghanistan, located in the Kabul Province. According to the 2008 official estimates, the population of Kabul metropolitan area is 2.8 million people." - Wikipedia

Coordinates	34.516667, 69.183333 (nga) ▸
Country	Afghanistan (nga) ▸
Location Type	Town (yahoo) ▸
MGRS	42SWD1682719461 (nga) ▸
Population	3043532 (geonames) ▸
Elevation	1,790m (5,873ft) (wikipedia) ▸
Area	30-100 kmsq (yahoo) ▸
Function(s)	Road Terminal,Airport,Postal Exchange Office (locode) ▸

Restaurant: Restauranti Marcopolo
source: aims

What I find most fascinating about this project is that it's the first practical application for 'linked data'. In contrast to the more academic approach *(http://linkeddata.org/)* , Cazoodle has attacked the problem in a much messier but more pragmatic way. A good example of this is how the system uses "fuzzy matching" to link data from different sources together.

That means if OpenStreetMap shows a mosque in a particular location, and a Flickr photo with coordinates a hundred yards away mentions a mosque in the description, then it's reasonable to assume they represent the same place, even though there's a small probability that's incorrect.

This means data may not be quite as vetted as a more traditionally sourced gazetteer, but it has much broader coverage and is far more dynamic. In many ways, it's like the tradeoff between Yahoo's original hand-edited directory and Google's chaotic but all-encompassing search index. By lowering the barriers to data entry, and in many cases using public information people don't even realize they're revealing, Cazoodle is able to create an effective guide.

The project is still being evaluated by the Army right now, and hasn't been used in the field, but it's not hard to imagine this approach becoming far more common as public data sources grow and multiply. It also illustrates the conflicts we'll face more and more frequently as this public data is used for completely unintended purposes. How will local photographers and OpenStreetMap editors feel if their work is reused by the U.S. Army? Christopher Albon *(http://christopheralbon.com/)* , a researcher into public health in warzones *(http://conflicthealth.com/)* also has some cautionary words on the limits of what can be done remotely:

> While an impressive start, Cazoodle's approach is missing the data that really matters. A map of a physical space only takes you so far. An Afghan village is no more a collection of mosques and houses than Silicon Valley is a collection of coffee shops and office space. What matters is a location's social, political, and economic structures; its human terrain. Who is related to whom? Who owns the fertile fields by the river or the rocky fields on the slopes? Who is healthy and who is sick? Cazoodle can not provide this type of information, leaving American soldiers to gather it the old fashioned way: talking to people door to door, face to face.

Photo by James Vaughn *(http://www.flickr.com/photos/x-ray_delta_one/4769639013/)*

179 Like 5 people like this.

Showing 2 comments

Sort by Newest first Subscribe by email Subscribe by RSS

Maksym 1 hour ago

If you're inetrested in geolocated photos for any given area, you can also try Geolocation.ws
Here is a map with Creative Commons and Public domain photos around Kabul:

http://www.geolocation.ws/map/...

Site pulls data from Freebase (another good source of stuctured open data with good search API available), Panoramio, Flickr and Wikimedia Commons.

There is also DBPedia, which provides data parsed from Wikipedia infobox templates

Like Reply

Ben 1 hour ago