<http://franklincenterhq.org/2541/geocoding-addresses-from-missouri-sex-offender-registry/>

\FranklinCenter\Projects\2011\KSHB-Sex-Offenders

**Geocoding addresses from Missouri Sex Offender Registry**

By Earl Glynn Tuesday, July 19th, 2011

This is the second of four articles about analyzing distances between sex offenders and child daycare centers in Missouri as part of a [*joint project with KSHB NBC Action News in Kansas City*](http://www.nbcactionnews.com/dpp/news/local_news/investigations/loophole-in-law-allows-hundreds-of-missouri-sex-offenders-to-live-near-church-daycares).

The previous article gave [*details of analyzing the Missouri Sex Offender Registry and preparing a data file for geocoding*](http://www.franklincenterhq.org/2536/analysis-of-missouri-sex-offender-registry-data/).Our goal was to compute distances from Missouri registered sex offenders to child day centers in Missouri, and in particular church daycare centers.  KSHB’s recent report explained that a [loophole in law allows hundreds of Missouri sex offenders to live near church day cares](http://www.nbcactionnews.com/dpp/news/local_news/investigations/loophole-in-law-allows-hundreds-of-missouri-sex-offenders-to-live-near-church-daycares).

To compute distances we needed the latitude-longitude coordinates from the offenders and the child daycare providers. But how do we convert street addresses to geolocation coordinates?

The process of converting addresses to latitude-longitude coordinates is called “geocoding” and Google provides an application program interface (API) for this.  The [Google Geocoding API](http://code.google.com/apis/maps/documentation/geocoding/) page provides many details of how this process works. Also note [Google has a number of restrictions on how this works](http://code.google.com/apis/maps/terms.html#section_10_12).

We used a similar geocoding process described below on both lists of offenders and childcare centers.

**Geocoding Example**

Let’s look at an example to understand how geocoding works.

One of the Missouri registered sex offenders lives at this address:  5007 S Fuller Ave, Independence, MO.  How can we use Google Geocoding to get the latitude-longitude of this address?

To geolocate this address simply enter the following URL into a web browser (it should be a single line URL):

<http://maps.googleapis.com/maps/api/geocode/xml?address=5007+S+FULLER+AVE,+INDEPENDENCE,+MO&sensor=false>

Note in the URL each space in the address was replaced with a “+”.

Google returns the following XML code for the address.  The fields we want to extract are shown in red:

<?xml version="1.0" encoding="UTF-8" ?>  
- <GeocodeResponse>  
< status>OK</status>  
- <result>  
< type>street\_address</type>  
< formatted\_address>5007 S Fuller Ave, Independence, MO 64055, USA</formatted\_address>  
- <address\_component>  
< long\_name>5007</long\_name>  
< short\_name>5007</short\_name>  
< type>street\_number</type>  
< /address\_component>  
- <address\_component>  
< long\_name>S Fuller Ave</long\_name>  
< short\_name>S Fuller Ave</short\_name>  
< type>route</type>  
< /address\_component>  
- <address\_component>  
< long\_name>Independence</long\_name>  
< short\_name>Independence</short\_name>  
< type>locality</type>  
< type>political</type>  
< /address\_component>  
- <address\_component>  
< long\_name>Blue</long\_name>  
< short\_name>Blue</short\_name>  
< type>administrative\_area\_level\_3</type>  
< type>political</type>  
< /address\_component>  
- <address\_component>  
< long\_name>Jackson</long\_name>  
< short\_name>Jackson</short\_name>  
< type>administrative\_area\_level\_2</type>  
< type>political</type>  
< /address\_component>  
- <address\_component>  
< long\_name>Missouri</long\_name>  
< short\_name>MO</short\_name>  
< type>administrative\_area\_level\_1</type>  
< type>political</type>  
< /address\_component>  
- <address\_component>  
< long\_name>United States</long\_name>  
< short\_name>US</short\_name>  
< type>country</type>  
< type>political</type>  
< /address\_component>  
- <address\_component>  
< long\_name>64055</long\_name>  
< short\_name>64055</short\_name>  
< type>postal\_code</type>  
< /address\_component>  
- <geometry>  
- <location>  
< lat>39.0301740</lat>  
< lng>-94.4355300</lng>  
< /location>  
< location\_type>ROOFTOP</location\_type>  
- <viewport>  
- <southwest>  
< lat>39.0270264</lat>  
< lng>-94.4386776</lng>  
< /southwest>  
- <northeast>  
< lat>39.0333216</lat>  
< lng>-94.4323824</lng>  
< /northeast>  
< /viewport>  
< /geometry>  
< /result>  
< /GeocodeResponse>

The [R XML package](http://cran.r-project.org/web/packages/XML/index.html) provides functions for parsing XML (or in theory one could also parse the JSON default code returned by the Google API using a different R package).

Parsing XML is still a bit of a pain using the **xpathApply** function to extract the various data items.  But to simplify parsing XML returned by Google geocoding, we wrote the function ***get.geocode***, which is in the [GoogleGeocode.R](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/GoogleGeocode.R) file.

Here’s an example of using the R **get.geocode** function to request the same information as shown above:

> d <- get.geocode("IDNumber","5007 S Fuller Ave","Independence","MO","")  
> t(d)  
[,1]  
id "IDNumber"  
status "OK"  
street "5007 S Fuller Ave"  
city "Independence"  
state "MO"  
state.code "Missouri"  
zip5 "64055"  
county "Jackson"  
lat "39.0301740"  
lng "-94.4355300"  
location.type "ROOFTOP"  
formatted.address "5007 S Fuller Ave, Independence, MO 64055, USA"  
result.count "1"

But how can we geocode a file of many addresses, such as a list of Missouri sex offenders or Missouri child daycare providers?

**Batch Geocoding**

The R script [GoogleGeocodeMOSexOffenders.R](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/GoogleGeocodeMOSexOffenders.R) reads a list of addresses from the tab-delimited file **geocode-MO-offender-in.txt**, calls the function get.geocode for each line, saves the information from geocoding, and when complete, writes the results to the file **geocode-MO-offender-out.txt**.

Most files of addresses can be geocoded this way by modifying the script appropriately.

The batch setup process is fairly simple.

Because there can be a conflict in column names, I normally use an uppercase letter as the first letter of each field in the original file and lowercase letters with the geocoded info.  That way ZIP is from the original file and zip is from geocoding.

Add these columns to the file you want to geocode — their exact placement does not matter:

status, zip, county, state, lat, lng, location.type, formatted.address, result.count

Modify the R script, like GoogleGeocodeMOSexOffenders.R, to assign quantities to street, city, state, and zip and pass those quantities to the get.geocode function.  The geocoding results are then assigned to the extra columns you added to your file.

The program uses a while loop to stop geocoding if you reach the daily limit.

**Limit of 2500 geolocation requests per day**

Google imposes a limit of 2500 geolocation requests per day and that’s tied to your IP address.  If somehow you can use more than one IP address, you can get 2500 geocodes per day per IP address.

[Google says this about their limit](http://code.google.com/apis/maps/documentation/geocoding/):

If you exceed the 24-hour limit or otherwise abuse the service, the Geocoding API may stop working for you temporarily. If you continue to exceed this limit, your access to the Geocoding API may be blocked.

My experience is that  Google knows to the minute when you exceeded their limit. Wait 24 hours before continuing.  Google knows.

To restart geocoding on additional days, edit the file **geocode-MO-offender-out.txt** and look down the status column. I normally edit the file in Excel.

Most status values  will say “OK” but the last one before Google quit will be “OVER\_QUERY\_LIMIT”.  Clear this cell in Excel and save the file back to a tab-delimited file.

I normally number the “in” files after they’ve been processed to “in1″, “in2″, etc., and then rename the modified “out” file to be the “in” file for the next batch run.

The 11,000 Missouri offenders took 5 days to geocode this way.

**Approximate  Matches**

Most geolocation assignments work well, but there can be problems.  Study the “[Status Codes” and “Results” on the Google page](http://code.google.com/apis/maps/documentation/geocoding/#StatusCodes).

Sometimes Google will return more than one address match (result.count >0), and sometimes Google will decide the address is in a different state — which doesn’t make sense to me when the state is specified in the request.

To “accept” a geolocation assignment from Google I normally require:

* status is “OK”, and
* state matches the specified state, and
* result.count > 0 (but result.count of 1 is best).

I often use R statements like the following to decide which rows to accept in data.frame “d”:

select <- (d$status == "OK") & (d$result.count > 0) &  
(d$state == STATE)

d <- d[select,]

For the results that I keep, I still often color a “fuzzy” match for approximate assignments where:

* result.count > 1 or
* location.type is not “ROOFTOP”

I find that location.type “ROOFTOP” with result.count of 1 gives an “exact” match, but many approximate matches are still acceptable.

I often use this R statement to define “fuzzy”:

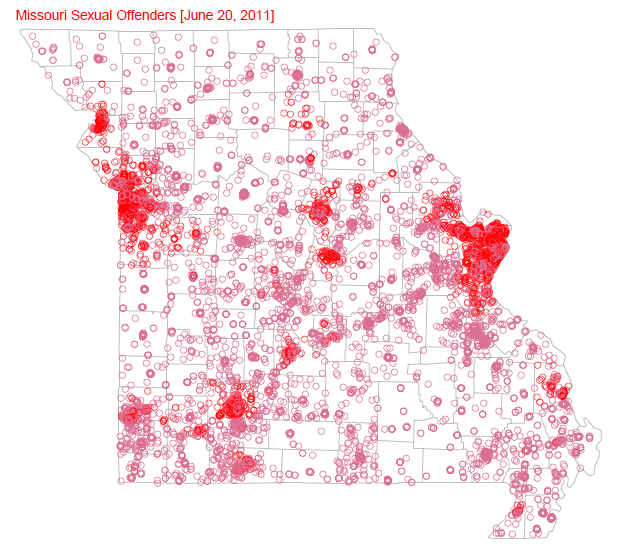
d$fuzzy <- (d$result.count > 1) | (d$location.type != "ROOFTOP")

**R Maps**

While the ultimate goal is to create an interactive Google map using the geocoded results, sometimes non-interactive maps of the results are useful during analysis and can be created using the [R maps package](http://cran.r-project.org/web/packages/maps/index.html).

If  an R object “offender” is a dataframe containing the geocoded results as described above, these R statements will create a static map that uses the “fuzzy” quantity to show better matches as a saturated “red” and fuzzy matches as the less saturated  “palevioletred”.

library(maps)  
map("county", "Missouri", col="grey")  
mtext("Missouri Sexual Offenders [June 20, 2011]", adj=0, col="red")  
points(offender.data$lng, offender.data$lat,  
col=ifelse(offender.data$fuzzy, "palevioletred", "red")



*Note the Google matches are “fuzzy” in many rural areas and are “ROOFTOP” mostly in urban areas.*

**Computing Distances**

Next:  [Computing distance matrix between Missouri sex offenders and child daycare facilities](http://www.franklincenterhq.org/2544/computing-distance-matrix-between-missouri-sex-offenders-and-child-daycare-facilities/)

**Program and Files**

[GoogleGeocode.R](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/GoogleGeocode.R) with function get.geocode

[GoogleGeocodeMOSexOffenders.R](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/GoogleGeocodeMOSexOffenders.R) script to process file geocode-MO-offender-in.txt  
(you will want to modify the basedir directory in line 13 to be your working directory)

Sample input and output files with 10 offenders:

* [geocode-MO-offender-in.txt](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/geocode-MO-offender-in.txt)
* [geocode-MO-offender-out.txt](http://cdn.watchdogmedia.org/national/computer-assisted-reporting/project/geocoding-and-distances/missouri-sex-offenders/geocode-MO-offender-out.txt)

**Related**

* [Loophole in law allows hundreds of Missouri sex offenders to live near church daycares](http://www.nbcactionnews.com/dpp/news/local_news/investigations/loophole-in-law-allows-hundreds-of-missouri-sex-offenders-to-live-near-church-daycares), KSHB NBC Action News, Kansas City, July 18, 2011.
* [Analysis of Missouri Sex Offender Registry Data](http://www.franklincenterhq.org/2536/analysis-of-missouri-sex-offender-registry-data/), Franklin Center CAR Blog, July 18, 2011.
* [Computing distance matrix between Missouri sex offenders and child daycare facilities](http://www.franklincenterhq.org/2544/computing-distance-matrix-between-missouri-sex-offenders-and-child-daycare-facilities/), Franklin Center CAR Blog, July 28, 2011.
* [Displaying Missouri sex offender/child day care facility proximity map using batchgeo.com](http://www.franklincenterhq.org/2563/displaying-missouri-sex-offenderchild-day-care-facility-proximity-map-using-batchgeo-com/), Franklin Center CAR Blog, July 28, 2011.

**Earl F Glynn • KansasWatchdog.org • Franklin Center for Government and Public Integrity** - See more at: http://franklincenterhq.org/2541/geocoding-addresses-from-missouri-sex-offender-registry/#sthash.i76hX5J0.dpuf