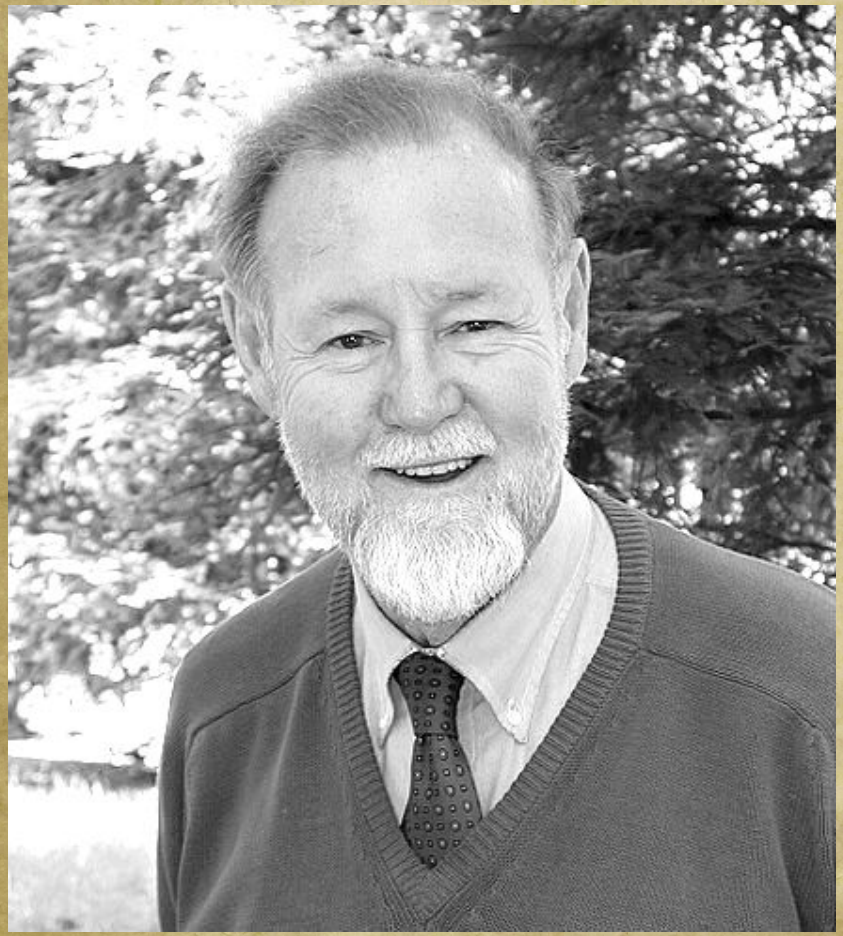
Befriending a Geocoder

Written by Diana Shkolnikov

Illustrated by The Internet



Our story begins a long time ago, in the 1960's, when the first Geographical Information Systems began to take shape in the faraway land of Canada.

Dr. Roger Tomlinson, a brilliant man, created a system that would catalog data pertaining to agriculture, wildlife, and forestry. He has since been acknowledged as the father of GIS.

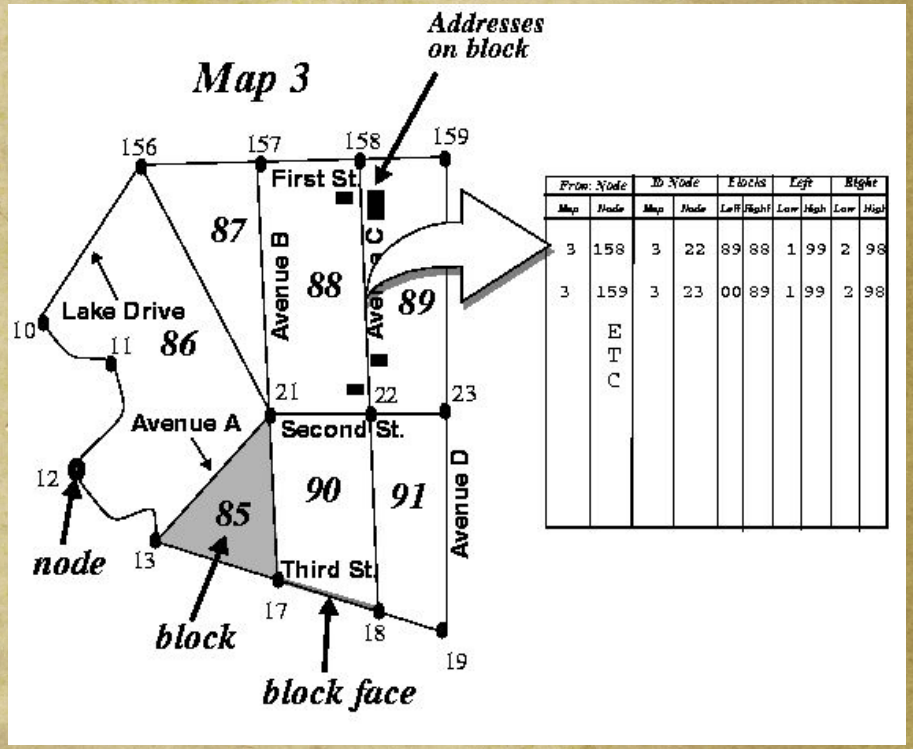
Dr. Roger Tomlinson

https://en.wikipedia.org/wiki/Roger_Tomlinson

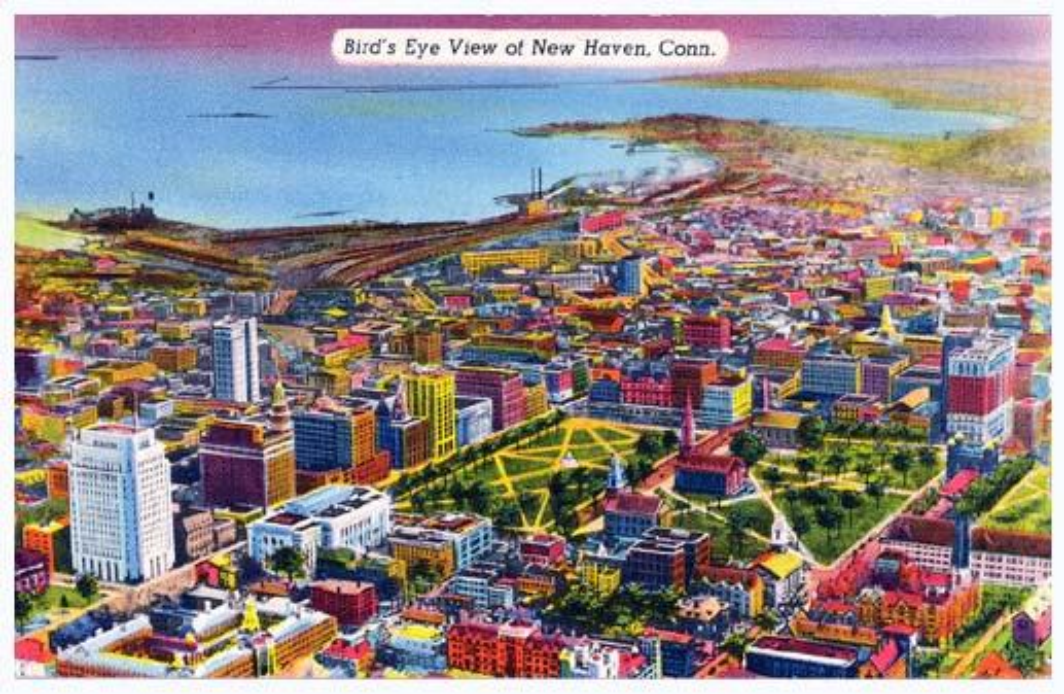
DIME/GBF System

Shortly thereafter some of the finest universities in the US, such as Harvard and Yale, followed suit and implemented their own versions of early geospatial search engines. A team of Yale graduates and students developed a protocol they called the Dual Independent Map Encoding, DIME for short.

This groundbreaking protocol paved the way for geocoding algorithms still used in some of today's most popular commercial geocoders, such as Google and MapQuest.



http://guides.library.yale.edu/GIS/geocoding
http://www.geog.ucsb.edu/~kclarke/G128/Lecture04.html
https://en.wikipedia.org/wiki/Dual_Independent_Map_Encoding



FUN FACT

New Haven, Connecticut, the home of Yale University, was the first city on Earth with a topologically integrated, geocodable, streets network database.

https://www.census.gov/history/www/innovations/technology/dual_independent_map_encoding.html
http://guides.library.yale.edu/GIS/geocoding

1968 Census Bulletin

THE FINISHED PRODUCT--George Farnsworth and Fadra Lewis, both of the Census Use Study office here, discuss features of a DIME or computer drawn map of the city of New Haven. Census information was tabulated and then translated into map form, showing, in this case, the geographical distribution of births in the city during one twelve-month period.

DIME Underwent Lots Of Testing Too

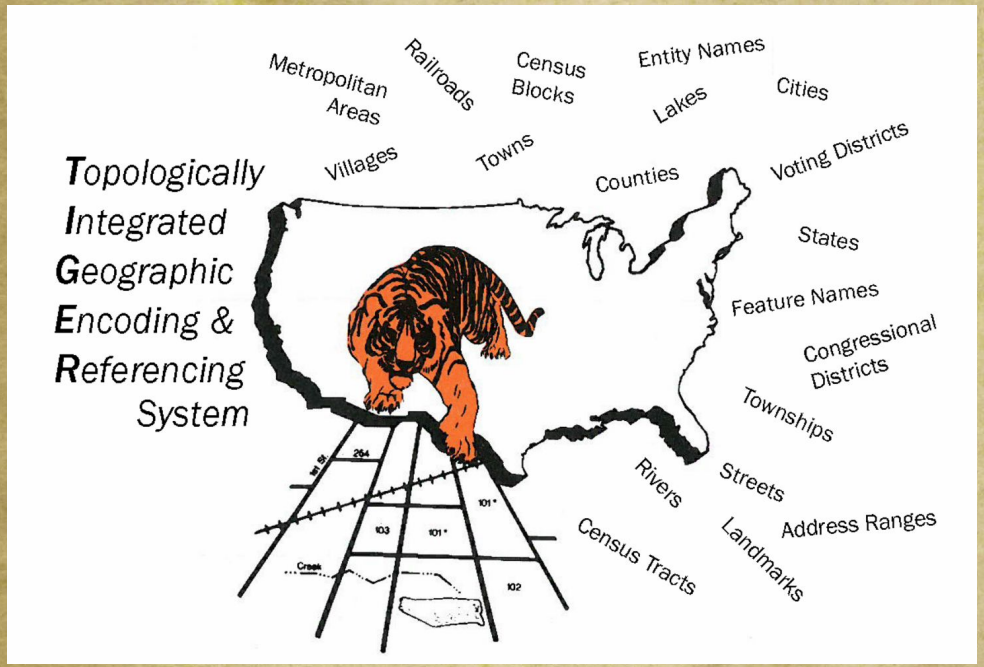
With preparations for the 1970 Decennial Census well underway, all Censusites are aware of the several recent tests and dress rehearsals involving enumeration and mail census procedures and data

numbers and the street address ranges on each side of the street segment is recorded on punch cards. Map coordinates are then assigned to the punch cards. The cards are edited by computer and the computer

Using US Census data from 1970 and 1980, the first automated system to store and retrieve city address data using city blocks and house number ranges was built. This allowed the user to compute the location of an address along the face of a city block.

https://www.census.gov/history/pdf/1968censusbulletin-dime.pdf
https://en.wikipedia.org/wiki/Dual_Independent_Map_Encoding

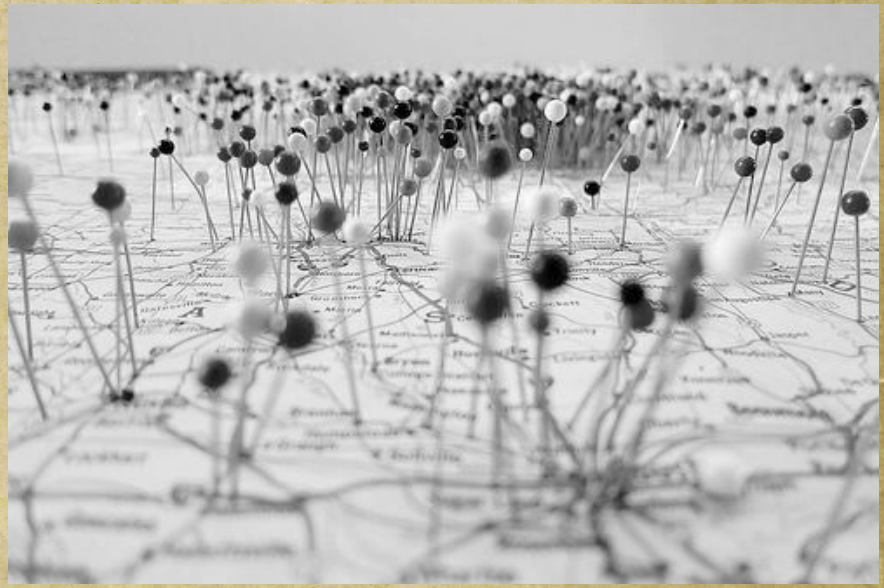
It became clear that there was a tremendous market for these amazing systems and commercial geocoding solutions began to emerge in the 1980's. Around that time a little company named ESRI began to offer commercial geocoding services.



http://census.maps.arcgis.com/apps/MapJournal/index.html

By the end of the century, with the help of the 1990 US Census, there was enough street segment data collected to generate a brand new, never before possible data set for interpolation of street addresses. This new set of data was named TIGER. TIGER was built on top of the principle laid out by the DIME system and quickly became the foundation of every serious US geocoding solution in existence. This dataset continues to be in use well into the present time.

Fast forward to the early 2000's the geocoding wizards began to focus on parcel and rooftop data. Combining the accuracy of parcel centroids with the flexibility of TIGER ranges allowed for a whole new level of precision.



http://link.highedweb.org/2012/11/location-based-services-where-are-we-going-next/map-pins

Those in possession of this amazing technology could easily locate an address on a map, which at that time was pure magic to the general population! These tools granted great power, but they were limited to a select few users, leaving majority of people at the mercy of crinkled up old maps and sketchy directions on the back of a napkin.

Luckily things began to quickly expand to the masses as personal computers and mobile devices became commonplace. The technology matured and commonfolk began to enjoy the benefits of these game-changing systems.

https://en.wikipedia.org/wiki/Alexander_(magician)

Years after the first strides in geocoding made the people believe in a magical future where any location in the world is accessible with just a few keystrokes, all the serious geocoding platforms were still owned and operated by large corporations and government agencies. This was mostly due to the fact that the data required to build a reliable geocoder was an insurmountable obstacle, requiring years of data collection in the field or millions of dollars to license such data from others.

The people were also limited in how they could share the information they retrieved using these mysterious geocoders. If the answers granted to them by the geocoders were incorrect, the people were powerless to improve them. Despite the restrictions, the geocoders were so powerful and great that the people went along with the rules and demands, for lack of alternatives.

https://en.wikipedia.org/wiki/Alexander_(magician)

Fortunately a seed was planted in 2004 by a visionary named Steve Coast. He saw a future where world wide geospatial data belonged to the very people that it aimed to represent. A project called OpenStreetMap was brought into the world with humble beginnings. By 2013 a remarkable milestone of 1MM contributors had been reached. That's 1MM individuals working towards the common goal of a free people's dataset. Together they accomplished something astonishing and the momentum continues to build as our story unfolds.

Once OpenStreetMap data reached a level of coverage that could be used to effectively geocode, even if only in the more populated parts of the world, several experts decided it was time to start making geocoders for the people based on the only dataset composed by the people. This saw the rise of a few prominent open source geocoding solutions, like Nominatim, Komoot's Photon, and recently Mapzen's Pelias, to name a few.

http://www.antiqueimages.blogspot.com

And the people rejoiced, for finally they had several geocoders of their own that they could nurture and train. But these geocoders weren't perfect and required much love and attention. So the dedicated OpenStreetMap contributors wanted to know what they could do to help improve these communal geocoders. What will it take to make these magical systems work like those well established wizard-owned ones. Can they be trained? Do they respond to positive reinforcement? Do they work for treats?

What geocoders thrives upon is point data. As history proved, adding parcel and rooftop point data to geocoding solutions dramatically improves their accuracy and reliability. Point data generally means one of two things: addresses or venues.

http://www.mygenealogyhound.com/vintage-postcards/vintage-independence-day-postcards/Independence-Day-Fourth-of-July-vintage-postcard.jpg

Although OpenStreetMap does not have much in the way of addresses, it offers tremendous value through its rich venue data! Places like restaurants, banks, schools, libraries, and public transit stops are really important in making a well-behaved geocoder. So to help train the geocoder to fetch those places by name and category, they should be added to the OSM dataset with proper tags. Some examples of tags that geocoders ingest regularly are amenity, building, shop, cuisine, public_transport, man_made, landuse, craft, and tourism, to name a few. If any of these tags are found on a record and there is also a name tag, the record gets added to things the geocoder will respond to correctly in the future. Geocoders are really predictable in this respect; they will only ever ingest things marked with tags it understands and knows how to classify.

http://blaurich.jalbum.net/Vintage%20Sign%20Co.%20Additional%20Images/General%20Store/slides/WestmgrocCo_Black.html

AGS

amenity
building
shop
office
public_transport
cuisine
historic
man_made
landuse
railway
sport
natural
leisure
waterway
aerialway
aeroway
craft
military
tourism

addr:housenumber + addr:street

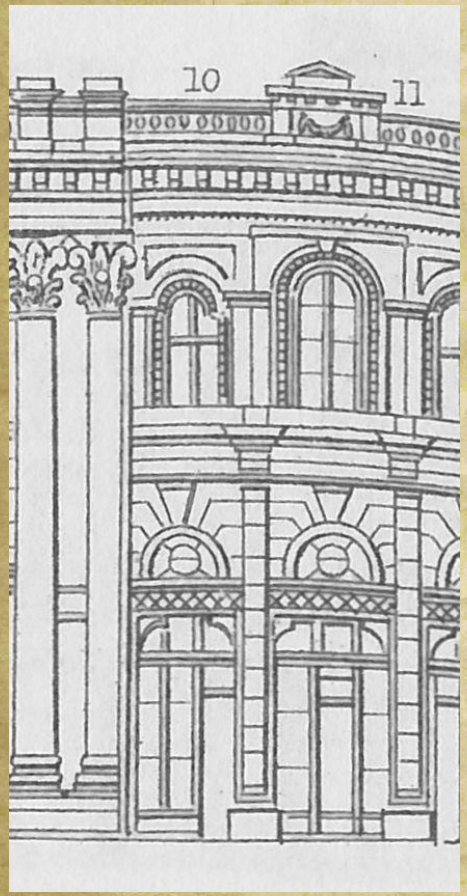
That brings us to another critical point: names are a big deal when geocoding! Unless it's an address, if no name is specified for a venue, that venue cannot be added to the list of things a geocoder knows about. That which has no name cannot be searched for and found successfully.

Another helpful training tip is to add alternate names for venues that might have them. These names could be in different languages, or just nicknames people often use to refer to a place. Basically the more names you can teach the geocoder to recognize a place by, the better it will become at fetching that place, no matter which one of the names you've decided to use in your request.

https://hawkebackpacking.com/thailand_chiang_mai.html

https://whitelocust.wordpress.com/2012/01/18/how-do-you-say-macrib-in-chinese/

https://en.wikipedia.org/wiki/McDonald%27s_Israel

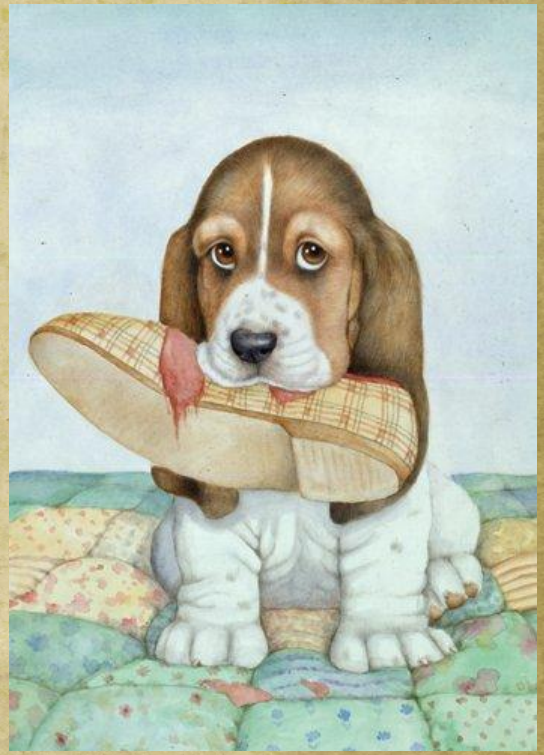


So if there were loads of venues in OSM and each of them had a name, the geocoder would be pretty obedient when searching for venues by name, but there would still be a problem when searching for addresses. Addresses are notoriously tough to add to an open dataset and so it's not realistic, even for a fairy tale, to expect that people would add all residential addresses manually. However, if a small step in that direction could be taken by ensuring that every venue has an address associated with it, progress would be significant.

As it stands, only a little over 30% of the cuisine venues have address tags, for example. Most restaurants have their addresses prominently displayed on doors, menus, and business cards, making it easy to look them up. Getting into the habit of adding those whenever a new venue is created, or updating existing records to contain an address would show the geocoder enough address points to make search results significantly better.

Still, even with all these venues and addresses, sometimes geocoders can get confused and misunderstand your requests, so you end up with results you didn't ask for; kind of like throwing a stick to a dog only to have it fetch you your slippers. It's trying to be helpful, but it's still just missing the mark.

This can be happening for any number of reasons, but to help avoid some confusion try the following tip: avoid abbreviations whenever possible. Geocoders get stumped by ambiguous abbreviations such as St, since it can stand for Street and Saint, or CT, which could mean either Court or Connecticut. If there is a directional in the name or street, it's best to spell the whole thing out, like North instead of N.



http://pussycatdreams.centerblog.net/3024734-Belles-images

Having these simple tips helped the people improve the data. Years into the future, all of their efforts paid off big and the people's data reached a level of coverage and accuracy that made geocoding with anything else a ludicrous proposition. Several loyal, dependable, and ever so magical geocoders emerged and served the people well. Mapzen's Pelias, being one of the first and by that time most established geocoder of the people, lead the pack onward and upward. The people were grateful and excited to have finally befriended a geocoder, and they all lived happily ever after.

https://moviesbeforeandafter.wordpress.com/2015/06/26/theres-no-place-like-the-homesman-a-real-western-in-the-strangest-sense/

@dianashk
diana@mapzen.com

@mapzen
mapzen.com

Come to our workshop tomorrow morning!!!

Info: mapzen.com/search
Docs: mapzen.com/documentation/search
Github: github.com/pelias/pelias
Blog: mapzen.com/tag/search

References

https://en.wikipedia.org/wiki/Geocoding
http://wiki.openstreetmap.org/wiki/History_of_OpenStreetMap
https://en.wikipedia.org/wiki/Roger_Tomlinson
https://www.census.gov/history/pdf/1968censusbulletin-dime.pdf
http://guides.library.yale.edu/GIS/geocoding
http://taginfo.openstreetmap.org/keys/cuisine#combinations