

# Using GIS to Find New Uses for Old Data

Eighty percent of municipal data has some kind of geographic component.

By Graham S. Hayes, Scott A. Mayers, Charles E. Penasack

**R**ecycling programs have made a huge impact on reducing land-fill volumes as well as raising public awareness of the importance of conserving resources.

The promise of recycling is to recover value from objects that were once determined to be worthless. While data is rarely considered a renewable resource, integrating technologies like GIS along with traditional business systems can create new value for old data. GIS can provide non-traditional visualization and analytical insights into spatial relationships between datasets that would never be realized from rows and columns of data traditionally locked behind legacy business applications.

GIS provides a link between maps and databases, as well as a wide range of tools to integrate, manage, and visualize data to make sound business decisions. The contents of any database can be displayed on a digital map if the data refers to a map feature (point, line, or polygon) in space (X, Y, and Z).

Just as there is value in recycling renewable resources like paper, glass, and plastic, many of our governmental clients unknowingly own and maintain a wealth of renewable data resources. Data that was originally captured for one purpose can be reused or repurposed to gain additional benefits.

## Legacy Data Integration

Today's wealth of technology integration platforms can enable utilities to recycle much of their existing legacy data by providing not only a spatial view of the data, but also trending and business intelligence about the data. Some examples of legacy data that can be used within a GIS include:

- Customer Service History

- Consumption and Billing Data
- Property or Tax Assessor Information
- Zoning Information
- Water Quality Databases
- Leak History
- Hydrant Databases
- Valve Databases
- Tank Databases
- Pump Performance Records
- Inspection Records

The legacy databases may be elaborate systems built on modern application architectures (ERP on a client server platform), housed on mainframes with data entry panels and reporting functions, or may be simple Excel® or Access® databases stored on individual workstations. In either case, chances are that when the data was first collected, it was never intended to be displayed on a map or even organized to maximize its query potential in a database structure.

## Making the Connection

Since the legacy systems were not designed with maps in mind, our challenge is to assess and recycle as much of the data as possible and to make a connection between the legacy data and a digital map.

80 percent of municipal data has some kind of geographic com-

ponent. The most common data elements found in disparate databases are street addresses. Address information can act as a primary or secondary data source. Databases for billing, meter reading, customer information systems, call center applications, and water quality lab information systems (LIMS), use addresses as a primary data element. Asset management and maintenance systems for valves, hydrants, pumps, tanks, leak reports, and complaint databases often contain address information in comment fields, but are generally not built around address components.

A process called geocoding takes an address from a record in a database and creates a point on a map based on an exact match to a parcel polygon or point layer or by interpolating along a digital street centerline. Geocoding engines operate by parsing an address into recognizable components including:

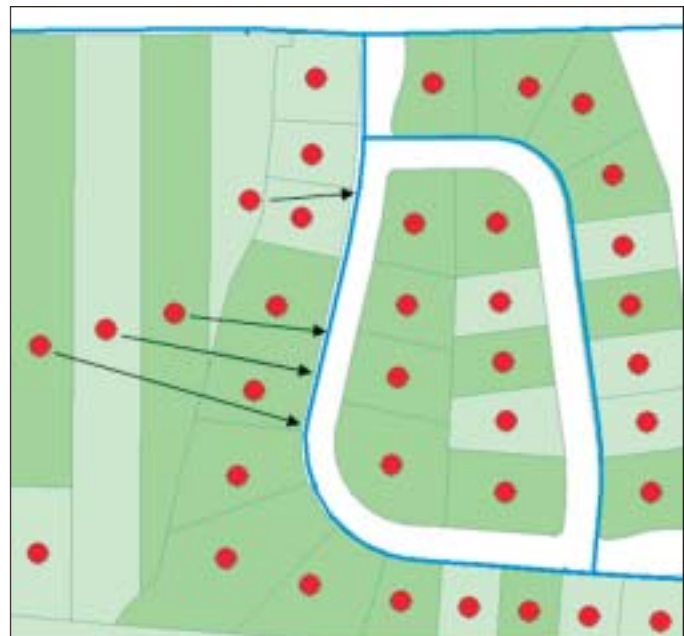


Figure 1. Parcel centroids snapping to wrong mains.

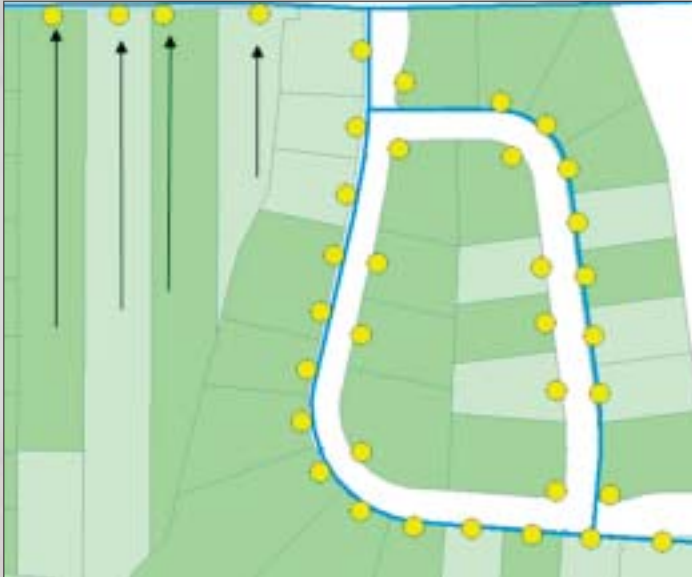


Figure 2. ROW points associated with correct main.

- House Number—2345
- Direction-Prefix—N/S/E/W
- Street Name—Main
- Street Type—St, Ct, Ave, etc.
- Direction-Suffix—N/S/E/W
- Zone—zip code or town name

The database address components are then compared to parcel or street centerline address components using a technique called a soundex. A soundex is a phonetic indexing system that creates a four character representation of a word based on the way a name sounds rather than the way it is spelled. Theoretically, using this system, you should be able to index a name so that it can be found no matter how it was spelled. As such, it can help resolve misspellings like North, Nroth; and inconsistencies like First, 1st; Parkway, Pkwy; OBRIEN, O'Brien, O Brien; etc.

Geocoding cannot directly process Post Office Box numbers or incomplete addresses (i.e., missing house numbers). Addresses based on landmarks (city hall, university arena, main street mall, etc.) can be processed, but a special landmark file needs to be established in advance. Apartment numbers or office suites are problematic for the geocoding address parser, and as such need to be isolated in their own fields in database records as part of the address cleanup effort.

The use of consistent address styles, correct street name spelling, complete street names (including type and direc-

tion a l prefix/suffix values), and zip codes or municipality names in the legacy address files will increase the percentage of features that can be successfully geocoded. Usual success rates for geocoding raw legacy address data are around 75 to 80 percent. As more care is taken to verify

the addresses, the higher the match rates will climb.

### Supporting Layers

The other component that will greatly affect the final success of the geocoding process is the quality of the geocoding reference layer. This reference layer is a point, line, or polygon dataset in the GIS that the geocoding engine uses to create an XY point when an address match is found. Obviously, the more up-to-date the reference layers are, including recent subdivisions, accurate street names, and complete address ranges and zip codes, the better the chances of successfully matching the addresses. The Census Bureau provides free TIGER (Topologically Integrated Geographic Encoding and Referencing) street centerline data, but the quality is less than adequate for robust geocoding. Commercial vendors like GDT ([www.geographic.com](http://www.geographic.com)), NavTeq ([www.navteq.com](http://www.navteq.com)), and others sell enhanced quality street centerline data by zip code, by county, by state, or by

country for both the United States, Canada, and Europe.

Parcel data layers can also act as geocoding reference layers if the parcel attribute tables contain property address information in addition to owner address information. Understandably, but unfortunately, there is often more effort placed on establishing and maintaining the mailing address to send the annual tax notices than on storing the address of the property itself.

Parcel data layers can be expressed as polygons or by points called centroids, which represent the geographical center point of the parcel polygon. Either data layer can be used for geocoding. However, when performing GIS functions like closest neighbor analysis, demand modeling, vehicle routing, or spatial joins between a water or sewer network and the nearest parcels, etc., points work better than polygons and are, therefore, preferable for geocoding purposes.

One of the downsides to using parcel centroids for spatial joins with water or sewer lines in the right of way (ROW) is that the centroids in odd shaped or deep parcels may not “snap” to—or find—the closest main (Figure 1). A workaround with a multi-purpose solution is to create a set of parcel ROW points positioned just outside the ROW along the parcel frontage line (Figure 2). These ROW points can be used for 911 address locations, vehicle routing, simple buffer searches, and spatial joins with objects in the ROW. The ROW



Figure 3. Finding the oldest building structures to estimate the age of the infrastructure.

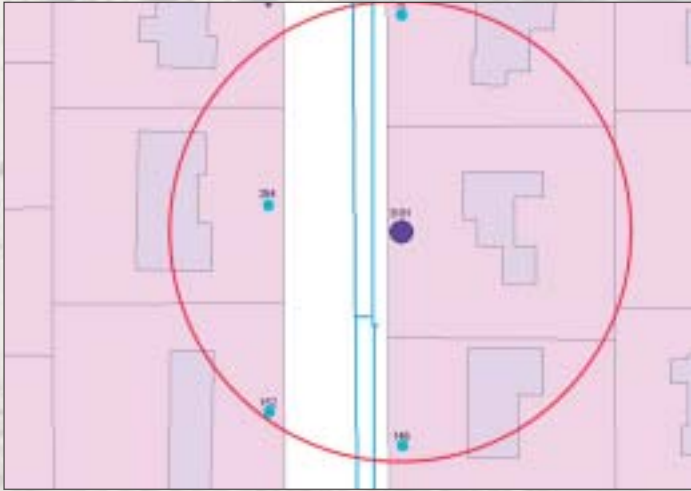


Figure 4. Comparison of nearest neighbor water usage.

points can also act as pseudo service connections to approximate the curb box or meter location. As crews visit these sites, the ROW point position can be adjusted and tagged with a feature-based metadata pedigree value to indicate the level of confidence in the service connection location.

## Geo-Relational Integration

Once the database address data has been assessed, scrubbed, and possibly reformatted to handle special cases and apartment or suite numbers, the data can be batch processed against the streets, parcels, or ROW points to create a layer of geocoded points. Whatever columns are present in the source address file (the file whose records are being geocoded) will also exist as attributes in the geocoded layer. Inclusion of common database keys (primary and foreign) will enable the geocoded points to be associated with other related databases and tables.

Because GIS operations are based on a Relational Database Management System (RDBMS), GIS tables can be joined or linked via standard Structured Query Language (SQL) through either native database connectivity, or to open systems through Open Database Connectivity (ODBC) compliant data interfaces. Direct access to legacy tables will be a function of the operating system and the availability of ODBC drivers.

With the primary keys established, these geocoded points can be tied to any accessible database through any com-

mon IDs (e.g., customer billing, meter maintenance, usage demand statistics, leak history, work order records, water quality test results, complaint records, etc.).

The primary advantage is not that the GIS can “see” into a related table,

though that is important, but rather, that the contents of the legacy database can be “seen” on a map

## Legacy Reuse, Visualization, and Analysis

**PARCEL/ASSESSOR DATA**—Assessor data tables contain a wealth of information that can be used directly or inferred to create baseline GIS utility data. For example, one piece of data that is often missing when conducting a condition assessment of water or sewer infrastructure is the installation date or age of assets in the ground. While records of recent repair work may indicate the age of the latest main replacement, records for older construction are often missing. However, many parcel databases may record the year built for structures on each parcel. By performing a spatial join between each pipe segment and the closest parcel polygons or ROW points, users can search the related assessor tables based on the parcel IDs and find the oldest year built date. By comparing this information to surrounding, connected mains, it can be inferred that the mains may not be any older than the oldest structure (Figure 3). Of course patchwork development and later construction might take place on undeveloped property that had utility lines installed years before construction, but this is one example of how an unrelated piece of information can be recycled and reused to infill gaps in the asset knowledge base.

Other pieces of data that can be

extracted from the assessors’ residential data tables and used for water and sewer rate studies, growth and demand models, or backflow prevention initiatives include: the presence or absence of septic systems; the number of bathrooms present; and the presence, absence, and depth of private groundwater wells. Having an accurate count of bedrooms and bathrooms in a residential property provides one more piece of data useful for comparison between the usage statistics from the billing database and the expected demand based on the actual plumbing. By assessing the number of residential properties on septic systems and performing spatial joins to the sanitary sewers fronting those properties one can predict the future load on the wastewater treatment plants if and when all residents connect to the sanitary network.

## BILLING SYSTEMS/CUSTOMER DATABASES

—Engineering efforts like hydraulic modeling can also benefit from the reuse of billing information by transferring customer usage statistics from the billing system to the demand nodes in a hydraulic model. Because residential water use can vary widely based on the number of occupants, socio-economic demographics, conservation habits of the residents, presence or absence of a pool, etc., a GIS-based spatial analysis can be helpful in predicting and displaying water use information. By performing a background spatial analysis that systematically compares the usage statistics of a given property to the four to five closest neighbors with the same property use classification (i.e., residential, commercial, etc.), usage values that are significantly different can be flagged for further research. Figure 4 illustrates the results of such an analysis. The property in question reportedly used over 3,000 gpd as opposed to the four nearest neighbors that only consumed an average of 156 gpd.

Significant differences between the expected and actual water consumption could indicate several conditions—a mis-calibrated meter, the presence of a leak on the customer side of the meter, wasteful or overly conservative customers, or, in case of extremely high water use, the possible presence of a

drug lab. Methamphetamine labs use an excessive amount of water and police departments often use water billing records to obtain search warrants. Properties with excessive water demand could be plotted on a map to look for spatial trends or patterns.

**CUSTOMER SERVICE CALL CENTERS**—GIS reuse of call locations can benefit customer service call centers by displaying all active calls on a map within a given time frame (Figure 5). This recycled view of the data allows dispatchers to assess and recognize trends or clusters in call patterns related to single or multiple outages, water quality, or other basic maintenance events (high bill, street leaks, etc.). By proactively recycling this legacy data, systems can be deployed to place outgoing voice calls to the customer base thereby lessening the volume of inbound calls.

**CUSTOMER NOTIFICATION**—Customer service location points can also be reused to notify customers in the event of service disruption. Using a spatial search, utilities can identify all customers affected by an outage and contact these customers using an automated outbound dialer. This would proactively warn customers in a region when a series of calls begin to emanate from a specific neighborhood. Other integrated GIS functions include zooming to the property address of the caller and displaying call history from the caller and surrounding neighbors. Integrating and reusing the location of all recent work orders would help to recognize and communicate changes in the system (i.e., flushing programs, leaks, pressure loss from fire flows, etc.).

**MINING COMMENT FIELDS**—One of the most misused data elements in any database is the free form comment field. Comment fields are useful for recording notes and observations, but are difficult for effective query and retrieval of information. However, sometimes the comment fields contain text regarding leaks, complaints, missing or found assets, measurements from building corners, etc. that might have a potential reuse. By scanning, filtering, and searching comment fields using key words, some useful information may be discovered. If the data value is signifi-

cant an automated parsing script or routine may be developed to extract the data into a more useful format by creating separate columns to indicate the presence or absence of key words.

**VEHICLE ROUTING**—The same ROW points can be reused to assess and reconfigure current meter reading routes to develop optimal routes that reduce vehicle operation costs, reduce the time required to read the meters, and potentially decrease the number of meter reading staff. Water quality sampling operations, and maintenance crews can all take advantage of known locations to schedule stops and to use the associated records assigned to those points for work order processing.

If inspection crews follow a calendar based schedule, it may be worthwhile to plot the locations of the inspection sites to determine if there is a strong case for changing the dates of the inspections to reduce the vehicle miles. We have seen many cases where crews are given a printout from a legacy system of valves, hydrants, meters, or manholes to inspect. Since the data from the legacy system had never “seen” a map, the ordered list would usually be organized by district, or grid, or project number—none of which would lead to logistical efficiency. Instead, the crews will drive by the same hydrant, valve, or manhole for days until they reached that asset on their ordered printout.

Obviously, evaluating and reusing the location of assets within the context of a maintenance system would be a great way to increase productivity.



Figure 5. Call Center GIS display of four low pressure calls on different streets.

## Conclusions

There is a wealth of data being collected and managed by different departments within a municipal organization. In many cases, data being collected by one group could have significant value when reused by other groups. By assessing the available data stores around their organization and looking for innovative, unusual uses of data, users can reap significant rewards.

The key components to integrating disparate databases include 1) common IDs to join the tables together based on common values (e.g., customer numbers, valve ID, sample numbers, meter numbers, etc.), and 2) a geographic component—either known XY coordinates or address information to support geocoding.

By standardizing primary and foreign keys, and centralizing and cleaning up a master address file, data from many different systems can be integrated, recycled, and repurposed for both business and mapping operations. Often the reuse of the data provides more value than the initial intended use. **GE**

*Dr. Hayes, a certified GIS professional (GISP) and the National GIS Practice Leader for Red Oak Consulting, a division of Malcolm Pirnie, can be reached at [GHayes@pirnie.com](mailto:GHayes@pirnie.com); Mr. Mayers is an Associate with Malcolm Pirnie/Red Oak Consulting and can be reached at [SMayers@pirnie.com](mailto:SMayers@pirnie.com); and Mr. Penasack, an experienced GIS Analyst with Malcolm Pirnie/Red Oak Consulting, can be reached at [CPenasack@pirnie.com](mailto:CPenasack@pirnie.com).*