

Modeling in Geographic Information Systems: Data Amalgamation and Integration

Allison Denby, B.A. GIS
Geographic Information Systems Consultant
Integrated Informatics Inc.
1313 7th Street SW
Calgary, AB T2R 1A5
403.829.7575
amdenby@integrated-informatics.com

Jason Humber, B. Eng.
Principal Consultant
Integrated Informatics Inc.
1313 7th Street SW
Calgary, AB T2R 1A5
403.619.8926
jlhumber@integrated-informatics.com

Abstract

Significant advances in the fields of remote sensing and surveying have increased greatly the variety, availability, and quality of spatial datasets. These advancements have led to new developments in Geographic Information Systems (GIS) allowing for broader use of GIS across the enterprise in an integrated manner.

The purpose of enterprise level integration is to model processes effectively and realistically. Even with the improvements in dataset availability and quality, data integration across the enterprise still requires considerable vigor and rigor. In order to produce realistic model outputs, it is necessary to amalgamate data from multiple sources and quite often occurs between datasets having different geometries, data types, accuracies, and resolutions. Overcoming such data diversity through amalgamation means that the amalgamation process must involve a well-devised and robust plan; a plan that will draw upon many specialized spatial data management techniques including georeferencing, geoprocessing, data calibration, data correction, extraction, reconstruction, classification, and much more.

High quality fused datasets are becoming increasingly important to companies turning to GIS for real-world models and analysis. This paper will focus on modeling in a GIS to support Emergency Response in the Oil and Gas industry. The keys to these models and analyses are held in data awareness, understanding, quality, validity, and integrity to ensure high quality and realistic results.

Introduction

Due to considerable advances in accuracy and resolution of datasets, GIS models now obtain the ability to converge with reality and can significantly support companies in accurate decision making. Models that are based solely on a single dataset typically lack robustness due to the complex processes found within the world around us.

Data amalgamation has drawn great attention in recent years and the ability to combine various datasets, from disparate sources, and in widely varying formats into a standardized single composite data set is a basic precept of GIS today. Unfortunately, there has been little research analyzing this process as a whole and no theoretical benchmark against which to determine the capabilities of combining data of differing types and accuracies. However, by following rudimentary processing steps, a successful output can be assured. The benefits drawn from data amalgamation are dependant on the procedures and techniques applied in the process.

Organizational Enterprise Solutions

Whether modeling in real-time or for decision making, an overall GIS plan for the organization is imperative to successful implementation. This organizational plan, or company protocol, determines what and why spatial datasets are important, how they will be collected, used and stored, and ensures that the results are properly interpreted and implemented. **Figure 1** outlines GIS in an institutional context. The company's effective use of GIS is dependant on all aspects of the life cycle. [1]

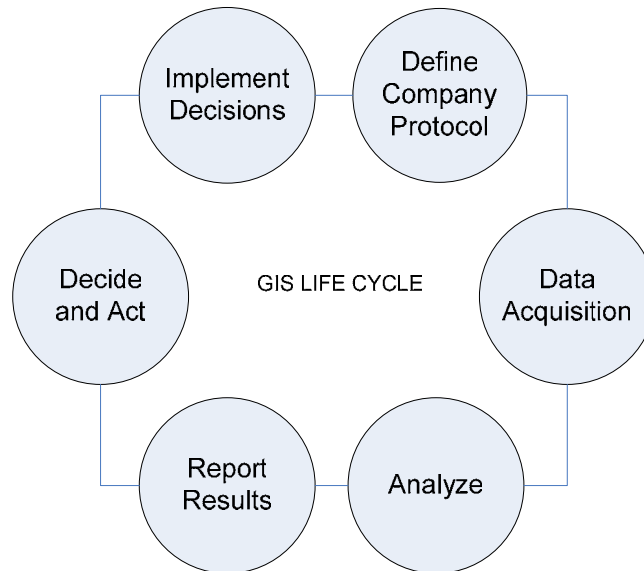


Figure 1: GIS Life Cycle [1]

Throughout the years companies have accumulated data in varying forms and formats and as technology has evolved so too have data collection and storage methods. The goal of the acquisition phase is to gather necessary and relevant datasets across the organization in a manner which lends itself to support over all business goals.

Once the acquisition phase is complete the largest challenge appears in the integration phase. This phase will set the stage for the effectiveness of the ability to amalgamate many different datasets. Data integration steps include determining the location where data will reside, development of a common reference system, and building a data dictionary. The development of these crucial components will allow data elements from various sources to be combined and accurately associated with other data elements.

The common reference system applies to the spatial reference and in some cases the linear reference system (a system for referencing two dimensional coordinates in a one dimensional system). The spatial reference system allows integration through X and Y coordinates. There are many coordinate systems available and a decision needs to be based on the organizations data extents and what aspects of the data are important for retention (i.e. length, area, direction, or shape).

Once a common reference system is established, data can begin to be logically integrated. The organization will need a mechanism to physically assemble data into useful information [2]. The typical solution is a data repository or a common database. Although the integration process is lengthy, once complete the organization can better manage new data and utilize updated data.

Data Amalgamation

For analysts to be able to create products, formulate analysis, and extract value from the enterprise data sets they must have an solid understanding of data mining and competency with processes that support specific operations. Analysts must understand the integrity of the data to include data sources, data structure and format, and spatial and/or spectral accuracy. Lastly, the analyst must understand the types of products that will be produced. [3]

Quality assurance rests on the shoulders of the analyst because it is their responsibility to situate checks and balances to ensure that no data error has been introduced. The single largest caution with data amalgamation is proper combination of data with different geometries, resolutions, accuracies, and of different type. Depending on the end goal of the operation, the user of the inputs must know how this affects credibility of the output. Moreover, the data must be measurable, quantitatively and qualitatively to assure integrity of the product produced.

For example, the process of merging two disparate datasets will produce a new set of metadata based on both inputs. Of particular concern, is what will the new accuracy be? There are many measurements of accuracy for spatial data. Digital forms typically have an accuracy statement associated with them, especially raster data. If this is the case for both datasets, then it is still a good practice to compare the imagery against another source. The comparison would also be the result of a qualitative analysis. If the accuracy or resolution of the data differs, the lower resolution will be the new accuracy.

Looking at another example, a digital map that was created at the scale of 1:100,000 and compared to raster data with 30 meter cell size will match up at certain scales. The two maps compared would likely match up at 1:100,000 scale, but if compared at a scale of 1:24,000, it would appear that a mismatch occurs. The analyst must be aware of how the products are created and the intended use of the dataset. The user must be aware of the potential use of original source datasets.

Timeliness of datasets offer concern to the amalgamation processes. When more than one dataset is used, they are often of different vintage. Everything around us changes at a faster pace than data is updated. For this reason, caution must be taken by the analyst when processing data. It is a case specific instance because for certain analysis the vintage of datasets may be irrelevant, while for other the analysis, vintage may be the top priority.

Emergency Response – the need and benefits of GIS for modeling

Analysts must be aware of the different data products that may be necessary for various spatial operations. In the case of an oil spill, emergency response regulations make it quite evident that GIS is the imperative modeling choice. Although not explicitly stated, the pipeline regulators in North America imply that GIS be used to determine affected high consequence areas (HCA).

The Department of Transport (DOT) in the United States has set a regulation, CFR 195 that clearly outlines the need for comprehensive identification of HCAs that could be affected directly or indirectly by a pipeline leak or spill. However, expressing this need is much simpler than the rigor required to answer the regulatory requirements in full. On the left side of **Figure 2**, are the extracted mandatory considerations from CFR 195. On the right side is the GIS interpretation of the recommendations as input datasets for an overland spill model.

Regulatory Recommendations	Model Considerations
1.Terrain surrounding the pipeline...	1.Digital Elevation Model
2.Drainage systems...	2.Hydrography
3.Crossing of farm tiles...	3.Crossings
4.Crossing of roadways...	4.Crossings and transportation
5.Product characteristics...	5.Fluid properties
6.Physical support...	6.Special Case
7.Operating conditions...	7.Operating conditions
8...Hydraulic gradient...	8.Hydraulic gradient
9...Potential release volume...	9.Potential release volume
10.Potential physical pathways...	10.Potential physical travel paths
11.Response capability...	11.Response time
12.Potential natural forces...	12.Probability

Figure 2: Regulatory Recommendations versus Model Considerations

Now from a regulatory standpoint, all considerations have been met. In order for the model to converge with reality there are additional datasets that need to be included. Additional factors that should be accounted for are surface cover, soil properties, and weather conditions. All of these factors will greatly affect the movement of liquid over land. For example, excluding surface cover would lead to a constant surface for liquid to travel. Obviously, the earth’s surface is more complicated than that. When including a dataset of this nature, one must remember the importance in accounting for accuracy. Most of the time, resolution of a land cover dataset is thirty meters. In raster datasets, only one value can reside per cell, therefore the highest accuracy of land cover would be thirty meters.

In designing a realistic model, consideration must be given to data availability and data quality. Higher quality datasets (i.e. model components) produce results that are more realistic. Resolution and timeliness of data are two major problems. Resolution can be easily accounted for and quantitatively justified in the outputs. In order for an organization to use the model outputs for decision support, a quality assurance test must be implemented. These reasons are enough that careful considerations need to be taken during the data amalgamation stages. The temporal aspect is fuzzier in determining the quality and relevance of data. The current data completeness depends on the developments or happening in the area. Some areas around the globe change at an expedited rate (ex: river banks, urban areas), while others areas change at a very slow pace (ex: farmer’s field).

Another issue that is broader than just this model is the availability of quality data. An effective model must be adequately adaptable to account for unavailable data, poor quality data, or for the use of overriding assumptions. A comprehensive model bases its methods on establishing realistic resistance to flow over ground surface and its algorithm creates a continuous surface of fluid flow by accounting for variations in surface cover, directional slope, fluid properties, and other inputs.

Once all the appropriate datasets are collected, preprocessing of data is required. This is where the actions of the analyst become extremely important. One simple miscalculation of the preprocessed data can propagate error throughout the remainder of the analysis. This particular model runs in a raster environment, therefore the first step is to convert vectors to raster. Once this is complete, reclassification and derivation of datasets must occur. Reclassification occurs on datasets such as land cover and soil properties. Datasets that needs to be derived are primary topographic attributes and finally the generation of a cost surface (the cost for liquid in effort to travel across a surface).

With the preprocessing complete, the model can then be run. Once the model produces outputs (**Figure 3**), efforts need to be shifted to data validation. If amalgamation steps were carefully followed, the output is ready to use within the organization to facilitate decision support and, in this case, produce results mandatory to the Department of Transport.

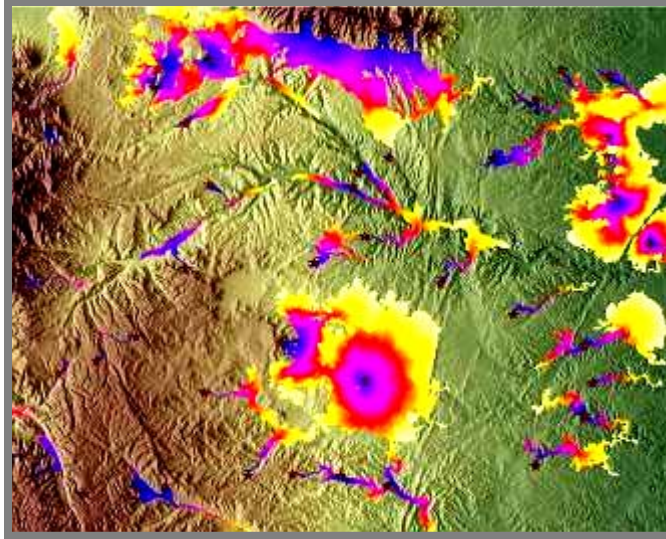


Figure 3: Overland Flow Model of Water from various spill points

Conclusion

Data amalgamation is quite the powerful process if completed procedurally correct. The ability to combine datasets gives new power to modeling within the GIS environment. Model outputs allow convergence with reality due to modeling with multiple highly accurate datasets. The advances in GIS are opening doors to new industry and process reengineering of current industries. With trained analysts following correct procedural steps in spatial modeling and amalgamation, it will only increase the reputation of GIS as a reputable source for decision making within organizations.

References

- [1]Bolstad, Paul, "GIS Fundamentals", May 2001, Eider Press: White Bear Lake, MN
- [2] Alexander, Jeff, "Integrating Data for Pipeline Compliance Part 1 - Concepts", September 2004, 13th Annual GIS for Oil & Gas Conference & Exhibition
- [3]Ware, Jared, "Geospatial Data Fusion: Training GIS for Disaster Relief Operations", National Geospatial Intelligence School

Biography: Allison Denby

Allison Denby joined Integrated Informatics in May 2004 as the lead Geographic Information Systems Consultant. Her activities within the company encompass many aspects of spatial data management including database development, analysis techniques, research and development for raster based modeling, and GIS implementation.

Allison has successfully completed her Bachelors of Applied Geographic Information Systems degree. Her work background is varied and has involved projects ranging from promotion of geomatics and GIS education awareness across Canada to process improvement for pipeline integrity data integration.

Biography: Jason Humber

Jason Humber founded Integrated Informatics Inc. in October of 2002 to provide data management and system design consulting services to the pipeline industry for new construction, operations, and pipeline integrity. As a Principal Consultant, Jason is responsible for corporate level development and delivery of Integrated Informatics unique suite of services.

In 1999, Jason began his career with the Natural Gas Business Unit of BP Canada Energy Company in Calgary, Alberta. His primary focus within BP was development of a business unit wide data management system that supported the analytic and integration needs of pipeline integrity. While working with BP, Jason also took on a pivotal role on the project management team within the Alaska Gas Producers Pipeline Team and helped to establish the processes required for Project Data Management. More recently, Jason has completed a similar advisory role with the Mackenzie Gas Project, and has broadened and implemented these data management approaches to encompass the needs of developing oil sands projects.