CHALLENGES IN CREATING A UNITED STATES GEOSPATIAL DATA FRAMEWORK

Stephen C. Guptill

U.S. Geological Survey
519 National Center
Reston, VA 22092 USA

Abstract

In the United States, plans for the National Spatial Data Infrastructure and a national geospatial data framework are based on the idea that digital spatial data collected by various producers can be integrated into a national coverage. By creating an environment of cooperative data production, agencies may be able to accommodate today's requirements for geospatial data.

Although the applications can vary greatly in geographic area, purpose, and content, the information requirements usually include several basic, consistent themes of data. This information, which includes geographic features such as roads, railroads, streams, and lakes, governmental units, cadastral data, and elevation, provides a framework for data collection and analysis. It also provides a base on which an organization can accurately register, compile, and integrate other themes of data.

The framework approach of collaborative data production should encourage many organizations to contribute to its construction and maintenance, but this framework will be viable only if significant technical barriers can be overcome. Technical advances that use permanent feature identification and tracking, object data bases, transaction processing, and multiple scale representations of features will make a geospatial data framework possible. Such technology, coupled with various content, quality, and procedural criteria, will make it feasible to produce geographic data that are valuable to a maximum number of users within any geographic area. A major institutional challenge is to create economic or policy incentives that encourage the collaboration of partners in the creation and maintenance of the framework.

1 Introduction

Business and government agencies are moving into an electronic environment for managing information and providing data and services. For most people, this conjures up a vision of thousands of computer files of textual reports and spreadsheets. However, an alternate vision exists, one filled with digital maps and electronic images of the Earth's surface. This is the conception of spatial data that is being incorporated into the broader discussions of a national or global information infrastructure.

Geospatial data, or data tied to locations on the Earth, are critical to solving today's complex environmental, economic, and social problems. The now pervasive use of geographic information system (GIS) technology for analyzing spatial problems has created a demand for vast amounts of digital geospatial data. These data, describing the characteristics of geographic space, must be current and of known quality. The data must be accessible and able to be manipulated with predictable results. Meeting all these requirements is difficult.

Although geospatial data can vary greatly in geographic area, purpose, and content, the needs for such data nearly always include a few basic themes. This information, which includes geographic features such as roads, railroads, streams, and lakes, governmental units, cadastral data, and elevation, provides a framework for data collection and analysis. Such *framework data* can orient a user and link the results of an application to the landscape. Framework data provide a base on which an organization can accurately register and compile other themes of data. They provide the geospatial foundation with which an organization can perform analyses and to which it can attach attribute information. These framework concepts have been developed by the Framework Working Group of the Federal Geographic Data Committee [1].

The concept of framework data was endorsed by the President of the United States in Executive Order 12906. The order instructs agencies to complete the initial implementation of a national geospatial data framework:

> *In consultation with State, local, and tribal governments and within 9 months of the date of this order, the FGDC [Federal Geographic Data Committee] shall submit a plan and schedule to OMB [Office of Management and Budget] for completing the initial implementation of a national digital geospatial data framework ("framework") by January 2000 and for establishing a process of ongoing data maintenance. The framework shall include geospatial data that are significant, in the determination of the FGDC, to a broad variety of users within any geographic area or nationwide. At a minimum, the plan shall address how the initial transportation, hydrology, and boundary elements of the framework might be completed by January 1998 in order to support the decennial census of 2000.* [2]

To date, because of inadequate investment, insufficient institutional arrangements, and the lack of a common technical approach, the spatial data producers of the United States have been unable to combine their resources to develop and maintain the framework data. The strategy called for by Executive Order 12906 is to establish an environment of collaborative production of framework data, where the potential exists to amass adequate resources to satisfy today's requirements for geospatial data.

2 Framework: Goals and Characteristics

A collaborative approach to creating framework data requires that many organizations contribute to its construction and maintenance. To encourage participation, the framework should evolve in response to the contributors' changing requirements and capabilities. The framework will be operated and maintained by participants who agree to provide digital geospatial data that are certified to meet various content, quality, policy, and procedural criteria.

The framework is a basic, consistent set of digital geospatial data and supporting services.

- It will contain the "best"[1] data available. The framework data should represent real world features (and not cartographic symbols). It should incorporate the high-resolution data

[1]The idea of "best" data, however, is a complicated one. Different applications require, or at least tolerate, different mixes of the qualities normally associated with the idea of "best" data: currentness, positional and attribute accuracy, consistency, and completeness.

collected by local governments, utilities, field offices of State and Federal agencies, and others.

- It will include consistently generalized, lower resolution data that are needed for regional or national studies. These data should be produced from higher resolution framework data. Links or references among different representations of features should exist.

2.1 Operational Context

In addition to data, the framework should provide the following operational characteristics:

- The framework must support transactional updating so that producers only provide change files and users only process changes. This approach reduces the effect of change on existing investments.
- Access to an official version of framework data (current and past versions) by information networks and digital media must be ensured.
- Updates to framework data should preserve investments in existing data to the maximum extent possible. Permanent identifiers should be changed only when necessary.
- Users should be able to find any data through a consistent data clearinghouse mechanism.

In addition, data contributions will cover a logical minimum areal extent so that the resources required to manage data holdings do not exceed the value of the data contributed. This extent will vary by theme.

2.2 Business Context

A goal for the framework is that it be widely useful. Framework data should have the following characteristics to attain this goal:

- Be free from use criteria or constraints. Licenses, copyrights, or other restraints will not allow the wide use needed for framework data to be effective and will lead to duplication of effort as those who cannot use the restricted data create their own. However, limitation parameters and suggested or optimal use of data need to be provided, and a disclaimer and liability structure should be firmly in place.
- Be available at the lowest cost possible, and no higher than a level sufficient to recover the cost of dissemination. The calculation of the cost to obtain framework data should not include costs associated with the original collection and processing of the information.
- Conform to approved standards and be available in public, nonproprietary formats.
- Be created and updated by organizations or partnerships that are knowledgeable about the data.
- Be certified to ensure that they meet the minimal standard for all framework criteria. A certification process of some form is essential; an independent assessment is needed to establish and maintain trust.

3 Role of Spatial Data Standards

Effective spatial data standards are necessary for any efforts in collaborative data production. Standards will increase the ability to share spatial data and preserve its original meaning, to create more complex applications, and to stimulate the commerce associated with GIS technology and spatial data.

Several areas require standardization. These include data models, data content, feature delineation, data collection, georeferencing, indirect positioning, data quality, metadata, and data transfer and exchange. Achieving agreement on standards in all of these areas is a significant undertaking.

The most effective standards are those that are widely used. How best to obtain this status, through de facto or de jure standards, is the subject of much discussion [3]. The standards process needs to balance the involvement of a broad community with the need to implement standards in an effective fashion.

4 Technology for Framework Implementation

Although it faces many organizational challenges, the framework will be realized only if significant technical barriers can be overcome. The technical problems associated with collaborative data production and use are beginning to be dealt with by the research community [4,5,6]. A design for the framework has been created that satisfies the goals of the framework and, although ambitious, appears to be technically viable.

The framework is based on a philosophy that considers a spatial data base to be, in itself, a multifaceted model of geographic reality. A conceptual data model based on the above philosophy organizes all user requirements for spatial data into a single framework, then specifies the components and relationships among those components.

The most fundamental aspect of the framework model, and the characteristic that distinguishes it from earlier geographic data models, is the existence of features. A feature is a description of a geographic phenomenon at or near the Earth's surface. Its digital representation, called a feature object, exists independently from any spatial elements (points, lines, areas, and nodes) to which it may ultimately be linked. Features, then, express nonlocational (attribute) rather than locational (coordinate) information.

Each occurrence of a feature, that is, a feature instance, is given a unique, permanent feature identification code. This feature identifier serves not only as a link to nonframework data bases of attributes, but also as the tracking mechanism for performing transactional updates.

The separation of feature objects from spatial elements ensures that the manipulations performed on the spatial data, such as vertical integration or coordinate transformations, do not affect the feature instances that have been defined in the data base. It also allows different features to reference the same spatial objects. This process alleviates needless replication of the feature information as the spatial configuration of points, lines, areas, and nodes is modified. Instead, adjustments are limited to the linkages between the features and the spatial elements to which they are tied.

When a feature is defined, it is further described by a set of attributes and relationships that reflect the properties of the real world entity that the feature object represents. Attributes define the

feature's characteristics, such as name and function. A feature may possess an unlimited number of attributes, or none at all, determined by the level of information needed to describe it adequately. Specific relationships are defined in the model to express interactions that occur between features. A substantial amount of research has been done on feature-based data models and implementations in object-oriented or extended relational data bases [7,8,9,10].

To meet the different needs of users, the feature-based design of the framework supports geospatial data at varying resolutions. Multiple resolutions of data (for example, data at different levels of generalization that have nominal positional accuracies of 50, 10, and 1 m) may exist at any given location. Where suitable higher resolution data exist, the lower resolution data will be generalized from the higher resolution data.[2] The data will be generalized according to a set of predefined rules for each theme. Alternate rules may be needed for a broad range of generalization [11].

As a general principle, the positions of contributed data will not be modified. For example, if a road crosses the boundary of two (otherwise equivalent) contributions, the lines at the common edge will not be geometrically joined. The disjoint lines that represent the location of the road will be associated through a common road feature, resulting in "logical seamlessness." Lower resolution data generalized from these data will be "geometrically seamless" (joined) if the alignment ambiguities present in higher resolution data sets can be resolved within the error tolerances of the lower resolution data sets. If resources permit, digital orthophoto data could be used as source material for updating and positioning features, particularly nonmatching features from multiple contributors. Features could be positioned at their proper geographic location and any new features appearing in the images would be added.

The process of revising (positional refinement) existing data, adding or deleting feature instances, and certifying the geographic fidelity and attribute accuracy of data provided by other parties is quite complex. To control this process requires the creation of a comprehensive transaction management system.

Consider the problem of maintaining different data bases in a consistent state. Multiple parties can be updating their holdings, probably without the knowledge of the others, and sending their transactions to the remaining partners. In a centralized, short transaction environment, locking is the strategy to prevent simultaneous and potentially conflicting changes to the data base. However, the multiple-party, long transaction environment seems to effectively preclude the use of locking or even a check in-check out strategy to maintain consistency.

Given this situation, a time-stamping solution could be introduced for each record type. The fields in each record type state the time the record becomes valid and stops being valid. The time stamps allow the transaction processor to leave the data base in a consistent state. Consider the case where a new road intersects an existing road segment. The existing road segment is split into two "new" segments. This process would generate the following sequence of transactions:

insert into road_fea_inst (id#) values ('107');	[the new road]
delete from road_fea_inst where id# = '16';	[the previous road]
insert into road_fea_inst (id#) values ('108');	[a road segment created by split of 16]
insert into road_fea_inst (id#) values ('109');	[a road segment created by split of 16]

[2]Decisions to store lower resolution data sets for later use, or to regenerate them "on demand," will be based on the state of technical means to generalize data and on business issues, such as costs and legal requirements.

What we want to capture is the information that road feature 16 has been replaced by 108 and 109, allowing the opportunity to transfer the attributes from 16 to the new road segments. The deletion of the road feature should not cascade to the associated feature attributes.

To achieve this requires several actions. First, the *delete* operator does not delete the record; the record is maintained, but its time stamp is set to "not valid." This is part of the strategy for version control in data base management systems such as Postgres [12]. Second, a field is added to the *insert* operation noting whether the new record replaces an existing one. Thus our transaction now looks like this:

```
insert into road_fea_inst (id#) values ('107');
delete from road_fea_inst where id# = '16';
insert into road_fea_inst (id#, replaces_id#) values ('108','16');
insert into road_fea_inst (id#, replaces_id#) values ('109','16');
```

If the *replaces_id#* field is not null, this could trigger the insertion of the attributes of the deleted feature into the attribute record of the replacement features. Although this example shows two features resulting from the splitting of one, the inverse case, the merger of two features into one, can also be described in this scheme. The time stamp fields allow the framework to accommodate the retention of past versions so that data are available for process studies.

Although the technical challenges in managing a distributed, heterogeneous data base environment are imposing, they are not insurmountable. Advances in spatial data modeling, data base technology, and network communications offer possible solutions. The problems, however, are so complex and solutions would be so useful, that an international consortium of mapping agencies may wish to pursue a common resolution to such questions.

5 Enabling Partnerships - The Policy Challenge

Building and maintaining the framework must involve a wide variety of parties in an ongoing, cooperative effort. Contributors must see clearly the benefits of doing more work on a data set than is required to meet their own needs. Users must understand how the framework would aid them.

Careful analysis of the benefits of the framework must be based on the recognition that organizations increasingly are both producers and users of data. Contributing data to the framework may require little more effort than an organization requires for its immediate needs, and the organization recoups this investment when it uses data from the framework, or data registered or linked to the framework, that others have provided.

The framework would help an organization to reach the following goals:

- Reduce expenditures for data collection and integration. Reducing redundant data collection and integration is cost effective and provides improved capabilities.

- Focus on its primary business ("back to basics"). As an organization sees that reliable information is or will be available, it can make more rational, less risky decisions to focus effort on what that organization does best. This argument becomes more telling as one considers the effort required to maintain data sets after they are acquired.

- Develop and operate critical applications for emergency response, natural resource management, and economic development more quickly and effectively because errors and uncertainty are reduced and the organization does not have sole responsibility for the entire required data set.

- Benefit more quickly and easily from data collected by others. Other organizations will use framework data as a base on which they register other themes of data or attach attribute information. Organizations whose data form the framework will find it easier to incorporate and take advantage of these other data.

In addition, the framework offers benefits to the entire community. Because it will allow better use of geospatial data, the framework will make the efforts of organizations beneficial to more than one community or set of customers.

6 Summary

Successful applications of GIS technology result from merging the capabilities of data providers and the requirements of data users. In the past, GIS users have created their own data bases, customized to their applications. Such data bases usually are of limited use to others, and they may duplicate data that exist elsewhere. The concept of a framework in which users share data holdings offers several advantages over user-created data bases: easier access to data, more complete analyses, lowered costs, and reduced duplication of effort. The data user is not just an information consumer, but also a provider of value-added information in return.

The technical problems of managing a confederation of different geographic data bases have been examined. Solutions that are based on the concepts of permanent feature identification, version control with time stamps, and a transaction processing system have been developed and seem feasible. The international mapping community may wish to pursue joint research on such issues.

The framework is a shared resource — defined, created, and maintained by an alliance of strategic partners. The partners, in their use of the framework, add attributes to features, update existing attribute data, refine the geographic accuracy of features, and provide notification of the creation or deletion of features to the other partners. These partners will form an advocacy group for the maintenance and growth of the framework.

References

[1] Federal Geographic Data Committee, 1995, Development of a National Digital Geospatial Data Framework, Washington D.C.: Federal Geographic Data Committee, 21 p.

[2] Office of the President, 1994, Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure, Executive Order 12906, *in* the *Federal Register*, Vol. 59, No. 71, pp. 17,691-17,674.

[3] Guptill, Stephen C., 1994, Spatial Data Standards and Information Policy, *Government Information Quarterly*, Vol. 11, No. 4, pp. 387-401.

[4] Frank, Andrew U., 1992, Acquiring a Digital Base Map: A Theoretical Investigation into a Form of Sharing Data, *URISA Journal*, Vol. 4, No. 1, pp. 10-23.

[5] Guptill, Stephen C., 1994, Synchronization of Discrete Geospatial Data Bases, *Proceedings of the 6th International Conference on Spatial Data Handling*, Edinburgh, Scotland, September 3-9, 1994, Vol. 2, pp. 945-956.

[6] Sheth, Amit P., and Larson, James A., 1990, Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, *ACM Computing Surveys*, Vol. 22, No. 3, pp. 183-236.

[7] Egenhofer, Max J., and Frank, Andrew U., 1992, Object-Oriented Modeling for GIS, *URISA Journal*, Vol. 4, No. 2, pp. 3-19.

[8] Guptill, Stephen C., ed., 1990, *An Enhanced Digital Line Graph Design*, U.S. Geological Survey Circular 1048, Reston, Va., 157 p.

[9] Guptill, Stephen C. and Stonebraker, Michael, 1992, The Sequoia 2000 Approach to Managing Large Spatial Object Data Bases, *Proceedings, 5th International Spatial Data Handling Symposium*, Charleston, S.C., August 3-7, 1992, Vol. 2, pp. 642-651.

[10] Worboys, M. F., 1994, Object-Oriented Approaches to Geo-referenced Information, *International Journal of Geographical Information Systems*, Vol. 8, No. 4., pp. 385-399.

[11] Beard, M. Kate, 1991, Theory of the Cartographic Line Revisited - Implications for Automated Generalization, *Cartographica*, Vol. 28, No. 4, pp. 32-58.

[12] Stonebraker, M., Rowe, L., and Hirohama, M., 1990, The Implementation of POSTGRES, *IEEE Transactions on Knowledge and Data Engineering*, March 1990.