Merging National Street-Network Datasets

Jonathan W. Lowe

With the acquisition of GDT by Tele Atlas, the companies are blending their MultiNet and Dynamap products and business models, undertaking what may be the largest spatial data merge ever.

On July 7, 2004, Tele Atlas acquired GDT (Geographic Data Technology). With that contract's ink now dry, the former rivals have spent the past three months merging their data, technical processes, and business models. In this column, Michael Gerling, former CEO of GDT and now COO for Tele Atlas North America (NA), explains how the two companies operated before they merged and how they now acquire and maintain data as a single entity. Though straightforward about the challenges, Gerling paints a convincing picture of the acquisition's benefits to both originally separate customer bases.

Corporate Navigation

Tele Atlas supplied only European street-network data until entering the U.S. market three years ago with the acquisition of Etak, a national street-network data provider based in Menlo Park, California. Following the acquisition, Etak became Tele Atlas NA with a North American dataset called *MultiNet*. After becoming one company, Tele Atlas Europe and Tele Atlas NA still had different data standards and models, but no areas of overlapping geographic coverage — data-wise the acquisition was nonredundant.

The more recent acquisition of GDT, however, is an entirely different situation: almost all of GDT's U.S. and Canadian street-network data, called *Dynamap*, overlaps that of Tele Atlas NA. Rather than maintaining separate datasets, Tele Atlas and (the former) GDT are blending MultiNet and Dynamap data management practices and business models. To understand their strategy, it helps to know how each company acquired and maintained data in the past.

Tele Atlas NA catered to markets more interested in navigation and dynamic traffic or weather content whereas GDT dominated the "traditional GIS" and geocoding markets. That traditional GIS market was the more mature of the two, as evidenced by GDT's customer base, which Gerling estimated to be 10 times larger than Tele Atlas/Etak's prior to the acquisition. (Of course, that customer-base comparison doesn't include a possible skew by NAVTEQ, which continues to compete with Tele Atlas NA for navigation customers.)

Though covering the same national territory, the two companies also had different cultures — GDT operated in a matrix management fashion whereas, at least on paper, Tele Atlas NA/Etak operated more hierarchically. Gerling reported that the two cultures are gradually blending. Perhaps more importantly, though, the two companies managed their data according to markedly different business plans, which Gerling characterized as "compilation-centric" (GDT) versus "field-survey-centric" (Tele Atlas/Etak).

The Business of Data Capture

To build the MultiNet dataset, Tele Atlas NA/Etak's field-survey-centric data-capture model relied on direct collection of street-network data, mainly by driving the streets with a GPS to capture street carriageway data. At its peak build phase 18 months ago, for instance, the Tele Atlas NA fleet was approximately 200 vehicles strong. Offering another point of comparison, Gerling speculated that NAVTEQ, which also applies a field-survey-centric business model, may have managed a 400–500 vehicle fleet during its peak build phase.

In contrast, a compilation-centric business model emphasizes indirect data collection through multiple external sources. Gerling called this business model the *Network,* referring to a network of external authoritative data observers (such as government agencies and other private companies) rather than a network of streets.

For instance, instead of driving city streets, a compilation-centric vendor

Net Results columnist **Jonathan W. Lowe** covers the role of emerging technologies in the exchange of spatial information. Lowe is the owner of Local Knowledge Consulting (Berkeley, California), where he designs and implements spatial Web sites. He can be contacted at info@giswebsite.com.

might form an alliance with a utility company that services that city. When the utility's mapping department creates professional survey-quality street data as part of its regular work, it would subsequently share that data with the compilation-centric vendor. GDT relied on such a business model, drawing on 30,000–35,000 U.S. and Canadian field relationships for ongoing input about the real world. Some direct field survey is unavoidable, but GDT's maximum fleet size was negligible compared with its field-survey-centric competitors. It never exceeded eight vehicles, and GDT dispatched each vehicle only for targeted collection.

Lacking a large field-survey fleet, GDT's compilation-centric business model hinged on rigorous data and metadata management practices, standardization, and delivery rather than direct collection. Because GDT gathered data from such a wide variety of sources, any given GDT street segment's geometry and attributes could originate from several different observers or time periods. Thus, GDT's compilation-centric Dynamap dataset became a mosaic of multiple geometries and attributes accumulated over time. Tracking each discrete element of data — geometry, attributes, and time of capture — was central to GDT's business model. As new data arrived, it only replaced existing data if GDT was more confident about its accuracy.

The Network

Financially, Gerling argued, maintaining a fleet of hundreds of cars is a less attractive business proposition than maintaining a Network of thousands of occasional contact people. In addition, with tens of thousands of sources and observers, change detection becomes much more manageable than driving the streets to detect change. As reasonable as that sounds, it's natural to wonder, what keeps that network alive? and what is the ongoing benefit to the contributors? For one thing, explained Gerling, the government is obligated to provide public-domain data to any citizen who requests it. So, if the contributor is a municipal agency, there doesn't have to be a compelling benefit — the public policy principle is enough.

However, not all contributors are government entities. Private utility companies, for instance, have some of the highest-quality data to share, but no legal requirement to do so. To this kind of contributor, the quid pro quo or exchange value of sharing local knowledge is two-fold. First, suppose a utility company uses Dynamap data to manage its facility. Naturally, the company wants its data to be accurate. The utility could maintain its own private dataset or modify its copy of Dynamap, but would have to repeat this process with each new Dynamap update. Why not make the changes official and leave the bulk of data management to GDT? As added incentive, contributors making significant improvements over time may even get Dynamap at a cheaper rate. Plus, for the contribution of a relatively small area, such as a city or neighborhood, the return is usually a much larger area — perhaps an entire county.

Sometimes little to no negotiation is needed, such as when the Network is an Internet crawler or robot. Any U.S. commercial street-network dataset contains millions of turn-restriction data elements. A turn restriction indicates what a motorist's options are at each intersection, preventing a routing algorithm from recommending the wrong turn into opposing traffic on a one-way street, for example. "The false myth," said Gerling, "is that turn restrictions can only be captured by actually driving the streets." On the contrary, he explained that turn-restriction change detection can frequently be automatically discovered without ever sending a company representative to the intersection being changed.

> "The false myth," said Gerling, "is that turn restrictions can only be captured by actually driving the streets."

For instance, imagine that a municipality decides to change a two-way street into a one-way street. Most municipal procedures require that such a proposal be posted multiple times in a legal arena, which, more often than not, may be on a Web page. Web crawlers (computer programs that scan the entire Internet, page by page, searching for specific text) may find such postings automatically. Even if no Web posting exists, construction companies or regional transit-planning authorities responsible for implementing such a two-way to one-way change typically maintain work orders. Those work orders may also be posted online for crawler discovery.

Though large and varied, GDT's Network of contributing participants tends to provide data of significantly higher quality — anywhere between 3-meter (from GPS capture) and 6-inch (from a professional survey) accuracy — than, say, the 20-meter accuracy of the U.S. Census Bureau's TIGER (Topologically Integrated Geographic Encoding Referencing) files.

May the Best Line Win

Following the GDT acquisition, when Tele Atlas initially compared the two U.S. streets datasets, it was able to correlate 90–95 percent of segments in the two databases. "Those matches are not all identical and certainly not error-free, but there is a high correlation," reported Gerling. In other words, for the majority of individual records, Tele Atlas knows when Dynamap and MultiNet line segments refer to the same physical strip of asphalt.

Following correlation, the challenge was to determine which of the paired records is the "better" of the two. Considering the two original data-capture business models, trying to thin each pair of correlated records may sound like comparing apples to oranges. But Tele Atlas isn't comparing just raw geometry or attributes, it is weighing levels of confidence in each. The level of confidence flows from the other major difference in the two former rivals' business models: metadata granularity. What kind of metadata capture processes did the two vendors originally deploy for their MultiNet and Dynamap products, and why?

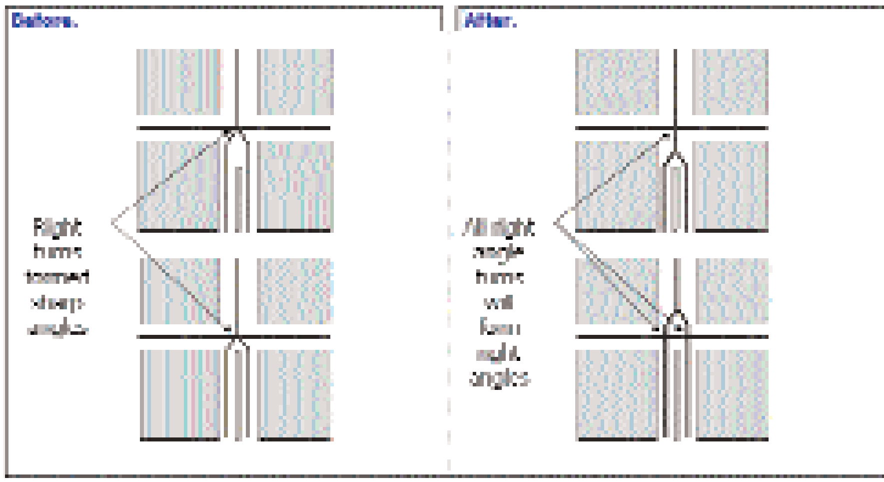

Figure 1. A data-model decision: should the transition from dual to single digitization occur at intersection nodes, before the intersection nodes (if the median ends before the intersection), or after them (if the median extends to the intersection or beyond)? Despite the increased complexity, Tele Atlas NA chose before and after rather than at because this model improves turn angles and actually simplifies some large, asymmetric intersections.

Elemental Metadata. Prior to the Tele Atlas acquisition, most of GDTs compilation-centric information came to them indirectly. GDT knew that it couldn't always be as confident about external data's accuracy or reliability as it would be if collecting them directly. But, if it could rely on the integrity of its sources, that external data might be even *more* accurate than any field-survey crew could hope to capture simply by driving the roads. Consequently, GDT's strategy was to ingest as much external data as possible, but to maintain metadata at a very fine-grained level that Gerling called *elemental*. GDT's greatest effort then went into assessing the quality of each element, and promoting the elements with highest confidence ratings to the master dataset.

By "elemental," Gerling refers to the fact that a street segment has both geometry and attributes (name, address ranges, and so on). GDT's metadata approach assigned a series of metadata records to the geometry and each of its attributes, all including a date of capture. So, when a municipality participating in the Network notified GDT of a road realignment, GDT might update that road segment's geometry while leaving its attributes untouched. The metadata for that segment's new geometry would then be updated to identify the source (a municipality, for instance), date, and so on. Any given line segment, then, could have multiple sources and different elemental metadata for each element — a melting-pot approach to metadata capture.

Project-Based Metadata. MultiNet field-survey crews captured metadata not element by element, but by time, process, and project according to the logistics of their field-survey-centric schedule. Tele Atlas' national data are the result of thousands of field-survey projects, each conducted according to specific rules and collected, overall, in three years. Accuracy standards and work instructions improved with experience during that three year period, so later projects have different metadata than earlier ones. Compared to Dynamap's elemental metadata approach, MultiNet metadata takes a blanket approach to all records in a specific field-survey project locality.

Consequently, unlike a GDT street segment's mosaic of elemental metadata, the Tele Atlas data for any given segment's geometry and attributes is all the same. For example, Pittsburgh's streets may have different metadata than Philadelphia's, but within the Pittsburgh project, all geometries and attributes are a result of the same standards and work instructions.

Confidence Game

Tele Atlas is now determining, element-by-element, exactly which Dynamap or MultiNet elements are the most accurate of each pair. To work at that elemental level, even though MultiNet data is at the time-process-project level, Tele Atlas simply assigns each MultiNet data element its own metadata entry based on the blanket specification. This allows for the finest grain of comparison to Dynamap metadata. At that elemental metadata level, more recent geometry may trump older, more positionally accurate geometry (or vice versa) depending on Tele Atlas' confidence-calculating algorithms.

When comparing any two MultiNet and Dynamap streets, Tele Atlas may find that the best geometry for the first block comes from Dynamap, while next block has more accuracy in MultiNet. Or that the third block retains the MultiNet geometry and turn restrictions, but integrates the Dynamap street ranges and signage text. To Gerling, all datasets are nothing more than additional resources entering the compilation process, which revolves around comparison of metadata. In summary, when two reports from two sources disagree, degree of confidence (based on metadata evaluation) helps resolve the "winner," which becomes the new gold standard. The "loser" is then archived.

But what about the 5–10 percent of records that can't be correlated, or the situations in which confidence is low for both members of a correlated pair? It's these targeted cases in which a field-survey fleet makes good sense. From a business-model perspective, Tele Atlas' decision to acquire both Etak and GDT may make the most sense not from a data perspective, but from a combined business-model perspective. Though benefiting from the Network model's cost-effectiveness, Tele Atlas NA can now also deploy its fleet of field-surveyors to resolve otherwise irreconcilable differences in a more focused fashion. Tele Atlas has reduced the former Etak's fleet size, but not to the bare-bones scale originally employed by GDT. According to Gerling, the resulting product delivers the best of both business models.

Choosing Data Models

Despite its approach to integration, however, Tele Atlas did face at least one set of decisions early in the process of unifying MultiNet and Dynamap datasets: how to resolve the differences in data models. A data model is a consistent approach to modeling a real-world feature, such as the intersection between a dual-carriageway road and a one-way street. Applications perform routing, geocoding, and various custom tasks based on a network's data model, so vendors tend to be cautious about making major changes to their models that might disrupt their customer's applications.

One approach to satisfying different data-model needs is to store the elemental core data in one format, but modify it automatically at the "product extract level," according to Gerling. For instance, if a customer needs centerlines rather than multiple carriageways, the core data elements could be multiple carriageways that are collapsed programmatically to produce single centerlines as needed. Similarly, network classifications can vary if comprehensively coded within the database. And the database can dynamically calculate the linear references needed by departments of transportation. "That's the difference between a 'data factory' and a cartography company," said Gerling.

Not all differing models lend themselves to this extract logic, though. Gerling reported that there was "already more in common than not" between the two models, but confirmed that of the hundreds of specifications, a few conflicts required an executive decision. As a simple example, Tele Atlas NA will now support 50-character, rather than 30-character, fields, which GDT had historically supported. At the geometry level, a simple example involves the model for delineating how divided roads merge back into single bidirectional segments (see Figure 1).

To make these decisions, Tele Atlas NA shared its plan with key customers before implementing the final model. And lest we blame an acquisition process for changes to a data model, Gerling wryly pointed out that GDT's data model had definitely not remained static during the company's history. The acquisition was a good excuse for further improvements to an already evolving product line.

It's rare to hear about an acquisition that went easily or smoothly, but if I were a deep-pocketed investor, Gerling's strategy would have convinced me that Tele Atlas' vision is sound. However, Tele Atlas NA will need every element of Gerling's convincing vision as the newness of the acquisition fades, and the hard work sets in. Let's wish him luck. If his team can walk its talk, the customers of both former companies stand to reap greater benefits than they could have separately. ⊕