# Innovative Ways of Using Geo-processing Techniques to Add Value to Police Crime Data for the County Open Data Portal

## Abstract of the Program

Montgomery County started an open data effort a few years ago, called dataMontgomery. GIS contributed directly to this effort with actual submissions of geographic data. However, perhaps less known and more notable are value added to non-geographic data submitted to dataMontgomery by other County departments using the power of GIS. The Montgomery County Department of Technology Services – Geographic Information Systems team (DTS-GIS) has automated and fine-tuned ways to improve performance of data uploads, add geographic details to tabular address data, anonymize often sensitive crime data, and greatly reduce processing time for all of the above.

While the solutions developed by DTS-GIS are used for multiple data sets on the Socrata-based dataMontogmery site, the most prominent tabular data set enhanced by the efforts of the DTS-GIS team are daily updated tables of crime locations submitted by the Police department, which can be viewed online at https://data.montgomerycountymd.gov/Public-Safety/Countywide-Crime-Map-with-Icons/u9k4-nwwu.

## The Problem or Need for the Program

The first problem with tabular data submissions that the dataMontgomery team brought to DTS-GIS was the need for a better way to upload tables with street addresses to their Socrata open data system so they could be displayed on maps in that platform. Many Departments supplied data with addresses that already could be uploaded to Socrata's cloud and automatically geocoded on their servers to display as points on maps; however, Socrata's servers were painfully slow in this process, often timing out with larger data sets that were uploaded frequently. The dataMontgomery team noticed, however, that the process on Socrata was significantly faster when latitude and longitude coordinates of a location were supplied to Socrata and used in place of address to locate the points. The dataMontgomery team needed an automated way to add latitude and longitude fields to the data coming in from various departments that only included a street address.

While the DTS-GIS team was processing the geographic coordinates, the dataMontgomery team also hoped they could add some other useful geographic data to new fields in the table. In which of the County's five Council Districts was the address located? For crime data coming in from the Police Department, in which large Police District, smaller Police Beat, and neighborhood-sized Response Area did the crime occur?

Also with the crime data, was there a way DTS-GIS could provide a generalized location in the output data without revealing the actual address of the crime, to provide some anonymity?

One of the open data features that dataMontgomery had looked forward to offering with the crime data from the Police Department was to allow users to filter crimes by city. Unfortunately, they discovered that the city name, often hurriedly entered by officers who had more pressing duties at the time, frequently contained typos that prevented grouping the data by city. Was there a way DTS-GIS could automate checking and fixing the city name field in the data?

Finally, after DTS-GIS had designed ways to accomplish all of the above, could it speed up the processing time so large data sets could be updated within a day?

**Description of the Program**

The initial request from dataMontgomery included adding fields for latitude and longitude coordinates of the map point, as well as checking the resulting geographic point against other existing GIS layers and filling in fields for other GIS polygons that contain the address. This was done by automatically geocoding – creating map coordinates for addresses – via .NET technology's multi-threading approach.

The above solution worked great for tabular data on Capital Improvement Projects submitted by the County's Department of General Services. The method was also used successfully for residential building permits.

However, the dataMontgomery Team realized unique challenges existed with the tabular data that would be coming in from the Police Department on crime locations that possibly also could be addressed using the power of GIS.

First, before determining the latitude and longitude of the point to be displayed on the map, the data on crimes should be somewhat generalized so, for example, a victim could not be identified because the map point was directly on his or her home. At the same time, the point needed to be geographically accurate enough so geographic processing with other GIS layers – in this case, Police District, Police Beat, and Police Response Area (PRA) - would provide accurate results in fields for those areas in the output data. Balancing these two requirements was tricky because, if a crime occurred on a street that happened to serve as a boundary for two Police Districts and the generalization of the address changed the street number from odd to even, the result would place the map point on the other side of the street and in the wrong Police District. *Figure 1* in the supplemental file *shows that **e**ach crime address must be geocoded twice – once using the actual address for assignment of polygons (such as Districts) and a second time using the hundred-block address to anonymize the point location.*

The solution was to create an automated workflow that actually geocoded the address of the crime twice. First, the original, unaltered address from the crime data was geocoded for the sole purpose of using GIS to determine the Police District, Beat, and PRA. For this purpose, DTS-GIS designed an address locator – a set of GIS rules for determining

the location of the point on the map – that offered the most geographically accurate point possible, in most cases placing the point directly atop the building footprint listed in the address. The GIS locator looked for the address in DTS-GIS' existing building footprint layer first, and if it didn't find it, another data layer of properties was checked. If the locator still didn't find the address, it would geocode according to DTS-GIS' street centerline data layer with a 30-foot offset. Again, the resulting location was used only to populate the new fields for the Police District, Beat, and Response Area.

Next, in an effort to somewhat anonymize the data as it would ultimately appear on dataMontgomery's map, it was run through a script to convert each original address into a generalized hundred-block address – such as changing "1234 Main Street" to "1200 Main Street." Most addresses were simply rounded down to the nearest hundred-block, with the exception of "unit block" addresses numbered lower than 100. These were instead rounded to one, because geocoding an address of "0 Main Street" would have failed to produce a point on the map.

A second address locator was designed for the purpose of generating the latitude and longitude fields, which were ultimately used to place the point on the map. This locator only used the street centerline data and, unlike the more precise first locator, was designed to place the point directly on the street centerline at the end of the block rather than with an offset to the correct side of the street. This way, it wouldn't even be clear from looking at the map on which side of the street the crime occurred or how far down the block, and thus achieved some degree of anonymity.

Care was given before running the addresses through either of address locators to first convert street type abbreviations to the standards used by the DTS-GIS team; for example, records in which the Police had input "8000 Wisconsin Av" as the address were changed via script to "8000 Wisconsin Ave" so the address locator would more readily recognize the address.

As the dataMontomery and DTS-GIS teams began working with the crime data, problems with the city name field populated by Police were discovered. First, most "cities" in Montgomery County are actually unincorporated areas with no actual boundaries, causing some areas of the County to be "gray areas" referred to using multiple names. Police entering data for crimes occurring in such areas would understandably use different names for the same place, resulting in one crime listed as "Silver Spring" and another on the same block as "Aspen Hill." Furthermore, even when the same city name was used, typos in the city name were very common. All of the above data issues with the City Name field contributed to severely limiting the usefulness of one of the most anticipated features of Socrata's system – aggregating the crime data by city. DTS-GIS added to its script a check of a zip code GIS layer disseminated by the State of Maryland, which included a field for city name, to normalize and correct spelling for city name.

When all of this was first set up using XML, a typical crime data 136,000 might take 40 hours to process. This was not acceptable to the dataMontgomery team, which planned to update the crime data delta daily and the entire data set quarterly.

DTS-GIS redesigned the process using a database-driven system in place of XML. The file-based approach had required each record to be processed inside RAM. The script read from the file and wrote back to the file through a process called "File I/O," which is very slow. In contrast, using an Oracle database, each record is read from the database, processed, and then sent back to the database, in a much quicker process.

*Figures 2a & 2b of the attached supplemental materials provide a visual representation of the work performed by the team in support of this program. Figure 2a depicts the earlier design using the XML file processing, while Figure 2b represents the new and enhanced design of using Oracle DBMS processing.*

**Use of Technology**

ESRI technology used in this project includes:

- ESRI ArcGIS Desktop software for maintaining our in-house street centerlines and building footprints data layers
- ArcGIS Server for hosting:
  - Map services for determining new polygon fields (such as Police District, Beat, and PRA)
  - Geocoding services (for both the precise locator for determining the new polygon fields and the hundred-block locator for providing some anonymity of the displayed point.)

The scripting was done using C# .NET. The original script accessed single-threading XML files. The later database model that improved performance approximately tenfold uses the Language Integrated Query (LINQ) feature of .NET to query an Oracle database. Again, this allowed quicker processing of the database itself rather than reading and writing to a file. A small investment of time in setting up Oracle schema and table for the Police Crime reports yielded a big dividend of processing efficiency.

DTS' Enterprise Service Bus (ESB) team then processes the resulting data, which now has corrected latitude, longitude, and other fields from the database into a format that is published to the Socrata based dataMontgomery website.

**The Cost of the Program**

The total cost to develop the program is estimated at $25,000 primarily in staff time. The project was completed over the course of 6-months. DTS-GIS staff leveraged existing hardware and software licenses to develop the program.

Jurisdictions seeking to replicate the County's efforts may be expected to invest $25,000 to $50,000 or more to develop a similar program. The cost will depend on such factors as the richness of existing GIS services, information and applications; labor costs; the extent to which stakeholders are engaged in the development process; the need for new software licenses and/or hardware etc.

**The Results / Success of the Program**

The scripts that added the new fields of latitude, longitude, and in which of several regions the address was located worked beautifully. The locators were fine-tuned repeatedly to examine addresses that didn't geocode properly and adjust the script so it would successfully process as many legitimate addresses as possible. In one test, out of about 124,000 total crime records, 120,000 locations were able to be geocoded by the system, for a 97.6% success rate. The few addresses that failed to geocode were found mostly to have been entered incorrectly in the original data by the Police officers.

The conversion from the original address to the hundred-block address also worked as expected, once particulars such as how to treat address numbers lower than 100 were addressed.

The efforts to use GIS' zip code map layer to normalize and correct spelling of the city name field successfully allows dataMontgomery users to organize the data based on the city name.

The attempt to speed up the processing time by switching from XML to database was also successful. Processing time for a typical crime data update of 124,000 records decreased from 32 to 40 hours down to four to six hours with the new method.

**Worthiness of an Award**

Montgomery County's Open Data program is widely hailed as one of the most successful and progressive programs in the nation. However, certain data sets provide much more value to constituents if they can be visualized. In some instances, visualization is further enhanced through the use of GIS maps and related mapping services. This specific program is an example of a highly innovative and low-cost use of existing GIS technologies and platforms to improve performance of data uploads, add geographic details to tabular address data, anonymize often sensitive crime data, and greatly reduce processing time. Citizens usually don't have the opportunity to view such detailed crime data, and the anonymization added via the GIS technique made this possible despite the sensitive nature of the data.

This program can serve as a model for other jurisdictions, small, medium or large, seeking to leverage the power of GIS in new and innovative ways.

**Supplemental Materials**

Please refer to the attached supplemental materials.