GIS Institute Geocoding Lab Exercise
By Jeff Blossom: jblossom@cga.harvard.edu
January 4, 2017

Introduction

The ability to assign specific geographic locations to textual information (the process known as geocoding) is available to anyone with a computer and internet access. The relative ease of geocoding and resulting accuracy can vary widely depending on a number of factors. What is the nature of the data? How 'clean' is it and what format is it in? What geocoding technique will be used? Determining a geocoding strategy that best suits a particular need is not always clear.

The process of geocoding begins with comparing data in text or tabular form to a reference data table in geographic format. The reference table is a dataset that has already been mapped, with established map coordinates. When matches between the input data and the reference data are found, the corresponding map coordinates are assigned from the reference data to the input features, thus geocoding them. A geocoding service is a program that allows for a user to input a batch of data contained in a table, search for matches as compared to a reference table, and output the result in a map or GIS layer format. The key to confidently geocoding data lies in understanding the reference table which the data is being matched to, how a match is found, and the resulting spatial accuracy.

Objective: The goal of this exercise is to give everyone practice performing geocoding using different methods, and to gain insight regarding how to successfully geocode.

Preparing address data for geocoding (best practices)

As a geocoding service compares input data to a table in an existing GIS layer in search of an identical match, it is of critical importance to eliminate **misspellings**, **special characters** (such as " ? # ' \ %), and **abbreviations** in the input dataset. Address data at a minimum should contain the Street Number, Street Name, Street Type, and Street Suffix Direction (if necessary – many street names don't have a Suffix Direction). Most address geocoders work on street intersections as well, and the intersecting Street Names and Street Types separated by an **&** in one field is the standard format for these. Geocoding services will require data to be in either multi-field, or single field format. **Multi-field** format involves the Address, City, State, and ZIP in separate fields, like the example below:

ADDRESS	CITY	STATE	ZIP	STORE_NAME
1171 PIEDMONT AVE NE	Atlanta	GA	30309	Ace Market
MAIN ST & 41ST AVE	Atlanta	GA	30309	Andrew's Gasoline

Notice for the first address that the Street Number (1171), Street Name (PIEDMONT), Street Type (AVE), and Street Suffix Direction (NE) are all separated by a space, and there aren't any periods or other special characters that may lead to confusion. Another standard format for geocoding data is to have everything in a **single field**, like this:

ADDRESS	STORE_NAME
1171 PIEDMONT AVE NE , Atlanta, GA , 30309	Ace Market
MAIN ST & 41ST AVE , Atlanta, GA , 30309	Andrew's Gasoline

This format involves having the Address, City, State, and ZIP all in one field, separated by a comma. As of this writing, the ArcGIS Desktop geocoders require input data in multi-field format, and the ArcGIS Online and Google geocoding services require input data to be in single field format.

As most address geocoders refer to a streets GIS layer that uses standard U.S. Postal Service abbreviations, following these conventions is also necessary in order to 'match' as many addresses as possible when geocoding. The U.S. Postal Service's standard Street Type, (or suffix) and State abbreviations are listed at: http://pe.usps.gov/text/pub28/28apc_002.htm
Zip codes are good to have as well, but usually a good 'match rate' can be achieved without zip codes. Notice as well the additional table attributes STORE_NAME, is included in the table. This, and all other fields will be retained in the output GIS table after the geocoding is performed.

Geocoding lab
In this exercise you will geocode single addresses and cities using Bing Maps. You will geocode lists of multiple addresses and cities using Google Fusion Tables, ArcGIS, and the Batchgeo.com website. You will create GIS datasets, online maps, and desktop maps of the results. You will learn how to evaluate the accuracy of the data using Batchgeo.com and ArcGIS.

Lab Setup: From the GIS Winter Institute 2017 page, download and unzip the **.ZIP file under the Day 2 Geocoding Lab topic.** This ZIP file contains a folder named *Geocoding_lab* unzip/Extract this folder onto your Desktop.

1) Geocode one address with Bing Maps
In an internet browser, go to http://www.bing.com/maps/
Type an address, such as the CGA's address"1737 Cambridge St., Cambridge, MA " into the window at the top, and click the Search button: 
The map will zoom to this location, and the latitude, longitude values will be displayed (see below). You can highlight these coordinates, right click, Copy, and paste into a word or other document.

Try searching for other places, like city names, water bodies, or famous locations (such as the Eiffel Tower). Bing Maps and most online geocoders will be able to locate "non-address" data as well. To find the latitude/longitude of any point on the map, just right click anywhere on the map. The two numbers listed at the bottom of the pop-up window are the latitude, longitude of that location. *This could be helpful if you want to find the location of specific points, (for example a park bench) instead of just the location the geocoder returns.* Another common way to geocode an address is to use Google Maps at http://maps.google.com. To view the latitude/longitude location of an address on Google Maps, you must right click on the location, and choose "What's here?".

2) Geocode with Google Fusion Tables

A Google fusion table is essentially a spreadsheet on the web that can contain thousands of rows of data. Fusion tables can be linked to one another based on common columns, similar to what

a database can do. Also, a column containing geographic information in a fusion table can be geocoded. In this exercise you'll load a couple of Excel spreadsheets into fusion tables, and map them.

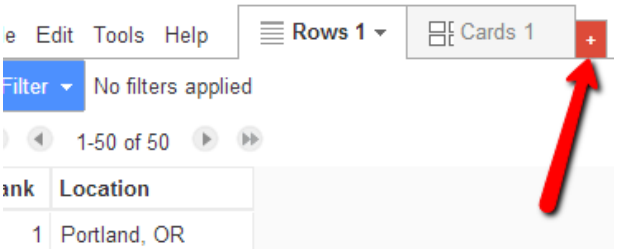
2.1 Open the World_Livable_Cities.xls spreadsheet from your Geocoding_lab folder in Excel, and notice that this contains a list of city, country locations, and a "livability" ranking and index. This list was generated by Business Week magazine, and downloaded in 2012.

2.2 Go to the Google Fusion tables website: https://support.google.com/fusiontables/answer/2571232?hl=en You'll have to log in with your Google account. Click on the "**Create a fusion table**" button.

This will launch the Import new table wizard. Click **Choose File**, and browse to the World_Livable_Cities.xls spreadsheet on your Desktop.
Click it, and click **OK**. Then click **Next**. The Import new table wizard will show a preview of your table. Click **Next** again, and **Finish**. Now you've created a fusion table with your data in it.

In the fusion table, click **Edit** > **Change columns**. Choose the **City-Country** column. Make sure the **Type** is set to **Location**. If it isn't, change it, and click **Save**. The table will recognize locational data, and highlight it in yellow.

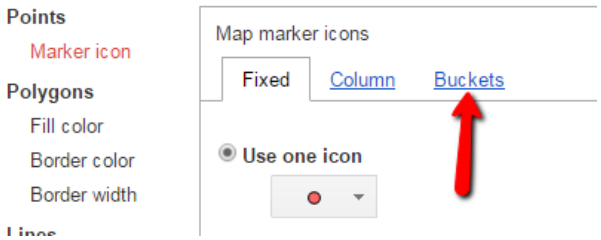
2.3 To geocode these locations, click the red button (see arrow on the graphic to the right) and choose **Add Map**. This will start the geocoding service, with Google trying to match each row of your data using the city and state listed in the City-Country column. Where matches are found, a dot on the map will be added at that location. When complete, an interactive map of the results will be displayed.



Now let's color the dots on the map based on their sustainability ranking.

2.4 Click the "Change feature styles…" button.

2.5 In the Change map styles window, click on Buckets.



Specify:

- 4 buckets
- Column "2007_Rank"
- Still within the Change map feature styles window, click "Automatic legend" which is the last choice on the left. Check the box next to "Show marker legend".

Click Save 

Now your cities will be colored according to these data classifications, or "buckets".

Some notes:

- The map and fusion table can be shared with others by clicking the Share button.
- The locations can be downloaded into a kml file by clicking File > Download. This KML file can be opened in Google Earth or QGIS, or added to ArcGIS by using ArcToolbox > Conversion tools > From KML > KML to Layer tool for further analysis or custom mapping (Not all KML files work with the ArcMap conversion).
- International addresses can also be geocoded with Fusion tables.

More tutorials on Google Fusion Tables:
http://support.google.com/fusiontables/answer/184641

3) Geocode multiple addresses and make a Google Map using Batchgeo.com
Open an internet browser, and go to: http://www.batchgeo.com/

Open the Customer_Locations.xls spreadsheet in the Geocoding_lab folder **and Copy** the 1st 25 rows. In the "**Step 1: Paste your data below**" window on Batchgeo.com, **left click** once, the data will highlight in blue, and then **right click** and choose **paste**. Click the **Validate & Set Options** button. Batchgeo.com should report that 8 columns and 24 rows were read in. Under "**Step 2: Columns & Options**" make sure the following fields match:

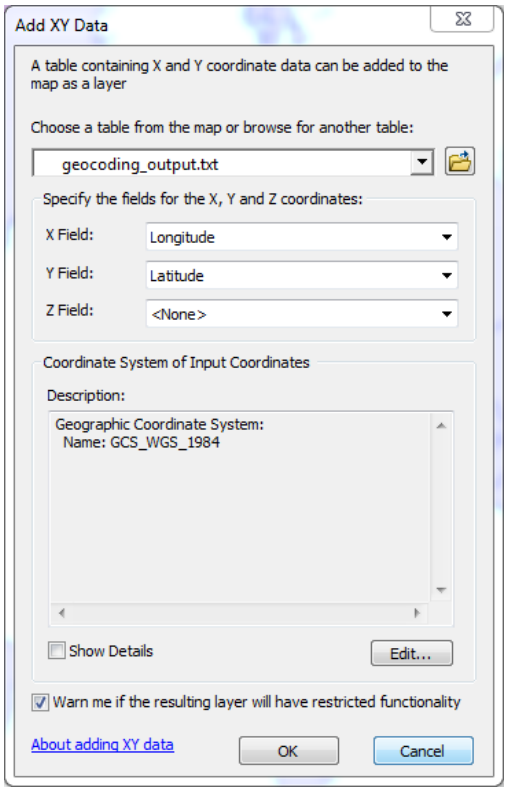
Location / Address: **address**
City / County: **city**
State / Province / Postal Code: **state**

Under "**Step 3: Geocoding**" click **Make Google Map**, and the addresses will be geocoded, and an interactive Google Map will be made of the result.
Click the **Save & Continue** button underneath the map. Enter a Title for the map, and your Email Address, and click **Save Map**. Examine the results on the interactive Google Map. This is a fully functioning Google map that can be shared with others by sending them the http URL. Scroll down past the map, and note that your geocoded points are also listed. Click on one on the list, and it will zoom to the map. Batchgeo also gives you the option to save your geocoded data as KML by scrolling to the bottom of the map page and clicking the **Save as Google Earth** KML link, and **save this file to your Desktop**. This KML file can be opened directly in Google Earth, QGIS, or ArcMap for other uses. For instructions on how to view the latitude, longitude values of the geocoded locations and save them to a shapefile using ArcMap, see the Convert_KML_To_Shapefile_in_Excel.docx document in the Geocoding_Lab folder.

4) Geocode using the Google API
Google's application program interface for geocoding can also be accessed through a Python script. It allows up geocoding up to 2,500 addresses for free, per day. A Python script that performs geocoding is in your Geocoding_lab folder, named "google_geocoding_free_python_2.py". **Right click this file**, and choose "**Edit with IDLE**" to open and view it. Note the Setup steps in this file. Setup steps 1 – 3 have already been completed for you. The formatted input address file is named "google_geocoding_sample.txt", and is in your Geocoding_lab folder. **Double click this file** to view it. Note that it contains the Atlanta addresses, but just in a different format. Close both files.

To geocode the addresses in the "google_geocoding_sample.txt" file, open a command prompt by clicking **Start** and typing in "**cmd**". This will open a black command prompt window. In this window type "**cd Desktop**" to change directories to the Desktop. Then type "**cd Geocoding_lab**". Now type "**google_geocoding_free_python_2.py**" and the script will run, geocoding each address in the output file. After running the script, **double click** the "**geocoding_output.txt**" file to view the results. Note the Location_Type field contains the precision to which the address was geocoded. These precision codes are explained here:

Note: To geocode more than 2,500 addresses per day with the Google API, use Harvard's license of the "Google for Work" api. Download this scrip at: http://gis.harvard.edu/services/blog/address-geocoding-harvard by clicking the "Google Maps API for Work" link.

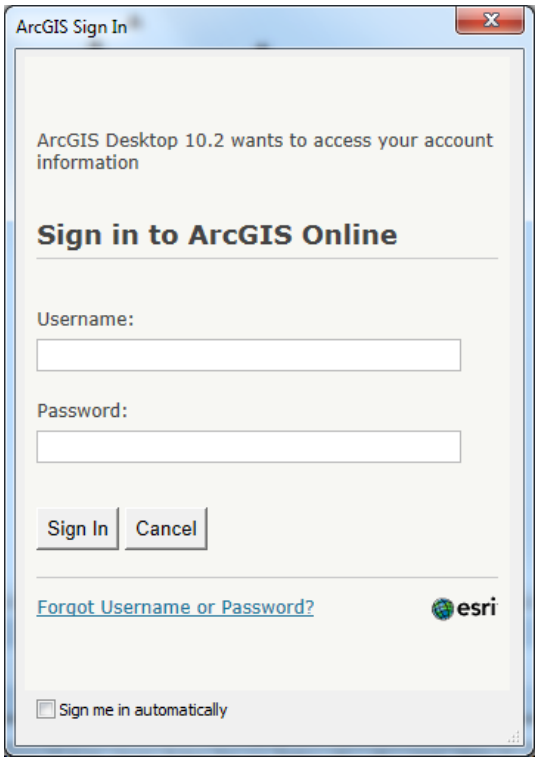


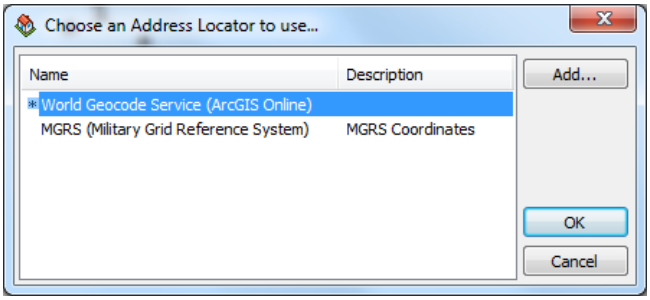
5) Add the geocoding_output.txt file to ArcMap

Open ArcMap, and click **File** > **Add Data** > **Add XY Data**. Specify your geocoding_output.txt file, and the X field as Longitude, Y field as Latitude. Click **Edit**, and change the coordinate system to **Geographic** > **World** > **WGS_1984** (see screenshot at right). Click OK, and your locations will be added to the map. Right click the resulting Events layer, and choose **Data** > **Export Data**, specifying to save the data as a shapefile in your Geocoding_lab/GIS_Data folder. This process can be used to add any latitude/longitude, or other coordinate data to ArcMap as a shapefile.

6) Geocode a batch of addresses using ArcMap

Now you will explore how ArcMap handles geocoding the **Customer_Locations.xls** spreadsheet in ArcMap. To geocode in ArcMap, click **File** > **Add Data** > **Geocoding > Geocode Addresses**. Choose the World Geocode Service (ArcGIS Online) and click **OK**.

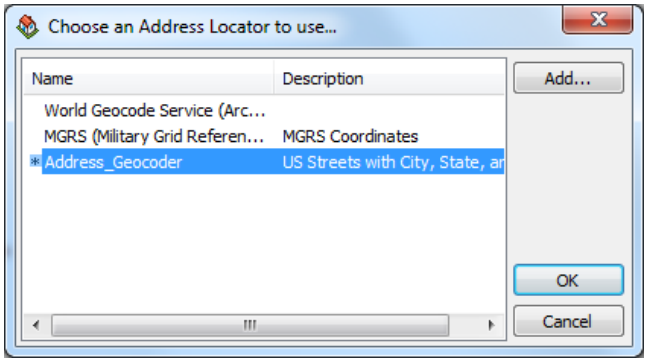




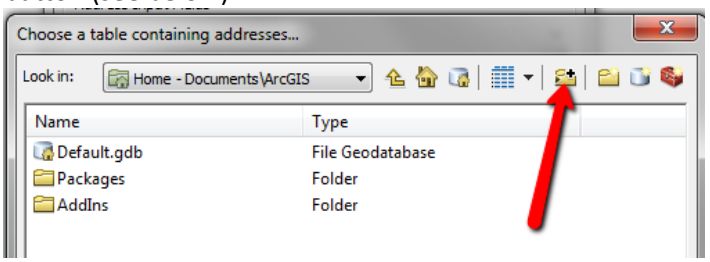
This will prompt you to login to ArcGIS online. To use the World Geocode Service, you must have an ArcGIS Online account. Using the World Geocode service is the most efficient way to geocode addresses into ArcMap. However, it requires you to create an ArcGIS account, which we won't do in this workshop, due to the time involved and hassle associated with everyone creating an account. You can sign up for an ArcGIS Online account at the CGA website here.

Instead, we'll geocode using a local geocoder, made from a GIS dataset containing street and address locations for our area of interest, Atlanta Georgia.

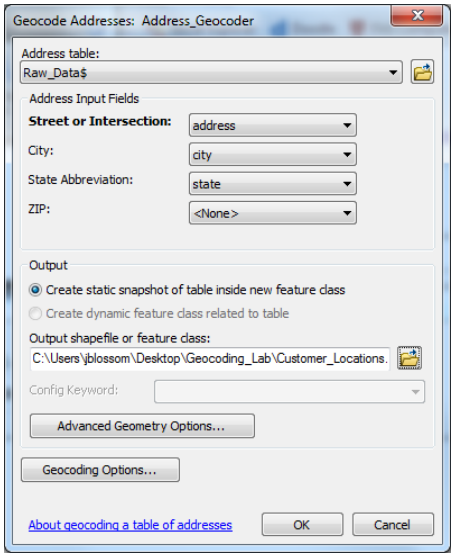
Click **File** > **Add Data** > **Geocoding > Geocode Addresses**. Click the **Add** button. Browse to your Geocoding_Lab folder, Address_Locater subfolder, and choose the Address_Geocoder. Highlight this, and click **Add**, and then click OK (see below).



Next, define your Address table in the Geocode Addresses dialog box by clicking the Browse to button: 
The "Choose a table containing addresses" window will appear.
In this window, click the Connect to Folder button (see below)



Browse to your Desktop\Geocoding_Lab folder and click OK. Now this folder will appear as a dropdown choice when adding or browsing for data from ArcMap. Specify the **Customer_Locations.xls** spreadsheet, **Data** workbook as the table containing addresses. Specify the following:
Street or Intersection: **address**
City: **city**
State: **state**



Specify the output location as your **Geocoding_Lab** folder, name the shapefile **Customer_Locations**. Your screen should look like the screenshot on the left.
Click **OK** to geocode the addresses. Note the resulting match rate, and click Close. Now a layer named Geocoding Result: Customer_Locations will be added to the map, displaying the results of the geocoding.

6.1 Inspect the geocoded results.
In ArcMap, click **File > Add Data > Add Basemap**, and choose the **Streets** basemap, and choose Add. Review the results by highlighting individual addressees in the attribute table, and seeing how they compare to street names listed on the map, or by selecting a point using the Identify tool:  Do the address locations match the correct street names on the Streets basemap?

6.2 Refine the geocoded results.
Right click **Geocoding Result: ArcGIS_Geocoding** and choose **Data > Review/Rematch Addresses**. Note the Status (M=matched, T=Tied, U=Unmatched) and Score (a value between 0 and 100 that represents how closely the text of the input address matched an address in the reference dataset), X (longitude) and Y (latitude) fields.
Address locations can be changed by entering different information in the Street or Intersection field, and clicking Search, highlighting one of the search results, and clicking Match.

This local geocoder was built from the Atlanta_Streets shapefile, contained in the Geocoding_lab\GIS_Data folder. Add this shapefile, open the attribute table, and inspect how it is put together. It is the reference data that the addresses are being matched to. Any reference data in shapefile format can be built into a Geocoder in ArcMap. This is helpful if you can't geocode your data online due to confidentiality / privacy reasons, or if you are geocoding to a specific dataset or historic time period.

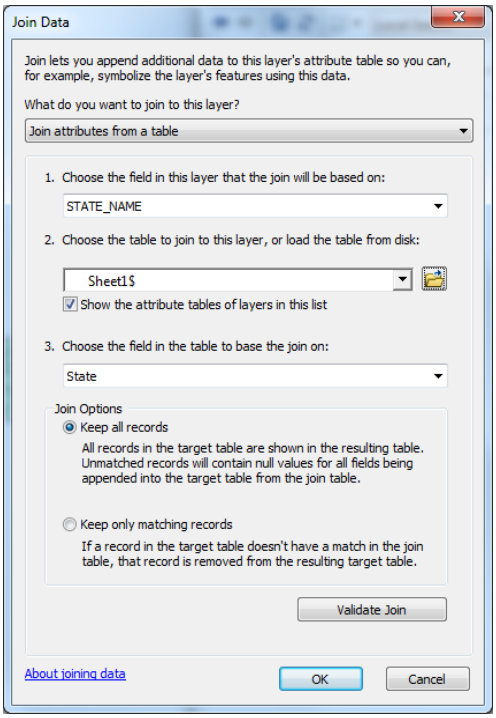
Now you've used both Google and ArcMap to perform geocoding. What's the best geocoder to use? This depends on how many addresses you have, and if they are U.S. or International. A guide to help determine the best geocoder to use is here:
http://gis.harvard.edu/services/blog/address-geocoding-harvard

7) Geocode to an existing dataset.
So far you've geocoded addresses and cities into point locations on the map. Often you may want to geocode data to non-point features, such as administrative areas. In this step, you'll geocode a list of US ski resort counts per state to an existing States shapefile. This list of ski resorts is found in the Ski_Areas_for_mapping.xls spreadsheet. Open it in Excel to examine the contents.

In ArcMap, **Add** the States shapefile using **File** > **Add Data** > **Add Data** and browsing to the Geocoding_Lab location on disk. Double click the **States** file to add it to the map.
Right click the **States** layer and choose **Open Attribute Table**. Note that the STATE_NAME field contains the name of the states. This field is what you'll use to geocode the Ski Areas to.
Close the States attribute table.
Right click the **States** Layer and choose **Joins and Relates > Join**. In the Join Data dialog, specify to "Join attributes from a table" on the top menu choice. Then specify:

1. STATE_NAME
2. Browse to the Ski_Areas_for_mapping.xls file and choose the Ski_Areas workbook.
3. State

Your Join Data window should look like this →
Click **OK** to run the Join.
Right click the **States** layer and choose **Open**



Attribute Table.
Scroll over to the far right of the table, and you should see the Ski_areas field at the very end.

Now you can symbolize your map by this statistic by **right clicking the States layer**, choosing **Properties**, **Symbology** tab, **Quantities** choice on the left, specifying the Value Field as **Ski_areas** and clicking **OK**. To save this geocoded data as a permanent shapefile, right click States and choose Data > Export Data. Specify a name and location for the new shapefile.

For tips on adding additional items to your map layout (like a legend, scalebar, title, etc.), and exporting it into an image or PDF format, refer to the **ArcMap Map Creation Layout Tips.doc** contained in the Geocoding_lab folder.

Conclusion - General tips to keep in mind when geocoding:

- Clearly determine your geocoding purpose, and the accuracy level required to meet this need. If you just need to geocode your data to the city or country level, don't complicate things with address information. However, if address level accuracy is necessary, take the proper steps to ensure confidence in your result.

- Have an 'iterative' mindset. Geocoding may reveal errors or typos in your data, or expose the shortcomings of the geocoding method used. Be prepared to re-geocode, and refine your data and geocoding process accordingly – several geocoding iterations may be necessary to achieve the desired result.

- Don't treat one basemap as the gospel. Compare your results to more than one basemap. For example, if geocoded in ArcMap, import the results to Google Earth to see if they match GE's basemap.

The CGA's geocoding best practices:
http://gis.harvard.edu/services/blog/geocoding-best-practices
A guide to help determine what geocoder to use:
http://gis.harvard.edu/services/blog/address-geocoding-harvard

For GIS and geocoding questions and help contact the CGA at: http://gis.harvard.edu/contactus