Flexible Data Models Strut the Runway

Jonathan W. Lowe

Geospatial industry leaders, such as ESRI (www.esri.com) President Jack Dangermond and Intergraph Mapping and Geospatial Solutions (www.intergraph.com/gis) President Preetha Pulusani, are predicting that the development of standard geospatial data models will be one of this year's key advances (*Geospatial Solutions*, "Market Map 2003," January 2003). Their clients and partners are mobilizing; experts in two dozen different vertical markets have already begun comparing their data and agreeing to standard representations of the real-world physical objects central to their disciplines. By communicating ideas with UML and other schematics, detailed text explanations, and sometimes sample datasets, vendors now offer their customers a blueprint or framework for storing and managing discipline–specific enterprise datasets.

This column covers the purpose of data models and steps through some examples of applying them to real-world problems.

What's sprouting

A quick inventory of the markets and sciences sprouting data models includes biodiversity, defense intelligence, utilities, environmental regulated facilities, forestry, geology, historic preservation and archaeology, hydrology and hydrography, land parcels, petroleum and pipeline, and telecommunications. There are also more generic data models on the drawing board, such as those for address, basemap, census administrative boundaries, marine, and transportation data that apply to several vertical geospatial markets.

"You're only as good as your data," the saying goes. That might have to be revised to "You're only as good as your data model."

Intergraph has developed data model templates based on work with its customers. The company cites the success of not just data models, but enterprise data models in particular. At customer sites such as Oncor (www.oncorgroup.com), and Florida Power Corporation (www.fpc.com), data models common to the whole company enable more efficient integration of formerly separate systems, such as those for assigning trouble tickets, outage analysis, call grouping, circuit tracing, and event monitoring.

ESRI's approach to data modeling relies on a seasoned domain expert as leader — for instance, Michael Goodchild (University of California, Santa Barbara, www.geog.ucsb.edu) in transportation, Nancy von Meyer (Fairview Industries, www.fairview-industries.com) in land parcels, and Peter Veenstra (M.J. Harden Associates, Inc., www.mjharden.com) in pipeline. These well-known leaders, vendors, and industry groups apparently anticipate enough benefit in a standard data model to devote their valuable time to its development. Something's going on here.

Nonstandard standards

We laugh about such adages as, "standards are great; everyone should have one," but we laugh ruefully. By now most of us are somewhat jaded by previous pronouncements that echoed down from lofty committees — "Behold! Thou shalt use our mighty new standard!"— but the standards were never widely adopted. The ESRI data modeling groups are taking a different approach by trying to assimilate as many different standards as possible in one flexible relational database model.

According to the ESRI transportation group, for instance, "Transportation standards such as NCHRP 20-27, FGDC, and GDF all deal with the basic transportation network in different ways. Our goal is not to begin a new standards effort, but to support these standards with a practical database design that works well with ArcGIS." Some ESRI-based groups have already posted extensive online documentation of

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. Lowe can be contacted at info@giswebsite.com.

Glossary

FGDC: Federal Geographic Data Committee

GDF: Geographic data file

LRM: Linear referencing method

NCHRP: National Cooperative Highway Research Program

UML: Unified modeling language

their models, including a UML conceptual diagram, a logical data model diagram, a sample dataset, and text-based documentation explaining the context of the model.

Sweating the geo-details

One of the problems ESRI's transportation data modeling group hopes to solve is the difficulty for many organizations, particularly smaller agencies and those without years of transportation GIS experience, to understand the best way to implement a data model. Their new twist lies in their definition of "best." Substitute "most flexible" and data models start to make sense. Start with a flexible example, and refine that model to match your organization's unique application requirements.

Because data models are arbitrary simplifications of our infinitely complex real world, there's always more than one way to simplify. For instance, a transportation model could represent a city's road network (see Figure 1) with thin polygons designating the two edges of the asphalt from curb to curb (see Figure 1a), or one line per carriageway (see Figure 1b), or with a single line running down the center of each street (see Figure 1c). Is any one of these the best transportation data model? Without understanding the transportation applications they will support, it's impossible to say whether any of these models are intrinsically better or worse than the others.

Applications in this context mean "a problem to be solved." For example, consider an application that has to map a traffic accident at the intersection of 11th Street and Broadway. Burrowing deeper into the previous three transportation models, how does each one store street intersections, and will they effectively support our example application?

In the curb-to-curb polygon model, just identifying intersections at all is ungainly — polygons are contiguous rather than intersecting (see Figure 2a), touching along lines rather than points. The multiple-carriageway model has several intersections for the same pair of streets if either street has two or more carriageways (see Figure 2b). How will our application consistently assign accident events to the same intersection when multiple choices exist? So, by process of elimination, the single centerline approach (see Figure 2c) seems to be the best candidate for consistently locating accidents at 11th Street and Broadway.

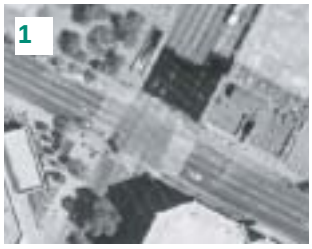
On the other hand, if the application was predicting traffic congestion or supplying detailed turn-by-turn driving directions, the multiple-carriageway model might be a more appropriate choice. Or if the application involved asphalt repaving calculations, then the areal extents supplied by a polygon-based model would make it the winner.

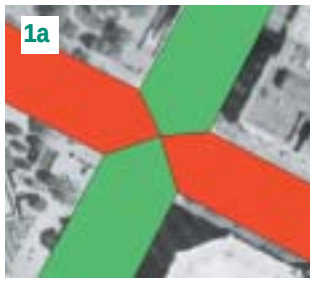
I'll be the judge of that!

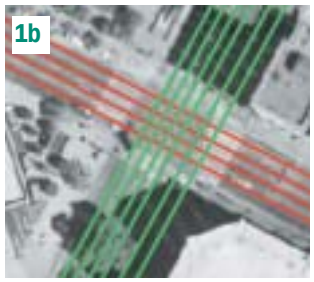
In other words, there's no way to judge a data model as good or bad without knowing how it is applied. Consequently, one goal of the groups building data models is to produce a structure that is flexible enough to satisfy the needs of as many applications as possible within the model's industry segment while minimizing data duplication and redundancy.

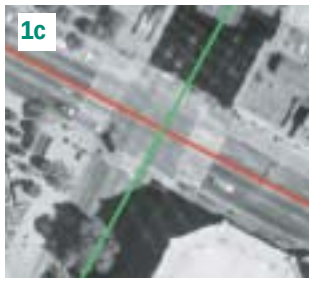
Revisiting our intersection example, the ESRI transportation data model combines a multiple-carriageway approach with the option of differentiating between logical intersections (where accidents should be mapped) and geometric network intersections (places that lines cross, but that are not formal cross-street points, such as where a private driveway meets a city street, or multiple carriageways intersect a cross-street) If you're willing to invest the time to identify which intersection points are logical and which are geometric, then you can have your cake and eat it, too.

Representing intersections with the single-centerline model is a self-maintaining process because intersecting points

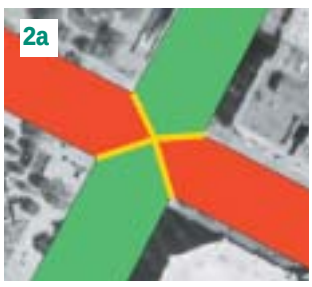


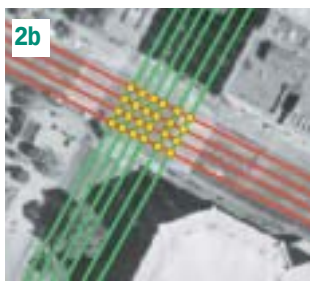


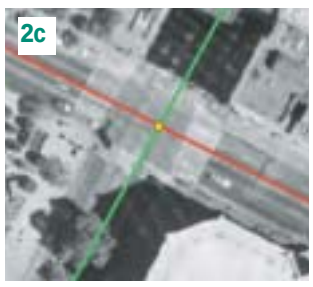




FIGURES 1–1c Data models can be constructed to fit the individual application. For instance, transportation models for the intersection of 11th and Broadway in Oakland, California (**1**) could represent the intersection with polygons from curb to curb (**1a**), with one line per carriageway (**1b**), or with one line for the entire street (**1c**).







FIGURES 2a–2c Again using the 11th and Broadway example, though all four polygons touch at only one point, there are several common edges that could be mistaken for *the* intersection (**2a**). Modeling multiple carriageways results in 24 valid but different intersection points at 11th and Broadway (**2b**). The single centerline approach has only one possible intersection at 11th and Broadway (**2c**).

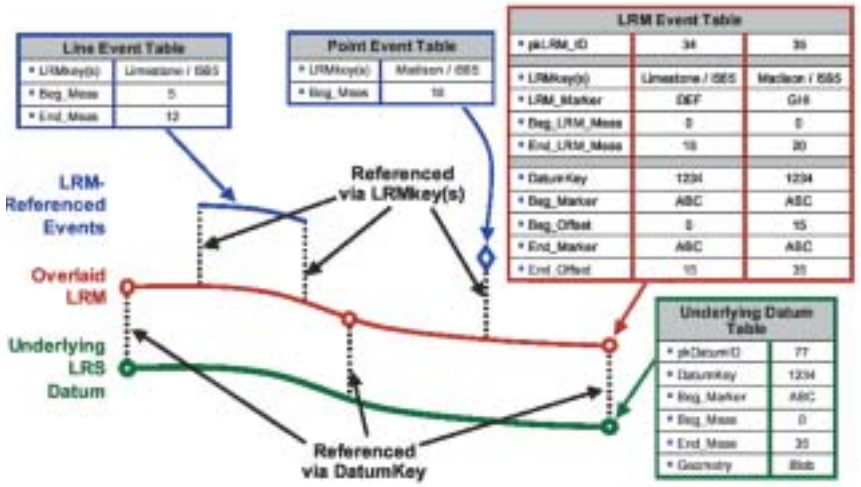


FIGURE 3 Models respect relationships when data changes. This Intergraph GeoMedia-based diagram shows a county milepost LRM overlaid onto underlying geometry. The LRM breaks at a county boundary even though the underlying geometry does not.

are implied everywhere lines cross each other. Data models based on a single geometry (lines in this example) are common in the simple environment of flat file data storage, such as shapefiles. Typically, GIS flat files store only geometry and its associated attributes, leaving rules and behaviors to the GIS software. If flat files hold the linear street network data, for instance, then the desktop GIS software must find the intersections, plot the accidents, and put the resulting points into a separate, unlinked file. If streets are realigned, automatically updating accident locations is the sole responsibility of the desktop GIS, not the flat file itself. With this simple data structure comes some limitations, especially as application designers attempt to model reality with increasing accuracy and sophistication.

More flexible yet more complex, both Intergraph's and ESRI's transportation data models replace flat files with databases, shifting some of the rules and relationships from the desktop GIS program to the data itself. Data structured according to Intergraph's GeoTrans data model, for instance, enables automatic cascaded intersection updates whenever related street lines are renamed, realigned, or deleted (see Figure 3). Back to our accident example, if an editor renamed Broadway to Chavez Road, the database and GIS desktop program would automatically update the intersection name to 11th Street and Chavez Road, also preserving links between the newly-named intersection and any accidents mapped there. Any domain rules or relationships residing in the database with the geometry and its attributes will enforce clean data practices regardless of who modifies that data or what application they use when editing it.

When relationships are an integral part of a dataset, and the GIS software can interpret and manipulate data according to those relationships, more sophisticated models are possible and the whole dataset stays accurate and current. Thus, if a transportation manager needs to map a tanker truck's spill of toxic material, the ESRI transportation model stores such an accident not only as a (typical) point, but also as a polygon defined by the extent of the spill. Though represented by different geometrics (point and polygon), both accident locators are linked by a shared accident identification number.

Parlez-vous UML?

Storing data relationships in a database for manipulation with GIS software is one challenge; communicating the structure of those relationships is yet another. UML diagrams and other schematics illustrate the relationships between tables in a model, showing how one parcel may have many owners, and how one owner may own many parcels (see Figure 4). UML shows tables as named boxes enclosing lists of attributes and shows relationships between the tables connecting lines. Where the lines touch the boxes, numbers or symbols indicate whether the table relationship is one-to-many, many-to-one, or many-to-many — capturing possibilities such as the owner and parcel situations discussed earlier.

Understanding a data model's relationship structure reveals whether it will support your organization's applications correctly. For instance, if one parcel can have multiple owners, then adding new owners should preserve any existing owners' names. In a one-to-one owner-to-parcel relationship, adding a new owner should overwrite the old owner.

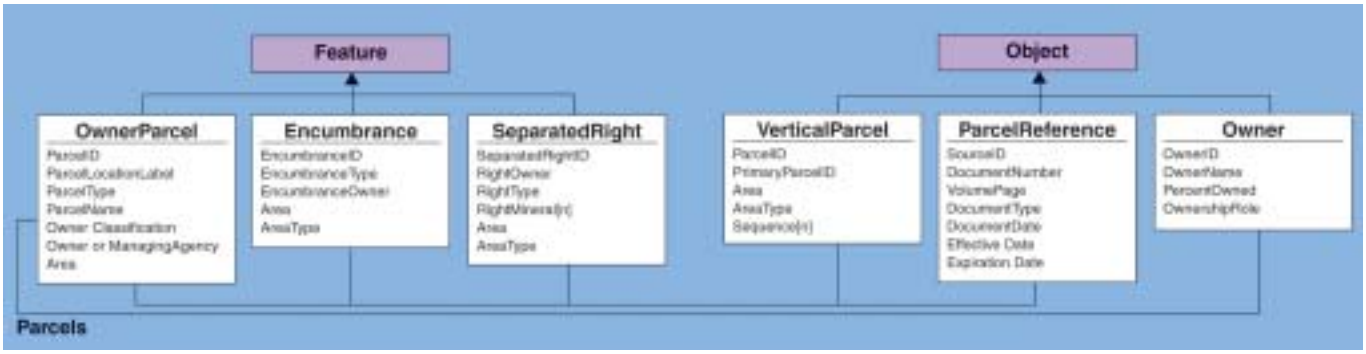


FIGURE 4 The ESRI ArcGIS Land Parcels Data Model uses UML to illustrate how parcels and owners can be stored in a geodatabase.

A data model's ability to anticipate events and automate responses to them inspires awe on first viewing. Remember that demonstration at your last conference in which the presenter moved a utility pole and the wires, transformers, and everything else connected to the pole moved with it, automatically? Later the presenter created a new utility pole and, again automatically, new transformer objects popped into existence, attached to that new pole.

Whether established by hand, as triggers and procedures in the database, or with a wizard-driven user interface, such as ESRI's ArcCatalog, these seemingly magical animated behaviors spring to life based on rules and relationships described by data models. Combining relational databases with spatially-aware application logic moves data from its former static flat file days to a complex "object" with built-in intelligence. As ESRI Data Modeling Specialist Steve Grise summarizes, "Thanks to object-relational technology, we have a basic working system that needs very little customization and does not need a traditional monolithic application."

Code	Description
01	Collision With Motor Vehicle
02	Collision With Pedestrian
03	Collision With Bicyclist
04	Collision With Animal
05	Collision With RailRoad Train
06	Collision With In-Line Skater
10	Collision With Other
11	Collision With Light/Support Utility
12	Collision With Guard Rail
13	Collision With Crash Cushion
14	Collision With Sign Post
15	Collision With Tree
16	Collision With Building/Wall
17	Collision With Curbing
18	Collision With Fence

FIGURE 5 The ESRI ArcGIS Transportation Model's coded value domain for accident type lists 32 different flavors of road accidents.

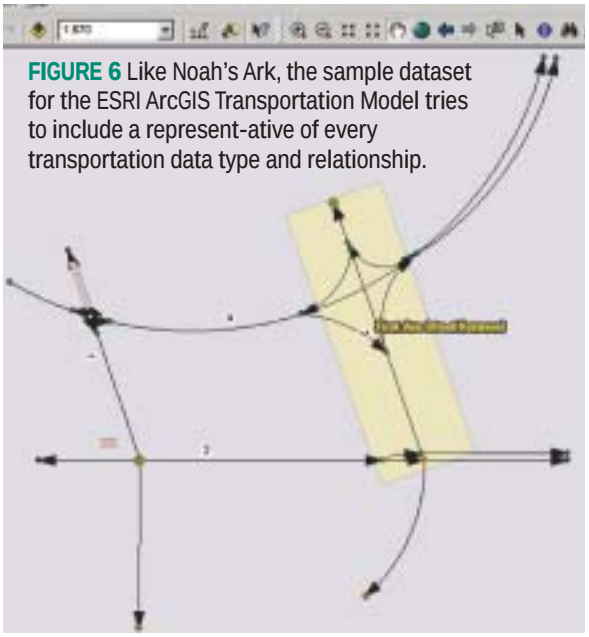


FIGURE 6 Like Noah's Ark, the sample dataset for the ESRI ArcGIS Transportation Model tries to include a represent-ative of every transportation data type and relationship.

Prewashed designer data

Even if an object-relational approach is overly complex for your application, a model's geometric representation of reality is only one of its elements. For many spatial professionals, the lists of value domains may be the most helpful data model component. For instance, how many different kinds of traffic accidents can you think of? There are 32 accident types offered in the ArcGIS Transportation Data Model (see Figure 5) including "collision with in-line skater" (a Darwin award candidate?). Getting all the possibilities into your database schema before putting the system into production prevents time-consuming enterprisewide changes later.

Furthermore, value domains can protect a dataset from inconsistent data entry and conserve disk space. Using the traffic accident types as an example again, if the people entering accident reports into the database select from a list of accidents rather than type in their own descriptions, the records will remain consistent. Plus, storing the lengthy text descriptions, such as "collision with Earth element/rock cut/ditch," only once in a lookup table and substituting a brief numeric code everywhere else results in smaller disk space requirements and faster searches.

From Missouri?

Residents of the Show-Me State will be happy to know that data model documentation sometimes includes small sample datasets. The ESRI transportation data model's sample (http://arconline.esri.com/arconline/datamodels/transportation) has a little bit of everything — bridges, freeways, streets, traffic events, and more — attached in a very small network (see Figure 6). The data unzips as two Microsoft (www.microsoft.com) Access .MDB files, directly readable by ESRI's ArcCatalog. All tables, whether empty or holding data, are included. Because data model samples ship complete with all tables and relationships intact, the curious can add their own data to any part of the model to test performance.

Inside out and outside in

At first glance, I assumed that the best use of a data model was to preserve the integrity of producing spatial data. The data models' authors want their work to be useful at that level, but also see wider integration opportunities.

Nancy von Meyer notes, "The publication [rather than the production] environment presents the best opportunity for building national consistency. Each organization has different structure, staff skills, and hardware, but if they transform their production information to fit a nationally consistent model, then their published data becomes available for much wider integration." In her own area of expertise, the cadastral community, von Meyer contends that "the national cadastral spatial data infrastructure is [already] here; realizing its potential is a matter of consensus on publication formats, recognizing that there doesn't have to be just one, and assisting local governments in generating the formats." Thus, implementing a data model that is consistent across your enterprise may have both internal and external value.