

Title: Data Modeling: Harvesting Trees of Information and Knowledge (Paper #2130)

Author: Jack Haefner, Major, US Army (NGA Release Case #07-135)

Abstract

The world is drenched with data. But how does one make sense of this forest of figures and derive trees of information: from root to tip, to fuel future systems, to drive our decisions, and to bring together and engender our mutual geography which is so very interrelated?

The Army Theater Geospatial Database (TGD) is but one example of how a well-structured logical data model can be swiftly taken from drawing table to physical implementation. As a result, the TGD fuels the efforts of others as they share a common core of data understand with others on a worldwide scale.

To further evolve these data modeling efforts, the Geospatial Intelligence¹ Standards Working Group (or GWG) is one such governing body to provide mutual purpose across the community. Thus, current data modeling efforts are progressing to inculcate geospatial intelligence across the enterprise domains of acquisition, engineering, data production, and data sharing.

Drowning in Data

When it rains it pours, for we now live in a world drenched with data. But how does one make sense of this forest of figures? Is it ever possible to have too much data? Most of us would always seek more—not less—data, but what are the implications? Yet even for simple, univariate information, humans are limited in their ability to sort through these immense volumes of data.² As a result, as our capability increases to gather, process, and store data (or gather, store, and process), so must our understanding of the greater data challenge.

So, if we're stuck with this challenge, how do we make sense of all this data? First of all, there must be an understanding that no data exists solely for itself, for every piece of data has a relationship to a larger geospatial fabric of immense design. For example, something as simple as electrical current is generated, stored, managed; it then flows through distribution lines, down power lines, through transformers that sit on poles, and through electrical service to consumers that fuels our lights, computers, and lives (and power bills). Even without physically touching that electrical power, it has touched many nodes, processes, and people on a larger geospatial

matrix. In other words, there is always a direct interconnection between each piece of data we may wish to collect—this is the essence of topology.

Secondly, there must be an understanding of persistence in our data context, data pedigree, and data topological relationships. In other words, these data types, objects, and attributes of today might no longer exist or be valid/relevant in the near future. For instance, as the US surveyed its western expansion over the last half of the 19th century, who would've foreseen that the automobile would assume the role of king from the railroads within the span of a few decades? As we inculcate these topological relationships at the map or database level, we need to bring focus to what data (both in structure and content) will change quickly and what will persist. It might be prudent to maintain some extremely temporal data elements outside of a larger data model, or perhaps build a data model which incorporates perhaps 80% of the uses. An example of this might be map markup (or what we call maneuver graphics), tracking of human intelligence, etc. In contrast, fairly mature features such as landforms, roads, and bridges are well-defined, temporally stable, and easy to model as inclusion candidates.

Yet there are development trade-offs for under- or over-designing a data model. Design function for a data model must bear these engineering trade-offs in mind. Not effectively data modeling in an enterprise environment (and we are all part of the enterprise in one way or another) is akin to developing your own spoken language with no regard to ever learning another. It is really best to understand the entire user base and incorporate the larger community³ adoption before full-scale development.

Data Modeling

The data modeling process allows us to make sense of all this data. Although an oversimplification, data modeling is much like organizing a silverware drawer: spoons go here, and forks go there. Of course, data modeling geospatial data is a bit more involved than a few minutes of cutlery organization. Yet, if the model is simple, elegant, understood, and socialized, economies will emerge.

One way of approaching the data modeling process is as it were tax preparation where it's best to have a system before you begin the next tax year than to do it all at the end. Data modeling is really the same where first you must have an understanding of where you are, then

where you want to get to, and then fill in the gaps. If you're like most people at tax time, you start with one big tax file. Yet, once you devise a system where everything has its place (such as income, expenditures, investments, etc), the process flows and you are finished (and have a return). This would surely be an easier process if one were proactive from the start.

Building the conceptual data model (CDM) is the first step in the formal standards development process.⁴ In effect, this involves describing all business flows, business rules and developing an understanding of what is needed from the data.⁵ At this stage, data is only described in general terms and taxonomy is often not formalized. The output of this stage is normally an entity-relationship model. (see Fig 1.)

Stage two requires logical data model (LDM) development where data schemas and attribution is formed to ultimately create the building blocks for final translation into particular database software.⁶ The output at this

step is normally a schematic diagram, most often in a Unified Markup Language (UML) construct using tools such as IBM® Rational® Rose® and Microsoft® Visio®.(see Fig 2.) Admittedly, most people would probably look at an UML diagram and suffer prompt headaches or drooling. For this reason, most UML diagrams are often translated into a Microsoft® Excel® spreadsheet which tells most of the story.

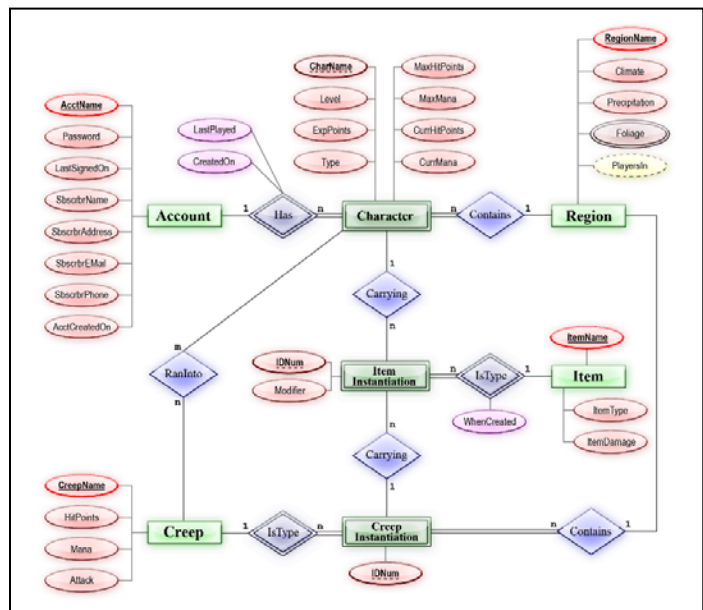


Figure 1. ER Diagram (example from “Entity-relationship model.”)

It is not until the last step where we construct the physical data model (PDM).

Where the LDM doesn't specifically account for the particular hardware or software on the user end, the PDM is the material instantiation of the data model using, for instance, ESRI® ArcSDE® Geodatabase™ on Microsoft® SQL Server 2003®. Of note, ESRI® ArcCatalog™ migration tools allow users to move from UML rendering of the LDM to the PDM fairly quickly, with accounting for designing domains within the geodatabase.

Implementing Existing Standards

Rapid program development is one compelling case for using accepted geospatial logical data models. For instance, the US Army, Pacific and US Army, Europe Theater Geospatial Database (TGD) is an example of how a well-structured logical data model can be swiftly taken from drawing table to physical implementation.

The problems faced in 2003 by both Army geospatial analysts in both Pacific and European theaters resonate with us all. Before any analysis occurred, there was a large physical effort spent just researching data, establishing project folders, and applying archaic naming schemes. To make matters worse, this data mining process often queried unchanged, previously mined data sources. Moreover, since little metadata⁷ was documented,

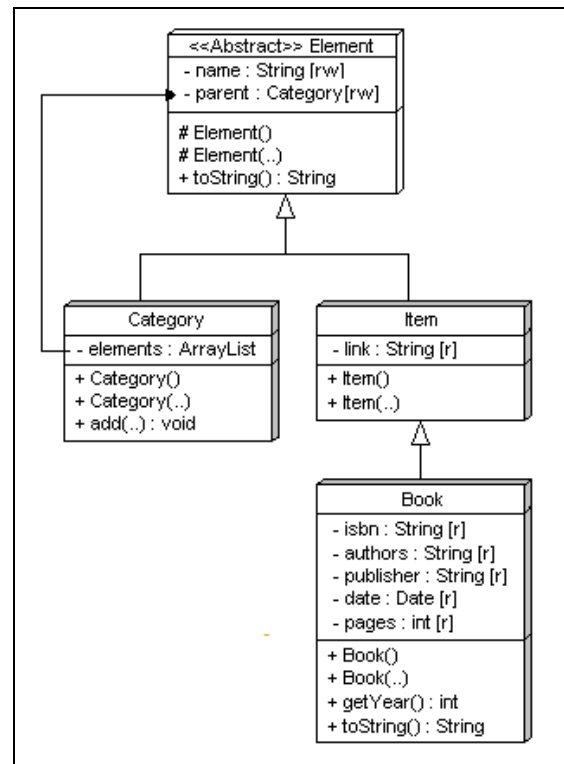


Figure 2. UML Diagram (example)

or the source lacked surety, a single data set might have been replicated multiple times over several projects. In essence, what was needed in US Army, Pacific and US Army, Europe was a migration to a truly enterprise geospatial intelligence system to:

- Collect/extract a feature one time and use many times;
- Create an understanding of the data by documenting the metadata;
- Have this data form the trunk from which all analysis and products are derived;
- Allow data check-out, subsequent improvement (or value-adding), and check-in capabilities;
- Documented workflows using a data schema where connected and disconnected users knew—and trusted—the details of the data schema;
- Replicate this data to national agencies where it could be leveraged across the defense and intelligence community.

These were not seen as lofty goals. But it didn't take a room full of systems analysts to see across the community—not just at two small geospatial analysis shops in Hawaii and Germany—that we were very inefficient in our workflow, redesigning the car from the ground up every time we needed a ride.

In effect, building the TGD was really nothing more than taking an existing documented LDM/CDM and producing a PDM instantiation. The TGD PDM was created from FACC (Feature Attribute Coding Catalog)⁸ as overseen by the Digital Geographic Information Working Group (DGIWG©).⁹ Moreover, our version of FACC was in use by NGA and others throughout the community. Thus, we felt that, by using the FACC+ LDM, the TGD could not only talk the same data language within its community, but we could also ship (value-add) data back to the national level (the National Geospatial-Intelligence Agency—NGA). Also as part of the TGD development effort, standards of extraction and workflow were documented, with the FACC data dictionary being used as authoritative source.

Since its 2003 origin, the TGD has been largely successful since it is 1) based on a known standard, 2) is a simple design, and 3) has a level of transparency to allow data sharing and mutual value adding (to lower-level producers as well as up to national producers). All this was completed, from concept to implementation—across both the Pacific and European theaters—in seven months.¹⁰ Clearly, this would've been a much onerous undertaking absent an existing LDM.

Yet, the TGD is not an existing Program of Record (POR). It was designed by the Army Service Components of both the US Pacific Command (USPACOM) and US European Command (US EUCOM) as an *interim* solution to facilitate an Army and Joint workflow until which time the NSG would have an enterprise geospatial system for increased efficiencies and federated/unified operations. As such, TGD data model maturation does not influence any other standards development. Maturation of the TGD data model is accomplished by inclusions and modification of the existing data model in annual and semi-annual working groups. After vetting, the contractor under data model development (ESRI®) makes the necessary changes and delivers the new data model and migration to same at both Hawaii and European locations. This level of maturation is relatively simple due to 1) a small, focused user group and 2) use of existing, off the shelf, standards.

Emerging Standards

Although FACC+ was a currently a documented LDM at the time of TGD development, development stopped in 2001 and community usage of the LDM ended in 2006 in favor of the Digital Feature Data Dictionary (DFDD; Baseline 2006-2). Due to documented needs within the community and the Army Future Combat Systems (FCS) program, the overarching NSG Entity Catalog (NEC) is now the emergent US geospatial standard. Consisting of the NSG Feature Data Dictionary (NFDD), the NSG Feature Extraction Guidance (NFEG) and other documents, the NSG Entity Catalog (NEC) forms a new community-wide emergent standard to allow a common ISO 191XX language of features and attribution for data production, dissemination, and data fluency throughout the community.¹¹ We fully expect at some point the TGD will migrate and fully adopt the NEC construct.

The Geospatial Intelligence Standards Working Group (GWG)

Requirements for the NEC is a fairly broad defense and intelligence community governance body—the Geospatial Intelligence Standards Working Group or GWG. The Department of Defense (DoD) Information Technology Standards Committee (ITSC) oversees the GWG and executes governance for “developing and promoting standards interoperability in support of net-centricity within the Department of Defense (DoD).”¹² The chief responsibilities of the GWG are to document GEOINT standards in the DoD IT Standards Registry (DISR) and to chair the GEOINT standardization forum.¹³ The GWG is comprised of eight focus groups¹⁴ that meet on a routine basis.

As the GWG forms the funnel for NSG standards development efforts, within the Army all geospatial standards must be vetted through Headquarters, Department of the Army before they are brought to the table at the GWG. The GWG votes these standards in technical working group sessions before they are brought in to the NSG¹⁵ for development. The voting pool is purposely limited with the Army getting only one vote among the group. All draft and final standards are posted in the standards registry (DISR).

Are we called to create one mother of all geospatial data models? Few would argue that our human geography—or interconnectedness of how our geospatial communities operate—does not matter or is not for a common good. However, those in this business of data modeling would argue readily that one size does *not* fit all. This is driven by the user base limitation—as adoption

is forced across a larger and larger segment, specificity decreases and full acceptance may decrease. What is important is to find these touch points or points of interconnection and begin the dialogue. In effect, we find that working across community segments, the highest level of satisfied participants only hovers around 80-90%. As a result, some parts of the data model can remain generic so that users can extend it later for their specific use case for the good of both the model and the community writ large.

Standards: Only a Slice of the Pie

Standards development is only a portion of the overall geospatial enterprise development effort. Likewise, there are a number of attendant challenges. First of all, there are immense cultural and practical differences within the Army between geospatial, operational, and modeling/simulations communities. Are there any geospatial analysts who haven't heard that "my Xbox can do that—why can't your GIS?" It's a world of difference between the synthetic world and that of operational battle space where lives are at risk. Moreover, much of this battle space is in areas where we are denied entry for obvious reasons. Sure, there are opportunities for a geospatial data model to share much with the virtual simulation data entities, but only the hard work of give-and-take will rue the day.

User communities also matter. There was a time when only GIS professionals touched a GIS or performed "computer mapping." Today, roles have shifted greatly as, for better or worse, viewers such as Google Earth™ are available to a broad client base. Perhaps we need to widen the scope, where our collector base is large, professionals conduct data management and some hardcore analysis, but, given well-documented metadata, dissemination is ubiquitous. Of course, technologies such as ArcServer™ can offload some of the more basic GIS analyses to leave the analyst more stringent tasks such as understanding and managing these data challenges and idiosyncrasies. Many of the Army's thoughts are obviously in this direction, but the "heavy lifting" implications of geospatial data needs considerable scrutiny: from conflation, to indexing/database tuning, and the tough work of management.

Lastly, much thoughtful consideration must be given to bringing the geospatial data model to other Army programs of record. There are still many systems that retain their own proprietary, closed technologies, data models, and data sets. We often times find that, to get these systems to function properly, one must take fairly well-formed and logical (and often

“standard”) data and redesign it just to feed this propriety system. That said, it is hard to imagine a user community within the Army that doesn’t have a stake in an Army Geospatial Data Model, but still there are many communities out there managing on their own.

Charge for the Future

It is an untenable position to not invest in geospatial data standards development. In fact, both the sheer volumes of data and net-centric operations demand deliberate and thoughtful data modeling development efforts. Where initiatives such as the TGD bear witness to the richness that a data model brings with regard to workflow development, value-adding, and feeding dissemination, we all need to continue to take these standards issues seriously. In an “all hands” manner, it is the yeoman’s work of working these issues together and then socializing the outcomes around the community that we will guarantee the information edge.

Author Information:

Major Jack Haefner is the author of several ESRI® User Conference papers including Theater Geospatial Database Value-Adding in a Distributive Environment and US Army, Pacific Theater Geospatial Database: A Model for Synergistic Geospatial Intelligence. He has served in geospatial positions both with NIMA/NGA and the Pacific. He currently serves with NGA Support Team to the US Army in Reston, VA.

Contact:

jack.haefner@us.army.mil

¹The term “geospatial intelligence” or GEOINT is “the exploitation and analysis of imagery and geospatial information to describe, assess, and visually depict physical features and geographically referenced activities on the Earth. GEOINT consists of imagery, imagery intelligence, and geospatial information.” Title 10 U.S. Code §467 establishes the definition of GEOINT. The document also provides an overview of the Functional Management/NSG role in leadership, guidance and functional management of GEOINT — as defined in DoD Directive Number 5105.60, Director of Central Intelligence Directive 1/8 and DNI Memorandum E/S 00245 — as well as roles and contributions of the constituent members of the NSG Community. As the NSG functional manager of GEOINT, NGA is responsible for GEOINT data standards design.

²This human limitation is somewhat misleading. Why is it that, after we have compiled extensive databases, performed complex analyses and the like, an experienced individual can look at the same piece of terrain and summarize all these interrelations faster than most machines? This is a truly amazing human capacity that should never be relegated to only machines; it underscores the importance of the human dimension when approaching our mutual geography.

³ Community and NSG (National System for Geospatial-Intelligence) is often used interchangeably. The National System for Geospatial-Intelligence (NSG) is “the combination of technology, policies, capabilities, doctrine, activities, people, data, and communities necessary to produce geospatial intelligence (GEOINT) in an integrated multi-intelligence, multi-domain environment. The NSG includes the Intelligence Community (IC), the Joint Staff, the Military Departments (to include the Services), the Combatant Commands (COCOMs), international partners, National Applications Office, Civil Applications Committee members, industry, academia, Defense service providers, and civil community service providers. (The National System for Geospatial Intelligence, Statement of Strategic Intent, March 2007)

⁴ It is worth noting that often times part of this process is reversed engineered based on need, especially when formerly divergent communities rejoin efforts or when pulling in the reigns on runaway development.

⁵ G. Lawrence Sanders, Data Modeling, (Danvers: Doyd & Fraser, 1995) 11

⁶ Sanders 12.

⁷ Metadata is “data about the data.”

⁸ More succinctly, FACC+ was used which is the baseline FACC with US/NGA specific extensions to that data model. FACC is documented in the DIGEST (Digital Geographic Information Exchange Standard).

⁹ The DGIWG© “was established in 1983 to develop standards to support the exchange of Digital Geographic Information (DGI) among NATO nations. The DGIWG© is not an official NATO body; however, the DGIWG's standardization work has been recognized and welcomed by the NATO Geographic Conference (NGC). Interoperability and burden sharing among nations are the goals of the group.” (<https://www.dgiwg.org/digest/About2.htm>)

¹⁰ Jack Haefner, US Army, Pacific Theater Geospatial Database: A Model for Synergistic Geospatial Intelligence, (ESRI International Users Conference Proceedings, 2004)

¹¹ Given the new NEC, there is still a degree of architecture, tradecraft, and workflow/business process development in order to use the NEC standard within the NSG and allow value-adding and federated/unified operations.

¹² Geospatial Intelligence Standards Working Group (Public Site) <http://www.gwg.nga.mil>

¹³ *ibid*

¹⁴ The GWG focus groups are: NITFS Technical Board (NTB), Motion Imagery Standards Board (MISB), Community Sensor Model Working Group (CSMWG), GEOINT Reporting Focus Group, Application Schemas For Feature Encoding (ASFE), Geographic Portrayal Focus Group (PFG), Metadata Focus Group (MFG) Metadata Focus Group (MFG), and the Information Transfer & Services Architecture Focus Group (ITSA FG)