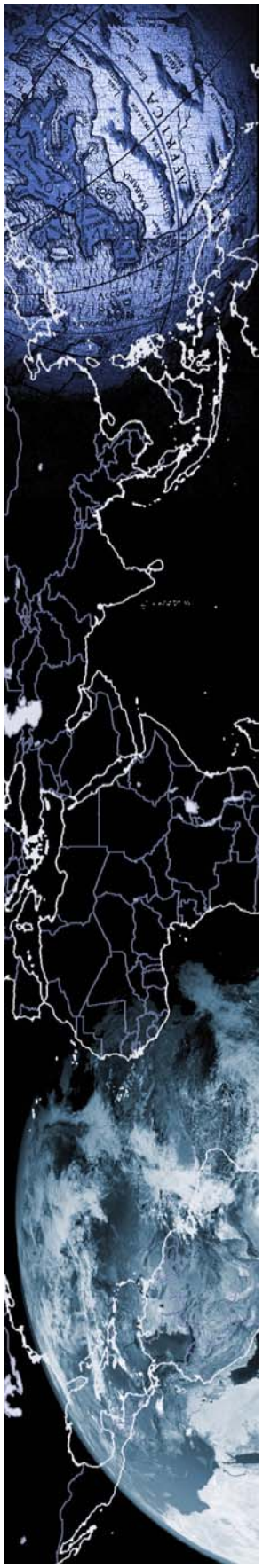


Georeferencing Workshop

March 21, 2008

Case Study:

Lessons Learned and Best Practices Realized from Geocoding Adjundicated Properties in New Orleans, Louisiana

Irina Gumennik, Harvard University, Class of 2007
irina.gumennik@gmail.com

Center for Geographic Analysis
Harvard University

Case Study:
Lessons Learned and Best Practices Realized from Geocoding Adjudicated Properties in New Orleans, Louisiana

May 14, 2007

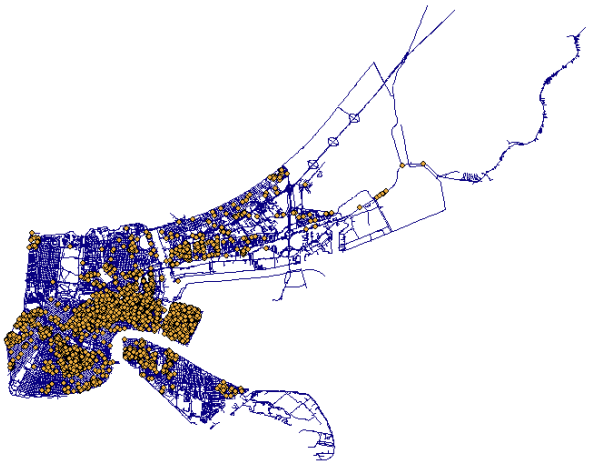
Irina Gumennik, Harvard University, class of 2007

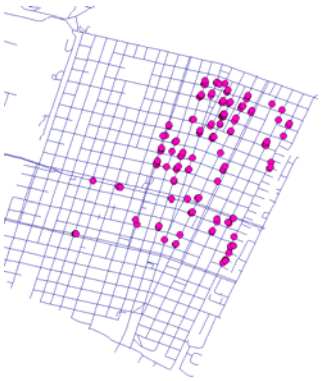
Introduction

Before I impart the lessons painstakingly learned from geocoding properties in New Orleans, I wish to share an experience from a previous attempt at geocoding, my first attempt as it so happens. I will never forget the satisfaction of seeing my first map of Superfund sites in Massachusetts emerge from nothing but a list of addresses. After preparing a list of addresses in Excel, creating an address locator, and setting the geocode parameters, I was amazed by the ability to visualize a tabular list of places on the Earth's surface. But then I began checking the accuracy of the red dots on my Massachusetts base map. My satisfaction quickly dissipated as I realized that the geocode had placed Salem, Massachusetts on Cape Cod. Actually, the geocode misplaced each of the thirty-something points except one. After this first failed excursion into geocoding, I quickly learned the value of good addresses and, most importantly, of validating the results.

I elaborate on this lesson and others learned from geocoding properties for this New Orleans project in three parts. Part I describes a variety of methods for creating a geocoded shapefile, partly to explain the advantages and limitations of a each method, but mainly to suggest several methods that may be combined for one project as a way of validating the results. Part II emphasizes validation during the editing process. In addition to validating the results by creating a number of shapefiles for comparison, validation should continue during the process of editing the best shapefile. The impatient reader may skip directly to Part III, the last section of this document, for a discussion of Best Practices relevant to the New Orleans project in particular and to projects with similar data availability and scale requirements in general.

Contents





Part I.

Creating the Shapefile, or "Validation, Validation, Validation"

A number of geocoding services exist to convert lists of addresses into spatial data. Part I discusses a few programs available to the geocoder, providing a description of the program, its advantages, and its limitations. As many of these programs as possible should be used for any project as a way to validate the results. Different services may provide the best results depending on the goals and data availability of a project (Part III describes the best process found for the New Orleans project after trying three of the services mentioned in this section).

ArcMap

Perhaps the most commonly used method, ArcMap contains tools for geocoding in its package of features. Locations can be geocoded from a list of addresses or mapped from a list of latitude and longitude coordinates using the "Add XY Data" feature. Since coordinates are not as commonly available as addresses, which can be obtained remotely without ever visiting the locations, and since geocoding technically refers to locating a list of addresses on a map, the ArcMap discussion will focus mainly on using the ArcMap geocoding tools. Nonetheless, the "Add XY Data" feature is a valuable tool and will be mentioned in other sections of this Best Practices discussion. I will refrain from going into details about carrying out a geocode in ArcMap because ESRI provides a number of documents, including a 192-page volume entitled "Geocoding in ArcGIS," to explain the geocoding process. Instead, I will talk about the two main lessons learned from going through the process – using the appropriate address locator and reference streetfile.

An address locator links the address data to the reference streetfile and must provide a seamless bridge between them. The correct style must be chosen to fit the format of the address data and the reference data (don't worry, if you are not sure if you have chosen the correct style, ArcMap will quickly let you know with an error message as you create the address locator). The ESRI documentation is useful for distinguishing among all the street locator verbiage, but a good rule of thumb is to use the basic US Streets style for creating an address locator. This style requires only a street number and name (combined into one field in Excel) and is compatible with the most commonly used reference streetfiles. If the address data has the address components in separate fields only, the concatenate feature in Excel can quickly combine them into one field. However, if your data does not have address information beyond the street number and name, the reference streetfile you use will have to be clipped to include only the geographical area that encompasses your address, or the geocode may match points incorrectly in other areas. If possible, try out different

combinations of streetfiles and address locators to obtain multiple shapefiles for comparison as a way to validate your results.

The ArcMap geocode will only be as good as the reference streetfile. Two common streetfiles are available on a national level, ESRI's own StreetMap (http://www.esri.com/software/arcgis/extensions/streetmap/index.html) and the U.S. Census Bureau's TIGER/Line files (http://www.census.gov/geo/www/tiger). Some states and municipalities provide shapefiles specific to their areas. Often produced by universities or government GIS departments, these shapefiles may be clipped versions of a national dataset like TIGER/Line or may be generated entirely by the university or GIS department. These shapefiles should be used when available over general, national-scale streetfiles because they may be more accurate and may help ensure that addresses are not matched outside of the correct geographical area.

The search for a good streetfile to use as reference data should begin with a search for organizations that may produce GIS information specific to your area of interest. A quick Google search or phone call to local officials and non-profit organizations can accomplish this task. Part III provides a list of Louisiana GIS information providers that can be consulted for the New Orleans project in particular. An example of another municipal GIS provider and the variety of shapefiles that may be available for geocoding is San Francisco's Enterprise GIS system (http://www.sfgov.org/site/gis_index.asp). After a free registration process, shapefiles with points for every intersection or every address in the city become available to use with a variety of address formats. While the TIGER/Line streetfile proved to work best for the New Orleans project, as discussed in Part III, geocoding should be performed with all available reference files as a way of validating the results.

As indicated above, the limitations of using ArcMap for geocoding reside mainly in the quality of the address data and reference streetfile. If the address data does not have information more specific than a street number and name, streetfiles with a geographic extent greater than the area that encloses all the addresses can produce incorrect results, as was the case with using the StreetMap data for the New Orleans project. Also, the geocoding tool may not locate an address at the exact location of that address, which would pose a problem for projects that require absolute accuracy. However, with some creativity, an ArcMap geocode can provide a good base to work from for projects such as the New Orleans one that do aspire to 100% accuracy (read on to Part II if you are curious to learn more).

The advantages of using ArcMap include all the advantages of using ArcGIS software, the speed with which several thousand entries can be geocoded, the interactive review and rematching process, and the variety of shapefiles that can be produced from different combinations of reference files and address locators. Some suggestions for making the most of the capabilities of the ArcMap geocoding tool would be to ask the geocode to provide the latitude and longitude coordinates of the matched points (these may come in handy) and to offset the points 5 to 10 meters from the road to place the points closer to the actual property. (Word of caution: Make sure to choose the correct units for the offset! Do not leave the

default "reference unit" option or you will find your data points enigmatically in a beautiful circle completely outside your geographic range.)

Google Earth Pro

Google Earth Pro (http://earth.google.com/earth_pro.html) geocodes addresses from an Excel file directly onto the aerial imagery provided by program. The address data must include city and state information, without which, 100 Main Street would be matched to any of the innumerable Main Streets in the country. The resulting KML file may then be converted into an ArcMap shapefile. The speed and ease of use of the program creates an enticing alternative to the opposite process of geocoding in ArcMap and then converting to KML (using an extension available for download from ESRI); however, initial experience with the program suggests that the results may not be as reliable as those produced by an ArcMap geocode. A Google Earth Pro geocode of the properties in New Orleans showed that the street numbers of the geocoded properties did not follow a logical progression in some instances. The program also does not offer the flexible geocoding parameters that ArcMap offers, such as side offsets or geocoding using only street numbers and names. Conversion from KML to a shapefile may also present problems for maintaining attribute data integrity.

This evaluation of Google Earth Pro is based on the free seven-day trial version, which only geocodes 100 address at one time. The program's full potential requires further exploration, and future versions of the Google Earth Pro will most likely expand the program's functionality and reliability.

While Google Earth Pro as a geocoder may not be the best method, the basic, free Google Earth package is a powerful tool in the geocoding validation process. Converting geocoded shapefiles produced by all of the programs in this section to KML files that can be opened in Google Earth serves as a quick way to visually compare the accuracy of the shapefiles. Google Earth's other functions, such as place marks and latitude and longitude coordinates can be creatively incorporated into the validation process (see Part II for more on Google Earth).

Batchgeocode.com

This free geocoding service is available online at http://www.batchgeocode.com/. The program quickly geocodes an Excel file that contains up to 500 addresses with attribute information (like a file that can be used with ArcMap, except the address information must include city/state or zip code information) into a number of outputs. The program produces another Excel file that contains latitude and longitude coordinates of the points, a KML file, and a webpage, and offers map analysis options, such as distance calculations. The results from this program, with knowledge about KML conversions and projections, can be converted into an ArcMap shapefile. Although the KML file can be converted to a shapefile,

experience from the New Orleans project suggests that the attribute data may not transfer during the process. Alternatively, the latitude and longitude coordinates may be used with ArcMap's "Add XY Data" feature to create a shapefile.

The additional work required to convert the results of this service into files compatible with ArcMap may limit the usefulness of the service for some projects. Additionally, the results may not be as reliable as an ArcMap geocode. Results from the New Orleans project showed that several address were plotted onto one point. Nonetheless, the ease and availability of this program merits a trial for any geocoding project.

GPSPhotoLinker

While this program was not used for the New Orleans project, previews of the program's functionality in other contexts suggest that it can serve as a powerful tool for projects where photographs will be incorporated into attribute data. The program is free and available for download at http://oregonstate.edu/~earlyj/gpsphotolinker/download.php. Use of the program must be incorporated into a project during the data collection stage. GPSPhotoLinker and its companion program MapThePhoto, also available for download from the GPSPhotoLinker website, link photographs to their geographic location on a map. The program is sensitive to flaws in the data collection, which must be carried out by carefully following the program's instructions.

Use of this service is specialized to projects that involve visiting and photographing locations, rather than working solely from a list of address, but for these types of projects, GPSPhotoLinker can prove to be a valuable timesaver and easily produce results that exceed original expectations.

Before moving on to Part II, a note on the data to be geocoded…
A few Best Practices for preparing lists of addresses

The more address information the better! If possible, include all address information available in addition to the minimal street number and name. Including city, state, and zip code information allows for more geocoding flexibility and can aid data sorting and searching in the resulting shapefile. Each address element should have its own Excel field, also to allow more flexibility. The address parts can then be quickly combined into one field as well, especially for geocodes that use the common US Streets address locator style. Finally, and perhaps most importantly, all lists of addresses should come with source information that clearly explains the meaning of the attribute fields, when the data was collected, who collected it, and where they obtained it from.

Part II. Editing the shapefile, or, "More Validation, Validation, Validation"

The extent of editing and validating depends on the number of addresses and the availability of manpower. For geocodes of several thousand addresses, the best way to validate would be to create several shapefiles and visually compare a select number of addresses to determine which shapefile appears to have more accurate placement. Inaccurate points or properties in the best shapefile that were not matched by the geocode should still be revised as much as possible given time and personnel constraints. For geocodes of several hundred addresses, more in depth validation and editing is reasonably attainable. This section provides two suggestions for editing and validation techniques beyond simply visually comparing a number of shapefiles. These methods are more suited for smaller numbers of addresses because they involve checking properties one at a time, but the methods should be applied to geocodes of larger numbers of addresses to the extent possible.

The first method, the one used for geocoding addresses of adjudicated properties in New Orleans, involves looking up the addresses using programs with address search functionality, such as the basic Google Earth software or address finders available on the Internet, such as Microsoft Virtual Earth (http://maps.live.com), Google Maps (http://maps.google.com), Yahoo! Maps (http://maps.yahoo.com), and MapQuest (http://www.mapquest.com). The geocoded location can then be visually compared with the location in one of these address finders, similar to the comparison of various shapefiles. While all of these programs may contain data flaws inherent in geographical data collection and representation, they generally provide a more reliable reference for comparison than other shapefiles produced for a project because these geographic search engines are professionally generated and maintained for daily use on the Internet.

This step can serve simply as another check for accuracy, but it can also be used during the data editing process, particularly for geocodes carried out using ArcMap. The location of an address in one of the address search engines can serve as a reference for choosing among the candidates ArcMap offers for interactive rematching of addresses. However, the positional accuracy of the results will still depend on the information within the reference streetfile.

A second method for editing geocoding results involves using Google Earth to position placemarks on the desired location. The shapefile to be edited is first converted into a KML file (using an extension available for download from ESRI) and opened in Google Earth (this method may also be used as simple visual check of the data). Points that place an address far from its actual location, based on visual determination (a point appears on top of a lake instead of a building, for example) or from searching for the address in Google Earth, are noted and the correct location is marked with a placemark. The placemarks can then be converted into an ArcMap shapefile. Adding attribute data may prove time consuming and difficult, but for projects that require utmost accuracy, this may be a viable method for achieving desired results. Alternatively, the latitude and longitude coordinates of the correct

locations could be recorded in an Excel list with any additional attribute data and then converted into a shapefile using the "Add XY Data" feature in ArcMap.

While this method sounds like just an inefficient way to manually conduct a geocode like one that Google Earth Pro could automatically do, this method is different from using Google Earth Pro to geocode addresses. Firstly, this method can be used to fix a select number of addresses rather than for geocoding all of the addresses. Secondly, this method does not necessarily rely on address information stored within Google Earth because the location for the placemarks can be determined visually from the aerial imagery. Thirdly, an automatic Google Earth Pro geocode may produce inaccuracies that could go unnoticed in a geocode of a large number of addresses.

This method and its variations could also be used to add a small number of addresses to an existing shapefile. Google Earth placemarks or latitude and longitude coordinates could be used to make a small shapefile of the additional addresses. This shapefile can then be combined with the larger shapefile in ArcMap.

For projects that only need points to be in the general area of the actual location, validation may end at a simple visual comparison of shapefiles to choose the best one. For projects, such as the New Orleans one, that aspire to 100% accuracy, the two editing methods suggested here may be used to the extent possible given the number of addresses. Nevertheless, with enough dedication and man power, even thousands of addresses can be converted into a highly accurate shapefile.

Part III. Best Practices for Geocoding Adjudicated Properties in New Orleans

Firstly, one quick word for those of you who skipped Parts I and II and jumped directly to this section – validation, validation, validation. If you are curious to find out what I mean when I use this word, please go back and take a look at Parts I and II.

The New Orleans project

The New Orleans project asked for a shapefile of adjudicated properties in the parish that could be used by property owners and non-profit organizations in their efforts to gather resources for rebuilding after the destruction in 2005 by Hurricanes Katrina and Rita. The project set a target of 100% accuracy for the shapefile because each property would be given individual attention by anyone using the web-based GIS database that would store the property information. The discussion in this section provides instructions for anyone wishing to replicate the geocoding process for future endeavors associated with this project.

The final shapefiles were created using ArcMap from two lists of addresses, one with over 6,000 addresses within all of New Orleans and another of about 160 addresses within the Ninth Ward. Ideas for improving the process and addressing the challenges posed by the high accuracy required and a large number of addresses may also be gained from the information in Parts I and II (if you still have not read Parts I and II, perhaps now your curiosity has been aroused enough that you will read them before moving on!)

Preparing the list of addresses

Both lists of addresses provided for the geocode contained street number and name information. Also, since all of the properties were in New Orleans, Louisiana, city and state information could easily be added to the Excel file. For projects with addresses in multiple cities, the city and state information should be included for each address. In general, including as much address information as possible in the original list allows for more flexibility during the geocoding process and may be useful for analyses conducted using the shapefile. Each address element should appear in a separate field to provide additional flexibility for geocoding and data sorting. Fields can be combined later in Excel if the need arises, such as for geocodes that use the ArcMap US Streets address locator. Source information should also be provided in the Excel file, explaining all attribute fields, listing the dates the addresses were compiled, where they were obtained from, and who carried out the process.

Choosing a reference streetfile

The TIGER/Line streetfile proved to be the best available streetfile. The TIGER/Line streetfile used for this geocode encompasses only the parish of New Orleans and is available for download from the Louisiana Oil Spill Coordinator's Office at http://lagic.lsu.edu/loscoweb. ESRI's StreetMap data was also used to geocode the addresses and produced comparable numbers of addresses that were matched, tied, and unmatched. The geocode from TIGER/Line streetfile was chosen as the better of the two because the TIGER/Line shapefile showed more street detail when compared to the StreetMap road lines. Additionally, the New Orleans streets were extracted from StreetMap data by clipping a square around New Orleans, and several points were mapped to the few areas within the square that were outside of the parish.

Additional reference files for Louisiana may be available from the following sources:
City of New Orleans GIS Data Portal <http://gisweb.cityofno.com/cnogis/dataportal.aspx>
Louisiana GIS Council <http://www.doa.state.la.us/lgisc/>
Atlas: The Louisiana Statewide GIS <http://atlas.lsu.edu/>
Hurricane Katrina and Rita Clearinghouse Cooperative <http://katrina.lsu.edu/>
Greater New Orleans Community Data Center (GNOCDC) <http://www.gnocdc.org/>
Louisiana Oil Spill Coordinator's Office <http://lagic.lsu.edu/loscoweb>

Creating an address locator

This step involves creating an address locator using the US Streets style with the TIGER/Line streetfile. This style requires only the street number and name, combined in one field. Address locators that use only the street number and name work best with streetfiles that include only the geographical area of interest, such as the New Orleans TIGER/Line shapefile used for this geocode. If a different streetfile is used for future geocodes, more address information may be required to obtain an accurate geocode with points falling within the correct geographical area. The ESRI StreetMap data comes with its own address locator for geocodes that use StreetMap as the reference streetfile.

The geocode…drum roll please!

The geocoding process available through the tools in the menu bar offers more geocoding options and is preferable to the geocoding tool available through the toolbox in ArcMap v9.2. Some suggestions for making the most of the ArcMap geocoding options include asking the geocode to provide X and Y coordinates in the output fields (they might come in handy) and choosing a 5 to 10-meter side offset. Make sure to choose the correct

units for the offset. Leaving the default "reference data units" option will result in a circle of points outside of the correct geographic range.

Once you've chosen the geocoding parameters, hit "OK" and let ArcMap work its magic!

Editing and validating

Before working on editing and validating your shapefile, try carrying out the geocode using some of the other services mentioned in Part I and compare the results. You may create a better shapefile (and can change these Best Practices recommendations) or affirm that ArcMap created the most reliable shapefile.

Both methods for editing and validating suggested in Part II are applicable for the New Orleans project due to its goal of 100% accuracy.

If you still have not read Parts I and II, this section should sufficiently entice you to go back and read them.

Writing up the metadata

Once you have created the final shapefile, write up a metadata document to tell everyone how hard it was! The metadata also comes in handy for anyone who will use your shapefile. The metadata should explain the attribute fields and source of the information (hopefully, this was provided with the address list), cite any shapefiles used in the creation of your shapefile, and provide a description of the methodology used and any possible data flaws. Spatial reference information can also be added by copying the information from the spatial metadata for your shapefile visible in ArcCatalog. Beyond this basic information, feel free to include any other information you feel is necessary to accompany your shapefile through its travels.

Final thoughts

This process produced a shapefile from the list of about 160 addresses that matched every address to a geographic location. The initial geocode matched 93% of the addresses to candidates with a score of 100, 5% to candidates with a score of 75, and 1% were left unmatched. Validating and interactively matching using Google Earth and Google Maps produced a reliable shapefile. Validation with Google Earth suggests that even more accurate results can be obtained using the somewhat rigorous process detailed in Part II.

The list of over 6,000 addresses encountered more obstacles, but this process produced a reliable shapefile given the available resources. The ArcMap geocode was only able to match about 60% of the addresses, leaving about 40% unmatched. Yet these results were more reliable than those produced by the Google Earth Pro trial version and batchgeocode.com, and ArcMap proved to be easier to use.

While these were the best results given the resources available to this geocoder, suggesting that these may be the Best Practices for future geocoders to follow for the New Orleans project, data and software availability, as well as project goals, may change. The various resources described in parts I and II should be explored further (especially GPSPhotoLinker!) and may be combined in different ways to produce even better results. The real Best Practice for this project and any other projects, for those of you dedicated enough to have read to the end, is to try a variety of methods and validate, validate, validate in order to have confidence in the results.