IMPLEMENTATION OF THE NATIONAL MAP ROAD DATABASE WITH CONSIDERATIONS FOR ORGANIZATIONAL AND TECHNICAL INTEGRATION CONSTRAINTS

by

BRYAN D. WEAVER

(Under the Direction of Lynn Usery)

ABSTRACT

The National Map (TNM) concept allows for a more effective and regenerative public mapping program but not without overcoming some significant challenges. Numerous federal initiatives attempting to consolidate spatial information continue to proceed with poor coordination. This thesis assesses TNM road database implementation plan. The major organizational and technical constraints to spatial data integration are presented. Literature and sample data are used to develop indices that measure road data integration complexity, potentially assisting policy-makers in the development of relative cost models for various integration strategies. Results suggest an overwhelming need for a comprehensive requirements analysis for TNM transportation theme. The establishment of an overarching authority over federal, domestic mapping agencies is recommended. Integrated public road GIS data and systems is a daunting goal but one that remains in the best interest of the nation and should be pursued by the administration in the spirit of responsible governance.

INDEX WORDS: The National Map, Data integration, Institutional integration, Information sharing, Transportation, Federal mapping programs, Integration measures
IMPLEMENTATION OF THE NATIONAL MAP ROAD DATABASE WITH CONSIDERATION FOR ORGANIZATIONAL AND TECHNICAL INTEGRATION CONSTRAINTS

by

BRYAN D. WEAVER

B.S., Ohio University, 1996

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2004
IMPLEMENTATION OF THE NATIONAL MAP ROAD DATABASE WITH
CONSIDERATION FOR ORGANIZATIONAL AND TECHNICAL INTEGRATION
CONSTRAINTS

by

BRYAN D. WEAVER

Major Professor: Lynn Usery
Committee: Steve Holloway
Xiaobai Yao

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2004
DEDICATION

To all persons selflessly engaged in the pursuit of responsible government.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>viii</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
</tbody>
</table>

## CHAPTER

1. Introduction
   - Purpose .............................................................. 2
   - Goals ........................................................................ 4
   - Thesis format ......................................................... 5

2. Program Description and Literature Review ........................................... 7
   - Introduction ........................................................... 7
   - Toward a National Spatial Data Infrastructure ................................. 7
   - USGS and The National Map ......................................................... 9
   - Integration and interoperability ..................................................... 11
   - Complexity and information theory in cartography ............................ 14
   - USGS road transportation plans for The National Map ....................... 15
   - The benefits of a collaborative mapping program ............................... 17
   - The challenges to collaborative mapping ......................................... 20
   - Summary .......................................................................... 25

3. Methods of Policy Assessment .............................................................. 27
   - Introduction ....................................................................... 27
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4.1: Primary Road Dataset Characteristics of 5 Southeastern States</td>
<td>40</td>
</tr>
<tr>
<td>Table 4.2: State Area and Positional Accuracy</td>
<td>56</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>The National Map Road Information Sources</td>
<td>18</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Virtual Agencies and Ease of Implementation</td>
<td>23</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Neighboring Participant Heterogeneity Measure</td>
<td>37</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Conceptual Data Flow: Integrating The National Map Road Data</td>
<td>41</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Local to Global Schematic Federation</td>
<td>45</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>Horizontal Integration Issues: Network Representation</td>
<td>48</td>
</tr>
<tr>
<td>Figure 4.5</td>
<td>Horizontal Integration Issues: Varying Spatial Extents</td>
<td>50</td>
</tr>
<tr>
<td>Figure 4.6</td>
<td>Neighboring Participant Heterogeneity Example</td>
<td>53</td>
</tr>
<tr>
<td>Figure 4.7</td>
<td>Neighboring Participant Heterogeneity Example, Incorporating Border Length</td>
<td>55</td>
</tr>
<tr>
<td>Figure 5.1</td>
<td>County Integration Boundaries</td>
<td>64</td>
</tr>
<tr>
<td>Figure 5.2</td>
<td>State Integration Boundaries</td>
<td>64</td>
</tr>
<tr>
<td>Figure 5.3</td>
<td>Transportation and Orthophoto Overlay</td>
<td>66</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>The National Map Road Data: A Recommended Conceptual Data Flow</td>
<td>76</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The United States Geological Survey (USGS) Geography Discipline has been the primary federal provider for base, domestic topographic information. In the past, it has centrally managed geographic data collection, representation (cartography), and dissemination. This process model has proven insufficient for most modern mapping demands. It simply takes too long to map and integrate geographic data at such a large scale and extent with the current process and level of federal funding. While the USGS has labored with its conventional mapping enterprise, other federal, state, and local agencies have developed independent solutions to contemporary geographic information systems (GIS) demands. As a result, the current public spatial data environment consists of vast amounts of information maintained within various, mostly autonomous, local, state and national agency systems. This has pressed the USGS Cooperative Topographic Mapping (CTM) program to develop a new strategy that better meets the needs of the contemporary geospatial community.

Aiming for a more effective mapping program, the USGS Geography discipline has made public a new strategy for supplying base, domestic, geographic information to the nation. This strategic change has come in the form of The National Map (TNM). The National Map strategy stresses the changing role of the USGS CTM. The USGS will no longer be the primary data source for much of the Nation’s topographic map content. Rather, it will act as a domestic mapping coordination agency, focusing on the coordination of local mapping offices, standardization, data integration and data creation in the absence of any existing provider.
Essentially, TNM will be a distributed database of integrated base map information provided, predominantly, by local mapping agencies (USGS, 2001). Thematically, content is to include orthophoto imagery, elevation data, hydrography, transportation, structures, government and administrative boundaries, geographic names, and land cover information. Data in each theme must be sufficiently registered with all other themes so that topology is maintained at a required level of accuracy.

Although the program is in the early stages of development, it is not too early to explicate and reason through the major challenges that this unprecedented collaboration will face. Such a vast partnering program requires a complex coordination effort. This is no small feat. Participants at all levels of government will have to manage technical data and process integration issues, negotiate cost sharing models, and abide by some fundamental standards for data development and institutional transparency.

**Purpose**

This investigation aims to address part of a major research need for TNM. Specifically, this thesis takes strides to improve understanding in geographic information sharing practices and spatial data integration strategy. Kelmelis et. al. (2003) highlight four broad research needs for successful evolution of TNM. Stated first among these needs is the definition of key data requirements, integration constraints and resolutions:

“What content, relationships, forms, and structures of geographic data and information are needed to respond to the various problems facing the Nation?

How will TNM best be accessed and distributed? How will its contents be
maintained, integrated, and quality assured? How can TNM be kept secure yet remain readily and reliably accessible?"  

While this thesis primarily addresses the third question in the above quote, it also highlights the federal government’s need to define fundamental data content and organizational needs as a prerequisite to TNM integration implementation. Attention is focused on the transportation theme due to its central importance in TNM framework.

No other geographic theme contributes more to the facilitation of public services than the transportation theme. Few, if any, themes more greatly facilitate the addition of new information or data layers to GIS. Take, for example, the utility of address geocoding. Although the value of richly attributed transportation networks is great, it is somewhat balanced by the cost of data maintenance. This cost is greatly increased when multiple, politically independent partners contribute their respective geographies to an integrated transportation network. Because of the central importance of transportation data content to the overall success of TNM and the complex challenges presented by its integration, this thesis focuses on strategy development for the TNM road transportation data. There are, however, important parallels among the various themes of TNM regarding the technical and institutional challenges of integration. These differences are highlighted where appropriate. Likewise, the interdependency of transportation data with other data layers is considered carefully in the policy evaluation and recommendations of this work.

Currently, the USGS seems to lack practical recognition of the complexity of local/regional/national data integration projects. This is due, in part, to the lack of fully implemented, national level GIS integration projects from which to draw conclusions. More importantly, however, is the lack of a practical USGS strategy formulation for meeting demands of such an ambitious data integration effort. Furthermore, there currently exist somewhat
redundant federal efforts aimed at improving governmental efficiency of GIS road data maintenance under the guise of the federal e-government initiative. This is a direct, yet ironic, contradiction of the stated agency missions. Above all else, this thesis presents a reasoned assessment of TNM promise and the need for a more comprehensive national strategy for public spatial data partnering.

**Goals**

The primary goal is to provide policy makers with measures that will enable relative comparison of costs among various data and systems integration designs. The second goal is to provide policy makers with a reasoned plan for reducing redundancy in the quest for national mapping efficiency. In achieving these goals, three specific objectives, listed below, are met:

1. A suggestion of a strategy on how best to organize bureaucracies for interagency spatial data collaboration.

2. The development of a conceptual transportation integration complexity measures for TNM. Such measures highlight the relative degree of difficulty with each major phase of the road data integration process.

3. The presentation of a recommended, long-term approach to TNM road data implementation.

First, it is argued that the existing political level and management structure of TNM is insufficient for effective project implementation. The project management structure should reside with a political body that is able to authoritatively coordinate integration efforts at an interagency level. A second main argument is that the current partnership strategy of TNM is under bound regarding its participant strategy. By under bound, it is meant that there exists
excessive data development freedom at the data provider level. This results in a great deal of diversity in data content quality, ultimately burdening the integration process. Such a data-sharing environment is likely to be cost prohibitive in the long run.

**Thesis format**

The problems confronted by this thesis require the application of technical data integration knowledge, best practices of institutional data sharing, and organizational development principles. The need to understand the physical technical aspects of geospatial integration is obvious. However, these technical issues are exacerbated by the coordination effort required by the participating agencies. New data flows and coordinated, cross-agency processing affects and is affected by the structure and interaction among partnering organizations. Fountain (2001, p.96) accurately notes: “Organizational, network, and institutional arrangements – and the embeddedness of behavior in them – play key roles in technology enactment”.

Given the high-level audience and reasoned consideration for both technical and institutional issues, the overall framework of this thesis takes the form of an applied policy analysis paper. The National Map implementation presents an ill-structured policy problem, meaning that there is little empirical information available to guide policy formulation for such a program. This is particularly true given the lack of synthesis between technical spatial data integration research on one hand and organizational and policy development research on the other. Therefore, an integrated form of policy analysis is used in this work. Integrated policy analysis attempts to bridge overspecialized disciplines using both descriptive and normative means to evaluate, forecast and recommend policy. Various assumptions are made, and stated, as policy problems are structured herein. Simple indices, or measures, for the major components of
the data integration process are employed as a means for forecasting program complexity and effectiveness. These measures are used in the development of a long-term policy strategy for TNM road data integration.

The organization of this thesis, as stated, resembles that of a policy issue paper – where the problem is defined first, then alternative solutions are developed and compared, and, finally, a recommendation is presented. The definition of TNM, along with its problems and challenges, are stated initially in the second chapter. Again, TNM road data are the main focus of this work. Following a definition of the program, the methods employed with this policy analysis are detailed in Chapter 3. A conceptual TNM road network data flow is presented in Chapter 4. The data flow will serve as a process map, aiding subsequent discussion of road data integration complexity measures and policy alternatives. Chapter 5 then assesses various policy alternatives, regarding the transportation data contributor community. A long-term recommendation is made in Chapter 6 based on the analysis of available data and literature from the various disciplines mentioned above. Chapter 6 also provides a conclusion, highlighting the salient aspects of TNM road data integration process and the means by which its challenges may be met.
Chapter 2
Program Description and Literature Review

Introduction

This chapter presents the need for spatial data consolidation and details the TNM vision for contributing to this end. The means by which the USGS currently expects to fulfill TNM vision is put forth. Regarding transportation data, the U.S. Census Bureau is expected to be a major data provider and coordinator by way of a Master Address File / Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Enhancement Program that has recently commenced. Therefore, significant attention is paid to this U.S. Census Bureau data update program. Beyond the existing program implementation description, literature on both technical and institutional constraints to data sharing and inter-agency partnership formation is reviewed. Particular attention is afforded to literature covering barriers to spatial data sharing.

Toward a National Spatial Data Infrastructure

Throughout the 1990s to the present, the geographic information science industry has rapidly expanded its presence within public governance. While geographic information has long been used in public administration, a number of factors have enabled the increasingly rapid adoption of geospatial technologies within government entities. First, innovations in hardware, software, and communications technology have made geographic information systems (GIS) faster, more interoperable, and capable, with greater data storage capacity. Second, while technological advances have improved GIS operational efficiency, the cost of hardware, such as
workstations and networking components, has decreased. Third, advances in spatial data
collection techniques have improved the precision, accuracy and, in turn, the value of GIS.
Fourth, demand for spatial information and analysis has grown as recognition of its utility in
enhancing resource management efficiency is realized. Evidence of this recognition is not
difficult to find. Nedvodic-Budic and Pinto (1999) describe local adoption of GIS technology as
a “growth surge”, citing a study that states seventy percent of local governments use GIS. At the
federal level, the National Academy of Public Administration (NAPA) (1998) identified twelve
federal functions that require spatial data, ranging from national security to economic and
community development (USGS, 2001a).

Such a rapid surge in the adoption of GIS technology has not come without cost. Many
organizations employing GIS have developed and maintained spatial databases mostly
independent of other organizations using GIS. Immediate, short-term demand for spatial data
overshadowed the need for coordinated, industry-wide standardization. This is apparent across
all levels of government and industry. One example of this problem is the existence of multiple
transportation databases among government agencies - each designed to fulfill a particular
agency mission. The TIGER data support the U.S. Bureau of the Census mission; Digital Line
Graphs support the basic topographic mapping mission of the USGS; while the U.S. Bureau of
Transportation Statistics maintains the National Transportation Atlas Database. Further, most
states and many counties have transportation networks that are maintained autonomously. These
datasets cover many of the same geographic extents and real-world features, but have been
designed and maintained disparately within distinct fiscal budgets. The result is not only
redundancy. With budget constraints commonplace, it is very difficult for any one agency to
maintain a current spatial dataset for geographies of significant extent. This has been particularly
true of the USGS and Census transportation database activities. The USGS states openly that, on average, its base 7.5-minute topographic maps are 23 years old, as of 2001 (USGS, 2001a).

Along with many professionals within the geospatial community, the U. S. Office of Management and Budget recognized this operational redundancy in 1990 and released Circular No.A-16, Coordination of Geographic Information and Related Spatial Data Activities (OMB 2002). This 1990 circular, revised in 2002, called for the establishment of the Federal Geographic Data Committee (FGDC) with the purpose of coordinating various spatial data activities among federal agencies. In 1993, the National Performance Review further recognized the importance of spatial information and reported the need to establish a National Spatial Data Infrastructure (NSDI). The vision of NSDI is to create spatial data partnerships across government functions and political levels (Guptill, 1994). As stated in the USGS Implementation Plan for TNM (USGS, 2003), the specific NSDI goals are to 1) reduce duplication of effort among agencies, 2) improve quality and reduce costs as related to geographic data, 3) make geographic data more accessible to the public, 4) increase the benefits of using available data, and 5) establish key partnerships with states, counties, tribal units, academia and the private sector to increase data availability. The FGDC was placed in charge of developing the NSDI (Executive Order 12906).

**USGS and The National Map**

The National Map mission aims to serve the NSDI mission. The evolution of GIS and complementary technologies coupled with the recognized need for spatial data consolidation has been the impetus for a paradigm shift within the USGS, the U.S. Department of Interior (DOI) entity responsible for maintaining basic, domestic geospatial information. Leaders of the USGS
GeoDisc argue that they must evolve from a map producer to a parent organization that helps to define national spatial data standards while managing data partnerships. This vision is encompassed in TNM. It aims to provide the public with basic geospatial data across the entire nation through the linking of existing spatial databases at the federal, state, and local levels and, where feasible, privately licensed data. Within the context of this vision, the USGS (2001a) will claim responsibility for:

1. guaranteeing national data completeness,
2. marketing the availability and utility of TNM,
3. creating and stimulating partnerships,
4. integrating, certifying, and assuring quality of data from all participants,
5. owning and producing content for TNM where no other suitable and verifiable source exists,
6. leading the development and implementation of national geospatial data standards.

Beyond an initial compilation of data, the USGS will also work with its partner’s to monitor data and organizational relationships, and build analytical tools for knowledge creation.

The stated content goal of TNM is to provide nationally consistent and integrated base topographic map information useful for any arbitrarily defined geographic area. The USGS stresses that TNM content will need to be current, seamless and consistent with respect to data classification. Data precision, resolution and completeness will vary depending on the geographic area and demand. Relatively high-resolution elevation data, for example, will be available for areas of subtle relief. All content and metadata of TNM are to meet standards established by USGS Geography and in compliance with the FGDC.
The target level of cartographic detail to be achieved is equivalent to or greater than the detail of the USGS primary series topographic products, or 7.5-minute 1:24,000-scale topographic maps (USGS, 2003). Since content for TNM is to be provided from various organization types, including federal, state, local, tribal and private agencies, it must be integrated. More information on the vision of TNM is available by the USGS at the following website: http://www.nationalmap.usgs.gov/nmreports.html.

To date, the USGS is funding TNM development primarily via the Cooperative Topographic Mapping (CTM) budget. Cooperative Topographic Mapping’s mission is to ensure that USGS topographic maps are available and current by working with partners in other federal agencies; in state, county, and local governments; and in the private sector (USGS, 2003b). The exact amount being spent on this national mapping initiative is unclear, although the Presidential Budget Request for CTM FY 2004 is $74.1 million (USGS, 2003c). The Presidential Budget Request for the USGS is available online (http://www.usgs.gov/budget/2004/justification.html). The funding for TNM is being used to support a variety of current activities, including the development of technical infrastructure, the solicitation of local government participation, and project administration.

Integration and interoperability

There are several levels of systems integration that apply to inter-organizational spatial data sharing and consolidation programs. The following list presents the forms of systems integration offered by Deuker and Vrana (1995, 153). They are listed in increasing order of integration comprehensiveness. Successful maturity of TNM will allow for greater and greater levels of systems integration.
Taxonomy of systems integration:

1. *Data Integration*. The process of unifying existing data sources representing some or all of the same entities into a single framework of data (Devoge et al. 1998; ESRI 2001).

2. *Applications Integration*. Bundling separate applications into larger, more comprehensive applications.

3. *Functionality Integration*. Consolidating applications that serve different organizational functions with the purpose of enabling the support of a broader range of functions.

4. *Organizational Integration*. The integration of various units within an organization to achieve a common objective. This could mean the combining of functions, applications, personnel teams, and office space.

5. *Mission Integration*. The integration of organizations for the purpose of achieving cooperation towards a shared vision. This is principally a political and cultural level of integration.

The first phase of TNM development is to establish a process for data integration, the simplest level of systems integration. Unifying the existing data sources that will comprise a single framework is the process of data integration (Devoge et al. 1998). With spatial data, integration is both horizontal and vertical in nature. *Horizontal integration* is defined as the joining of spatially adjacent datasets with the purpose of generating expanded geographic coverage for a particular data theme. This term is intentionally broad, as project requirements determine the specific criteria for horizontal integration. For example, a data integration project may have geometric edge-matching of border road segments as the sole criteria for two datasets.
to be horizontally integrated. A more rigorous application may require that all participant data be interoperable, meaning road segments have unique object identifiers, road paths be consistently attributed, and have geometric edge-matching in order to be considered horizontally integrated.

*Vertical integration* refers to the overlay of data themes that originate from more than one source. Topological integrity across themes and data models is the goal of vertical integration procedures. As with horizontal integration, vertical integration success is measured by the error tolerance threshold defined in project standards and application requirements. Although there is no simple solution to either type of integration, horizontal integration may present the greater challenge to the transportation theme, due to the variety of datasets to be integrated. While this thesis alludes to vertical integration issues, its focus is on the horizontal integration of transportation data. For the program to be sustainable, data must also be interoperable to some degree (USGS, 2003; Kelmelis et al, 2003). By interoperable, it is meant that data must be usable by the same applications regardless of the data provider.

There exist three main strategies for accomplishing spatial data interoperability offered by Devogelle et al. (1998). The *first* approach is to integrate the data manually, specifying the data from participating databases that is to be merged with the global application. Global application processes receive the component data and synthesize the information to meet the global application requirements. A *second* approach to interoperability is through the application of standards. Standardization can be applied to both data models and schemas. Essentially, standards facilitate data exchange among systems. With a wealth of valuable information already existing among the participant community, however, the problem of converting present, unstandardized, data to a standard format remains a problem with this approach. The *third* strategy for achieving interoperability is to develop a software system that ties together existing
data model and schema designs. This is referred to as a federated database system (FDB) approach. It requires that all local schema differences be resolved via a global, virtual schema (Devogele et al. 1998, 336). This would be a daunting task for TNM considering the number of participants. A federated system may be developed with either a bottom-up approach – where local schemas are translated to a global schema, or a top-down approach – where local data providers are mandated to comply with a global schema, (Laurini, 1998).

**Complexity and information theory in cartography**

As described by Manson (2001), *mathematical complexity theory* and *information theory* contend that the complexity of a system is a function of the difficulty of describing the system. Further, he defines *aggregate complexity* to be concerned with how individual elements of a system work in concert to create systems with complex behavior (Manson, 2001). Complexity theory and information theory have been applied to cartographic research in various ways. Information theory has been used to describe the volume of information contained within a map (Tobler, 1997) and to determine more optimal map generalization routines (Battersby, 2002). To date, however, there have been no measures developed that attempt to estimate the complexity, or difficulty, of cartographic data integration. Applying the basic concepts of aggregate complexity, this thesis attempts to measure the complexity of TNM integration process through an examination of the component characteristics of TNM network. Both institutional and spatial data components are accounted for in these measures.
USGS road transportation plan for The National Map

Leveraging the Census Bureau’s Enhanced TIGER database

The USGS has identified the U.S. Census Bureau’s MAF/TIGER Enhancement Program as the long-term provider of road transportation and administrative boundary data (Broom et al, 2003). No formal cooperative agreement, such as a Memorandum of Understanding (MOU), has been signed by the two agencies but leaders of TNM program and the TIGER Enhancement project recognize overlapping objectives in their respective missions. The Census Bureau’s MAF/TIGER Enhancements Program has five strategic objectives (U.S. Census Bureau, 2002):

Objective 1 - Improve address/street location accuracy and implement automated change detection;

Objective 2 - Implement a modern processing environment;

Objective 3 - Expand and encourage geographic partnership options;

Objective 4 - Launch the Community Address Updating System (CAUS);

Objective 5 - Implement periodic evaluation activities/expand quality metrics.

As evidenced in the first objective, the primary goal of the MAF/TIGER Enhancement Project is to improve the positional and attribute accuracy of the TIGER database. This improvement will provide the Census Bureau personnel (as well as others depending on TIGER data) equipped with Global Positioning System (GPS) to precisely and confidently locate address information in the field. Census states, “the horizontal accuracy required for realignment of TIGER coordinates (including the required structures) is to be such that the geographic positional coordinates will correctly place an enumerator, relying on a mobile GPS-equipped computer:

- At the desired structure 100 percent of the time;
On the correct side of the street (i.e., in the correct census block) 100 percent of the time;

In the correct relationship to legal boundaries, other boundaries, and neighboring structures 100 percent of the time.”

Census statisticians have calculated that these requirements mandate an average road-centerline accuracy of ± 7.6 meters. This level of spatial accuracy is more than adequate for the current requirements of TNM (Broom, 2003). Through a contract with Harris Corporation, the Census Bureau expects to have completed this first objective for all 3,232 counties and statistically equivalent entities by the end of fiscal year 2008, in time for incorporation into the 2010 Census. Once the initial enhancement is complete, the information in MAF/TIGER is to be maintained to a currency of 1 year or less at all times (U.S. Census Bureau, 2000).

The Census Bureau cites that this massive TIGER update effort will be a cooperative effort, enlisting the aid of other federal, state, and local agencies. They will actively lobby for local and state mapping agencies to share current spatial datasets that may be used as update sources for the project. A June 2000 publication from the Census Bureau states: “The Census Bureau will coordinate the interaction among various government agencies to find and obtain such files; the Contractor [Harris Corporation] will be responsible for coordinating comparable interactions with commercial firms.”

Objectives two through five allude to a permanent change in TIGER maintenance policy. Not unlike the evolving USGS TNM policy, these objectives aim to develop a new culture of collaborative mapping. The future TIGER update environment will be based on automated network change detection via multi-source digital data, such as expert systems analysis of
imagery and local vector datasets. The parallel developments of the USGS and U.S. Census Bureaus’ mapping policies are highlighted in the Chapter 5 policy recommendation discussion.

Figure 2.1 generally illustrates the data source flow from the U.S. Census Bureau’s TIGER enhanced database to TNM. This diagram also highlights the overlap in data solicitation by both the USGS and Census on local mapping agencies. Throughout this enhancement process, the Census Bureau will attempt to develop and maintain informal (non-legal) partners at all levels of government.

Although an enhanced TIGER will offer adequate spatial accuracy and geometric precision, the U.S. Census Bureau will not populate its new database with all of the attributes currently required for TNM. This means USGS will have to source such information from, perhaps, the same partnership pool of agencies. The alternative is to centrally maintain this information within the USGS. The later alternative counters the stated philosophy of the USGS 21st Century mapping operations (USGS, 2001).

Considering the need to source additional, non-spatial data for TNM, the USGS will continue to rely on a bottom-up strategy for data integration that is most closely related to a Federated Database approach. USGS staff will translate local schema designs to TNM global schema via maintenance of a cross-reference table for each unique local schema.

The benefits of a collaborative mapping program

The arguments for inter-organizational sharing of spatial data are strong. Dueker and Vrana (1995) suggest three classifications of potential improvements from shared GIS initiatives.
Figure 2.1 – Multiple agencies at various levels of the political hierarchy will contribute data and coordinate relationships for TNM.
Although they focus on intra-organizational collaborations, the same potential benefits apply to inter-organizational initiatives such as TNM:

1. Efficiency – pool efforts and allow for data maintenance at a lower per-unit cost,
2. Effectiveness – new and higher quality products, services and analysis and decision making,
3. Enterprise benefits – improved communication and shared knowledge across the expanded community, providing opportunities for further integration and collaboration.

Conversely, the absence of sharing impedes technological advances and the social adoption of GIS (Pinto and Onsrud, 1995). The lack of data sharing forces organizations to expend more resources on data collection and maintenance as opposed to developing analytical applications and results. Further, applications for spatial data analysis, once developed, are not easily distributable within GIS communities having extensive system heterogeneity.

Reducing this heterogeneity is achieved through process and data standardization across agencies. Data standards, when developed in coordination with the stakeholder community, afford many benefits. According to Fountain (2001):

“Data standardization, catalyzed by the Internet, represents a significant rationalization of agency and interagency processes. First, standardization renders redundancy transparent. Second, standardization weakens the rationale for having different agencies collect and store highly similar or identical data elements. Third, data standardization suggests new forms of analysis that may lead to changes in the structure or organization of agencies. Fourth, structural changes in the federal bureaucracy are inevitable
as redundant data collection, storage, and analysis by different agencies is eliminated. The political battles revolve around which agencies will win and which will lose ownership of data.”

Her last point, regarding winners and losers in the political battle for process ownership, points to one of the many challenges of collaborative mapping.

**The challenges to collaborative mapping**

Integrating spatial data from a number of sources presents many challenges. Such challenges are both technical and institutional in nature. It is often difficult to distinguish the technical from the institutional challenges, as one challenge often interacts with or exacerbates one another challenge. That said, technical issues are associated with integration impediments resulting from spatial and non-spatial data characteristics. Institutional impediments refer to those integration problems associated with managing projects, such as coordination, cooperation and data ownership issues. Negotiating agency requirements and establishing workgroup accountability are more specifics examples of organizational, or institutional, challenges. The following discussion first focuses on the technical impediments to road data integration.

**Technical impediments**

Combining data from various, disparate sources present many potential data integrity problems. With non-spatial datasets, data definitions must be translated into a common, federated code, allowing for communication across all sources. Source discrepancies specific to geographic information present further complication to a distributed database design. The
typology of spatial database integration issues presented below is based on a list developed by Laurini (1998, 380):

1. Diversity in spatial representations,
2. Diversity in global projections,
3. Diversity in values for the same items located at different sites,
4. Diversity of spatio-temporal sampling,
5. Variability of definitions over time and space,
6. Discrepancies in coordinate values,
7. Discrepancies in boundary alignment (zonal fragmentation),
8. Variability in content quality,
9. Variability in data maintenance procedures,
10. Discrepancies in spatial metadata

Spatial data are not represented consistently among organizations with regard to schema, attribution, or geometry. For transportation networks, there may be significant overlap in geographic extent between neighboring partners with many of the same road segment features represented differently. Geometric inconsistencies across datasets also lead to topological accuracy issues when multiple data themes from multiple sources form a composite map for a particular geography. Developing processes for successful integration in the midst of such heterogeneity is a major challenge of TNM.
Institutional impediments

While it is commonly recognized that both technical and organizational impediments to spatial data sharing and integration exist, much research suggests that the organizational obstacles are more difficult to overcome (Croswell, 1991; Masser and Campbell 1995; Nedvodic-Budic and Pinto 2000). Masser and Campbell (1995) highlight some key obstacles to GIS sharing:

1. variation in participant priorities,
2. variation in GIS experience and technical ability,
3. differences in spatial data handling skills,
4. disagreement among participants regarding data openness, leadership, data standards, equipment and training.

Given the complexity of a national spatial data integration program, these issues are very difficult to surmount. Meredith (1995) has shown that the greater the number of participants in a data sharing program, the greater the organizational complexity. In addition, an inverse relationship between the interdependency of sharing organizations and the likelihood of project success has also been noted (Azad and Wiggins 1995, Fountain, 2001).

Collective action theory further highlights the multi-agency institutional dilemmas presented by TNM. In most cases, all relevant parties would be best served if they were to cooperate. Collective action theory argues that, in the absence of an overarching authority to enforce appropriate behavior and ensure commitment, individuals tend to avoid the risk of cooperating (Fountain, 2001). Unless individuals perceive the coordination cost to be less than the benefits derived from collective action, individuals will remain independent. Figure 2.2 is a graphic from Fountain (2001 illustrating the relationship between virtual agencies, operational
complexity and institutional barriers. Geospatial One-Stop (http://www.geo-one-stop.gov/), an intergovernmental spatial data portal developed to support of the President’s e-government initiative, can be considered an interagency website. The National Map is a cross-agency integration and systems development effort. This visual clearly depicts the development of a virtual, cross-agency implementation as the most difficult to accomplish.

**Figure 2.2** – The level of institutional change and the degree of operational change required for virtual agency implementations is positively related.
Complementing the issues highlighted by collective action theory are those issues brought to light by exchange theory. Currently, there is no legal and little political mandate for cooperation among agencies. In the absence of such legal or political mandates, long-term participation in an inter-organizational data-sharing program requires the perception of mutually reinforcing benefits (Azad and Wiggins, 1995; Cook, 1977). This argument is known as exchange theory, and it is likely the most popular reasoning for organizational cooperation within the public sector (Azad and Wiggins, 1995). The National Map is an initiative internal to USGS but its requested participant list reaches far beyond the USGS, even beyond the Department of Interior, for that matter. So how can the USGS, or the federal government more broadly, develop such mutually reinforcing benefits?

The institutional challenges facing TNM implementation are not solely attributed to inter-agency dilemmas. Resistance to change is inherent within institutional bureaucracies as well as between them. Longstanding institutions have developed and used GIS data for many years to fulfill their own missions and find it difficult to adapt to supporting a broader mission requiring organizational change, despite acknowledgement of such a need. Amidst a society swiftly discovering new ways to gather, store, analyze and share information, the institutionalization of key elements of organizational structure provides a means for resisting disruptive implementations of information technology (Fountain, 2001).

Further, Daft (1989) demonstrated that bureaucracies are more efficient than open and flexible. This may be due, in part, to the tendency of government to continue expansion rather than consolidation of capital. Fountain (2001) observed that government budget appropriations are more difficult to rescind than private budget appropriations once they have been used to establish programs. Within governmental entities, major changes must typically be initiated with
a political mandate before requisite attention and inter-agency cooperation is achieved (Craig, 1995).

Summary

The ubiquitous proliferation of GIS has given rise to “information silos” of redundant spatial data. Much of this redundancy exists among public agencies with distinct missions yet overlapping spatial demands. The USGS aims to help reduce this inefficiency by way of TNM. Through coordination of existing and distributed federal, state, local and tribal assets, TNM will provide the Nation with an integrated base map. The U.S. Census Bureau is in the process of enhancing the accuracy and maintenance procedures of its MAF/TIGER database. The spatially enhanced TIGER information will be the primary source for TNM road and boundary themes. However, given TNM current requirements, additional non-spatial attribute information will need to be sourced for local or state mapping agencies.

Most would agree that there exists a reasonably strong argument for developing a consolidated national mapping program (OMB, 2002; NAPA, 1999; Jensen et al. 1998). There is a great deal of public money to be saved and benefits to be realized in doing so. However, most also realize that there are many obstacles to overcome before TNM vision can come to fruition (NRC, 2003). These obstacles include technical integration issues - specifically, horizontal and vertical integration of spatial and non-spatial data attributes. Institutional obstacles, articulated in part articulated by collective action theory and exchange theory, present challenges to the long-term success of TNM.

This thesis recommends strategies for dealing with the major integration challenges facing the implementation of TNM base road data. Since the successful implementation of a
transportation theme is dependent, in part, on the broader program approach, many of these recommendations address basic program structuring. Chapter 3 will focus on the approach used to assess TNM plan and develop a recommendation.
Chapter 3
Methods of Policy Assessment

Introduction

In this assessment of TNM implementation, a general policy analysis framework has been employed. The National Map implementation presents an ill-structured policy problem, meaning that there is little empirical information available to guide policy formulation for such a program. Given the lack of synthesis between technical spatial data integration research on one hand and organizational and policy development research on the other, an integrated form of policy analysis is considered the most suitable approach. As mentioned in the introduction, integrated policy analysis attempts to bridge overspecialized disciplines using both descriptive and normative means to evaluate, forecast and recommend policy (Dunn, 2004). Constraint mapping and simple index creations are two, broad policy analysis strategies employed in this research. The former, constraint mapping, is used to identify and classify limitations and obstacles that hinder the achievement of program success (Dunn, 2004). The developed indices, or measures, for the major components of the data integration process are constructed as a means for communicating program complexity and effectiveness. These measures are used in the development of a long-term policy strategy for TNM road data integration.

The final project results reflect a reasoned assessment of the USGS description of TNM objectives contrasted with the difficulties presented by data integration and organizational development. Essentially, the methodology will consist of three main stages of research:
1. Literature review and project experience;
2. Qualitative data analysis of Georgia and Missouri data; and
3. The development of measures that express the degree of difficulty for specific integration alternatives.

These three stages are discussed below in greater detail.

A review of literature provides information on both TNM initiative and problems that have historically confronted similar, albeit smaller scale, partnership initiatives. By way of this review, the goals and current approach to TNM development is understood. Technical data integration literature is referenced and compared to the current expectation of TNM. A review of geographic information partnering incentives and constraints will be accomplished in this stage. More broadly, government agency partnering as well as e-government literature is cited.

Second, analysis of sample road data from various government entities provides the basis for identifying similarities and differences across potential partner datasets. Both the literature review and data analysis, collectively, provide the foundation for the development of a conceptual integration data flow model for TNM road data theme and complementary integration complexity measures. These measures constitute the third stage of research. The measures are used to logically compare the several road data implementation alternatives.

**Understanding TNM and information sharing**

To be effective in researching partnership programs, the researcher must play two, somewhat conflicting roles. First, the researcher must remain an objective evaluator. But the researcher should also witness the inter-workings of the program by attending meetings and holding informal discussions with the project staff (Ventura, 1995). I had the opportunity to
observe TNM project while also participating in operational level meetings and strategic
discussions regarding TNM implementation during the summer of 2003. During this time,
discussions with personnel of various functional responsibilities and at various grade levels (GS-7 to GS-14) provided a broad perspective of how TNM is being implemented within USGS. That experience has been helpful for the identification of potential integration issues. However, no specific conversations or observations are be used or cited in this research.

In order to develop a more complete understanding of TNM goals and project status, a review of published and unpublished documents dealing with TNM is critical. Independent evaluations regarding TNM implementation offer external perspectives on how the USGS may contribute to the NSDI (NRC, 2003; URISA, 2003; USGS, 2002). Beyond vision statements and program evaluations, it is also necessary to attempt to understand the project requirements of TNM. To complete this mission, several sources are referenced, including the Homeland Security Infrastructure Program (HSIP) publication (NIMA, 2003)\(^1\), various transportation standard documents, and peer reviewed publications regarding TNM.

As mentioned, various articles and books are referenced in relating TNM to similar spatial data collaboratives and e-government initiatives. An attempt is made to tie the theory learned from spatial data sharing projects and government partnership initiatives to TNM initiative. Literature of published geographic information sharing, partnering, standards and spatial data integration topics form the foundation of the implementation strategies presented in this thesis.

\(^1\) As of August 2003, this document was classified For Official Use Only.
Transportation data analysis

Beyond a literature review, state, local and commercial transportation data sets will be referenced. Local, state, federal and commercial data are visually referenced for the St. Louis, Missouri and Atlanta, Georgia geographies. Although these data provide a geographically limited sample of potential data provider road information for TNM, they represent potential contributions from the various political scales. These data are used to help identify potential data integration problems.

Specific procedures are used for identifying integration problems include both spatial and non-spatial dataset analysis. Regarding the spatial characteristics of the datasets, samples of both Georgia and Missouri datasets where evaluated against the HSIP 133 urban area color orthographic images (1 foot resolution) during the summer of 2003 at the USGS Rolla, Missouri facilities. These road networks were first evaluated based on their vertical alignment with the imagery. Second, judgment was made regarding the currentness of each dataset by comparing the imagery with the road networks; b) referencing the metadata of various sources and c) assessing the difference in topological content between each network. Datasets from neighboring political authorities were referenced and topological disconnects at the borders of the neighboring datasets were sought. Comparing the feature attributes of datasets to one another provided a means for the non-spatial attribute data analysis. Both conflicts and consistencies across the organizations’ datasets are referenced in the discussion of TNM conceptual data flow.

Conceptual integration process model and measures

Determining the relative costs of alternative integration approaches is a key component to TNM decision-making process. To assist in such a comparative evaluation of cost, a conceptual
integration data flow diagram is developed based on an assessment of the technical and organizational constraints. This model aids in the explanation of the data integration process and highlights the most complex aspects of the process.

Through practical consideration of TNM integration data flow, along with the cited literature and data analysis, key integration complexity measures are developed. The issues affecting the cost of integration will be reflected in a series of simple indices, or measures. An attempt was made to combine a few of these measures into an aggregate index. However, a lack of both empirical data and historical research on large-scale spatial data integration severely limits the accuracy and meaning of such an index. Various implementation alternatives are substantively evaluated by changing the implementation assumptions on the input of each measure.

Some indices are derived from the formulation of theoretical interaction matrices. These matrices represent the presence, absence or degree of interoperability that may be required between adjacent data providers, given various assumptions. This may provide a means for forecasting the maintenance complexity of TNM. The results of the each measure provide a means for understanding the relative difference in integration complexity and long-term viability for various partnership strategies and data requirements.

**Summary**

Overall, TNM integration presents an ill-defined problem for policy makers and policy researchers. There is a lack of public sector experience regarding empirical research on both technical and institutional spatial data integration programs. Therefore this research attempts to reason through TNM implementation strategy by mapping its organizational and institutional
constraints and, subsequently, developing simple indices which assist in communicating the complexities of such constraints. Due to 1) a deficiency of existing empirical research on spatial data integration, 2) limited access to datasets and, 3) the scope of this project; an evaluation of existing transportation datasets was limited and mostly qualitative. Indeed, most of the salient aspects of this thesis are deductive in nature.
Chapter 4

Defining The National Map Road Data Integration Process

Introduction

The first aspect of this chapter is to define the derived measures for integration strategy complexity. The complexity measures are simple indices that are fairly easy to comprehend. Each measure presents policy-makers with a theoretical approach to understanding the relative difficulty of alternative integration strategies. This leads to the second aspect of the chapter: presenting a conceptual transportation data integration process for TNM from the perspective of its major complicating forces. To achieve this, a conceptual data flow diagram is presented and explained. The explanation of each step, or phase, also points to one or more of the measures of integration complexity presented in the first section of this chapter. In Chapter 5, various implementation strategies will be evaluated with the aid of the proposed measures. A final implementation recommendation informed by this discussion is then provided.

Measures of integration complexity

With the literature, data review, and some applied reasoning, seven integration complexity measures have been developed. These measures can be used as a means for comparing complexity among different partnership strategy alternatives and requirements. Empirical testing of most of the measures discussed below requires substantially more data than this project had the resources to gather. Also, the, thus far, ill-defined nature of TNM product would require additional assumptions to be made in deciding which road network attributes are
most important regarding seamlessness and consistency. Discussion following the description of these measures will make assumptions on the relative index value for various alternative integration approaches. These assumptions are accounted for in the recommendation presented in Chapter 5.

1. **Number of participants**

   *Measure Definition:* The total number of participants contributing transportation data for which horizontal data integration is required. Here, a participant is defined as any entity that is a direct provider of transportation data to TNM. A participant may be a county government entity or an acting local or state partnership office in lieu of each individual county comprising the partnership.

   *Importance:* Base on collective action theory (Fountain, 2001) and organizational collaboration research specific to spatial data Meredith (1995), the greater the number of participants in a data sharing program, the greater the organizational complexity.

2. **Number of participant Schemas**

   *Measure Definition:* The total number of participant *schema types* contributing transportation data for which horizontal data integration is required. A *schema type* is a database characteristic, or set of characteristics, deemed to be important and distinguishable. For example, participant schemas exhibiting the same road segment attributes, data accuracy characteristics, and database maintenance procedures may all be attributed by the same *schema type* value. How *schema type* values are determined is dependent on TNM data consistency requirements and integration capabilities. Unless joint database development within the participant community exists, the number of schemas will, most likely, be equal to the number of participants.
Importance: The greater diversity in schematic designs within the partnership community, the greater degree of difficulty in federating the community schemas (Laurini, 1998).

3. Global schema requirements

Measure Definition: There are two definitions, or types, of global schema requirements for which measure may be of use.

1. The total number of attributes required for each federated road segment in the global schema.

2. The accuracy or precision requirements of the federated dataset. This may include the definitional precision of road attributes, such as road class, or the positional accuracy requirements of the road centerlines.

Importance: Greater attribute detail in the global schema requirements leads to greater costs and a lower likelihood of successful data integration (Laurini, 1998). This measure is most useful when evaluated with the local-to-global compatibility measure (measure 7).

4. Aggregate participant boundary length

Measure Definition: The total length of coincident participant boundaries. This measure is a summation of the linear distances of the participant boundary segments.

Importance: The length of the aggregate participant boundary directly and positively relates to the extent of required horizontal integration. In addition, an inverse relationship between the interdependency of sharing organizations and the likelihood of project success has also been noted (Azad and Wiggins 1995, Fountain, 2001).

5. Neighboring participant heterogeneity

Measure Definition: The spatial configuration heterogeneity of schema types among participants. For any particular required database characteristic, or set of characteristics,
neighboring participant data heterogeneity is calculated via an interaction matrix summation. Figure 4.1 illustrates a simple example of this concept.

(1) Each participant is attributed by schema type. The schema type attribute represents a database characteristic, or set of characteristics, thought to significantly impact the efficacy of the integration process. For example, key road segment attributes used for automated integration processing (e.g. route name), data accuracy characteristics, and database maintenance procedures may all constitute one schema type attribute or, alternatively, distinct schema type attributes. The definition of each schema type attribute is dependent on TNM data requirements and the integration process capabilities. A schema type attribute may be a nominal, ordinal, or continuous measure. This example demonstrates how a nominal measure of schema type may be used to calculate the neighboring participant heterogeneity measure.

(2) An interaction matrix is developed from all neighboring participants. The matrix grid values for each participant pair are determined by a) spatial adjacency and b) a comparison between participant pair schema types. If two, geographically adjacent (share border) TNM participants are attributed with the same schema type value, then their participant pair grid value within the matrix would be zero. If adjacent participant schema type values differ, then their matrix grid value would equate to 1, assuming a nominal schema type variable. An ordinal or continuously defined schema type variable may be appropriate for some participant characteristics. A sum of the matrix rows followed by a sum of the row summation column would yield a participant heterogeneity measure. Schema type attribution may be sourced from individual participant metadata.
Neighboring Participant Heterogeneity Measure

Participants
Represented in Geographic Space

Participants
Database Representation

<table>
<thead>
<tr>
<th>Participant-ID</th>
<th>Schema Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
</tr>
<tr>
<td>7</td>
<td>C</td>
</tr>
</tbody>
</table>

Participants
Spatial Interaction Matrix

<table>
<thead>
<tr>
<th>Part.-ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Row Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

Neighboring Participant Heterogeneity = 7

Figure 4.1
This Diagram shows three different representations of a theoretical participant network attributed by local dataset characteristic class; in a) geographic space, b) database table, and in c) a spatial adjacency matrix. In the spatial interaction matrix, 0 represents no difference in schema type while a value of 1 represents a difference.
(3) Weighting each matrix cell by the length of the coincident border shared by the two participant entities would yield the length of participant border with conflict regarding the database characteristic of interest.

*Importance:* The greater the heterogeneity and poorer the aggregate data quality among bordering participant networks, the greater the difficulty in reconciling differences during the horizontal integration process. With empirical data highlighting the dataset characteristics that contribute to integration cost and matrix cell weighting based on the participant border length, the neighboring participant heterogeneity measure can be used to estimate project edgematching cost. An example of this technique is provided later in this chapter.

There are two measures of data quality presented that indicate, to some degree, the level of data integration difficulty. These measures may refer to both spatial and non-spatial participant database characteristics.

6. **Geometric integration complexity (GIC)**

*Measure Definition:* This measure estimates the relative complexity of topological integration ascribed to the spatial and non-spatial accuracy of the participant networks. This measure is an area-weighted summation of the accuracy of all data providers. Formally,

\[
\text{GIC} = \sum_{i=1}^{n} A_i \times \mu_i
\]

Where GIC is the geometric integration complexity measure, \(A_i\) is the percent of total area data provider \(i\) contributes to the integration project, and \(\mu_i\) is the mean dataset accuracy value for data provider \(i\). With this measure, a mean accuracy value for the provider population is provided while accounting for the differences in geographic size of each provider. High GIC values translates to a low aggregate data accuracy.
Importance: The greater the mean spatial and non-spatial accuracy of participant data, the more difficult the integration procedure will be.

7. Local-to-Global Compatibility

Measure Definition: The sum of all required global schema characteristics not met by the local schema. This may be an *a priori* or *post prior* measure of integration complexity. Formally:

\[
\text{Total Compatibility} = \sum (F_R - P_i)
\]

Where \(F_R\) is the number of required federated attributes and \(P_i\) is the number of federated attributes correctly translated from a participant to the global schema. For example, the global schema might require all *Class 1*-road segments (as defined by the global schema) to be distinguishable from non-*Class 1*-road segments within participant schemas 100% of the time. If a *Class 1* road class cannot be correctly derived from a participant’s dataset 100% of the time, then this federated attribute would increase the \(P_i\) value in the above equation by one. A representative sample of participant schemas may provide enough information to estimate the average number of mandatory attributes correctly mapped to the global schema per partner.

Importance: Greater complexity and stringentness within the global schema requirements leads to greater sourcing costs and a lower likelihood of success.

Table 4.1 presents dataset characteristics for five southeastern states. Real-world examples of the measures presented in the next section of this chapter will call on the information contained in this table. The assumption for each example is that these five states desire an integrated transportation network comparable in content and quality to the proposed TNM road network. Additional assumptions and information are presented to illustrate specific measures. All information was gathered from the metadata provided by the state government agencies responsible for their respective road GIS databases.
Table 4.1 – Primary Road Dataset Characteristics of 5 Southeastern States. Y=Yes, characteristic is present and N=No, not a characteristic of the participant dataset.

<table>
<thead>
<tr>
<th>Dataset Characteristics</th>
<th>Alabama</th>
<th>Georgia</th>
<th>Florida</th>
<th>North Carolina</th>
<th>South Carolina</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network vector (arc-node) model</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Persistent Unique identifiers</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Primary route number or road name</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Road jurisdiction</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Record modify date</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Address range</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Data Content Quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All public roads within extent present</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Positional accuracy (meters)</td>
<td>5</td>
<td>10.1</td>
<td>12.2</td>
<td>12.2</td>
<td>12.2</td>
</tr>
<tr>
<td>Update frequency 1/year or greater</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

The following section attempts to define a conceptual integration data flow for TNM transportation. A discussion of the constraining forces affecting this process, both organizational and technical, is aided by the theoretical inclusion of the measures defined above and samples from road dataset analysis. Various political levels of implementation are evaluated for each phase of the integration process following a general description of the data flow diagram. In addition, examples of how the more abstract measures can be calculated are provided at the end of each phase description.

**Conceptual integration data flow**

The data flow proceeds in a general direction of left to right (figure 4.2). Content is virtually acquired from providers represented at the left, including the U.S. Census Bureau and local government agencies. This information is evaluated and the local road attribute schemas are mapped to a federated, TNM road attribute schema. Once the required data are appropriately classified, they are horizontally integrated, or edge-matched. This integration effort is affected by
Figure 4.2 – Road data integration workload is affected by the quantity and quality of participant data as well as the requirements of the TNM federated road dataset.
both data quantity and data quality. Once the horizontal integration is complete, a preliminary U.S. road network is produced. Data conflicts and inconsistencies that do not meet the required level of data quality are then addressed in a second integration iteration. Once all intolerable issues are resolved, TNM road network version is finalized. The National Map network is then used as a base for the integration of subsequent network generations. It is important to bear in mind the continuous update nature of TNM information. This diagram may represent the process for integrating simply one transportation data partner to an existing federated road base. Conversely, it may represent a single mass-integration of all transportation data partners. Further, components of the topological integration effort, represented by the center of the diagram, may be completed in parallel with the federated translation phase rather than in strict, chronological succession. Each major aspect of the data integration process is evaluated below.

Assessing each phase

1. Data Acquisition – Identify and acquiring “Best Available Data” assets

As TNM vision statement (USGS, 2001) calls for, the USGS will find and acquire the best available data source for all areas of the U.S. With the USGS – U.S. Census Bureau agreement, much of the data collection effort will be under the control of the Census and its contractor, Harris Corporation. However, Broom and Godwin (2003, p.1122), both with the U.S. Census Bureau, note that “the USGS’ list of desired features was more extensive than that needed by the Census Bureau.” They further noted that the Census Bureau “would provide a location within its data structure to store any features that the state/local/tribal GIS files that could be of possible interest to TNM and other groups…” Further, the Census Bureau has
reviewed its Census Feature Classification Code (CFCC) so that it will collect and store basic attributes and feature-level metadata (Broom and Godwin, 2003).

This means that the Census Bureau will, most likely, identify and acquire a bulk of the transportation and boundary information required for TNM. Additional sources, beyond that provided by the Census Bureau, may need to be collected by the USGS to fulfill TNM requirements. Both the global schema complexity and the participant attribute quality influence the difficulty of the TNM sourcing effort. Generally, the greater the minimal road segment attribute requirements, the more difficult it will be to acquire the necessary sources.

2. Federation - Local-to-Global attribution translation

As discussed, TIGER transportation data, and therefore TNM, currently relies on a bottom-up strategy for data integration that is most closely related to a FDB approach. From a data attribution perspective, USGS staff will translate local schema designs to TNM global schema via maintenance of a cross-reference table or a technical equivalent. The benefits of this strategy include limited disturbance of local participants’ existing processes. Any standards that are applied with this approach are applied at the overarching level. This translates into less of a need for USGS to provide incentives for local adherence to standards, monetary or otherwise.

There are many disadvantages to this integration approach, however. A bottom-up approach is much more complex than a top-down approach due to the heterogeneity of existing spatial databases (Laurini, 1998). Responsibility for tying together partner data is left to the overarching authority, along with the cost for doing so. Of primary concern is that this process does not address the problems of inconsistency among the various jurisdictions. An enormous amount of work is left up to the organizing entity, i.e. the USGS with some major issues left
unresolved. Figure 4.2 illustrates the basic philosophy of this mapping service structure and some of its limitations. Figure 4.3 illustrates some classic complicating factors with the bottom-up, federated approach. The first step in the process is to establish a partnership with a state or local authority. Second, each stream of information from the participants WMS or database needs to be interpreted and cross-referenced to TNM schema before it can be rendered by TNM display engine. A complete integration of schemas, where local schema discrepancies are resolved, may be extremely costly, dependant on TNM requirements and the volume of participants. The cross-reference table above shows how the local semantics might only partly translate to the global semantics. There will rarely be a conflict-free match of feature classes between the local and global data structures. As an example, Missouri Department of Transportation (MoDOT) road segment objects with a ‘BU’ or ‘CO’ Road Class attribute value require additional attribution in order to determine which TNM Road Class value is appropriate.

In the case of Missouri county roads (‘CO’), the definitional difference between TNM ‘Class 3’ and ‘Class 4’ road segments must be discernable within the Missouri dataset. This problem is made less severe since MoDOT also uses the U.S. Census Census Feature Class Code (CCFC) attribution, allowing for most county road segments to be correctly translated to either a Class 3 or Class 4 within the global schema. Unfortunately, there exist 23,560 road segments that lack the CCFC attribution and 1,370 lack both CCFC and MoDOT Road Class attribution within the MoDOT dataset. Those 1,370 road segments are guaranteed not to automatically translate directly to TNM schema. Many participant datasets do not contain CCFC, and the appropriate Road Class translation may be much more difficult.
Further complicating the federation process is the need to understand the capture conditions and procedures with which local participants maintain attribute information. The dynamic nature of organizational procedures requires the federating agency to know when critical capture conditions and procedures are modified by data providers. With the number of partners envisioned coupled with the heterogeneity of mapping schemas and systems, maintaining thousands of partnerships will be awfully burdensome.

Attempting to determine the complexity of this aspect of the integration process requires consideration of the **Number of participants**, and the **Number of participant schemas** measures. It is important to remember that a participant schema may be defined to include the
capture conditions of critical attributes as well as the definition of attributes. Even if participant schemas translate to a global schema without definitional conflict, the quality of dataset attribution affects the cost of translation. With some local dataset data mining, the **Local-to-Global compatibility** measure may also provide valuable insight into the likelihood of success for a federation strategy.

**Example measure values**

Assume that the five states listed in Table 4.1 are to integrate their respective transportation datasets. The number of participants is clearly equal to 5, assuming that only the primary road datasets at the state level are to be integrated. The number of participant schemas is equal to 5 as well, considering that no significant interstate data development program currently exists between any of the participants. If major metropolitan or county governments were to become data contributing participants, both the number of participants and, most likely, the number of participant schemas would increase.

Summing all instances in which a global schema characteristic is not met by a participant schema derives the local-to-global compatibility measure. To produce this measure, the required global schema characteristics must be defined and pertinent local schema characteristics must be available. For this example, it is assumed that all characteristics listed in Table 4.1 are the complete requirements for the global schema. The positional accuracy requirement is not included in this measure when it is applied to the road segment attribution federation. This example, then, results in a local-to-global compatibility equal to 11. Alabama, Georgia, Florida, North Carolina, and South Carolina do not meet four, one, two, four, and zero of the required global schema characteristics respectively. That sums to 11. This measure is correlated with the
number of participant schemas but is sensitive to both the specificity of the global requirements and the contributing content quality relative to the requirements.

3. Integration – edge-matching of partner data

The current strategy for dealing with horizontal integration of transportation is to develop automated and semi-automated techniques for linear feature edge-matching. Although the initial integration function is to be completed by the TIGER Enhancement Program, attempting to understand the complexity of this process is important for managing expectations and maintenance protocols. A few dataset examples from potential partners are provided to highlight the difficulties of this process. In the process, several of the measures defined above will be used to assess various implementation alternatives.

Technical Constraints

Horizontal integration is complicated by variance of topological representation and accuracy of the participant community. Topological representation refers to the way in which features are stored and depicted in geographic space. It also refers to database standards and maintenance procedures. Figures 4.4 and 4.5 diagram these major spatial database characteristics that complicate horizontal integration efforts for TNM.

Figure 4.4 shows a map of the border area between Cherokee County, GA on the north and Cobb County, GA on the south. The county line is displayed horizontally across the middle of the map. This map shows a discrepancy in road entity representation between the two counties.
Figure 4.4 – Differences in the way agencies represent road entities present complications to the cartographic federation process.

In the county border area around Site A, the limited access road shows double-digitized representation within the Cobb County database contrasted with the single road segment
representation within the Cherokee County data. Such differences need to be resolved in the integration effort from both a standards perspective (should the double digitized entity remain so or be reduced to a single-line representation) and from an edge-matching perspective (the road segment objects within the respective database management systems must be paired appropriately).

Figure 4.5 shows a map of the border between Bartow County on the north and Cobb County on the south. The county border is displayed immediately below the scale bar. Sites A, B, and C demonstrate three additional edge-matching scenarios that occur between neighboring datasets. First, **Site A** shows a significant overlap between the Cobb County data and the Bartow County data. Specifically, Cobb County seems maintain its road database with a one kilometer border into neighboring counties. And according to the metadata of the two datasets, the Cobb County data are more recent and spatially accurate. In this situation, which representation should be captured for TNM or TIGER? The converse of the overshoot situation is the undershoot scenario. With an undershoot, a road segment does not extend to the dataset boundary when it should. Second, **Site B** shows no such overshoot; this is, perhaps, the most ideal edge-matching situation from a topological perspective. **Site C** is another example of the difference in topological representation; with a double versus single line representation for a limited access highway. This site also highlights the reasoning for the data overlap; there exist important network entities that lie near political borders. It may be useful for a neighboring county (in this case, Cobb Co.) to maintain important network entities that fall within the neighboring county. This might allow Cobb county personnel to route to important network points efficiently. If an important interstate access ramp exists just across the Bartow/Cobb County boundary, should not
Figure 4.5 – The extent of neighboring datasets need to be accounted for in the federated map display.
the Cobb County E-911 system store it? To complicate matters further, often, a mismatch between two road segments is valid.

Both the spatial and attribute accuracy of the participant networks may present challenges to data integration as well. As the spatial accuracy of participant datasets decreases, the success of automated edge-matching routines decreases. Essentially, poor accuracy datasets offer a greater risk of incorrect road segments being edge-matched. Significant line geometry changes may also be required to “snap” the corresponding road segments at the participant border. Likewise, poor attribute accuracy inhibits automated integration routines, where entity matching is highly reliant on non-spatial attribute values.

Considering these technical constraints to cross-agency integration, the following integration complexity measures may be useful in evaluating strategy. The list of these measures is followed by an example of how each is calculated.

1. **Aggregate participant boundary length** – the longer the inter-participant border, the greater the integration work.

2. **Neighboring participant heterogeneity** – The spatial configuration of certain, critical participant data characteristics contribute to horizontal integration complexity. The participant boundary length measure may be combined with this measure to yield the length of coincident boundary presenting significant integration challenges.
3. Geometric Integration Complexity – The spatial and non-spatial data accuracy of participant data directly contributes to the difficulty of integration.

Example measure values

1. The aggregate participant boundary length measure is demonstrated in chapter 5. If a state level horizontal integration of road networks were to be completed for Alabama, Georgia, Florida, and South Carolina state datasets, then geometric discrepancies among the datasets would need to be resolved along the 1,700 kilometers of adjacent state border length.

2. The neighboring participant heterogeneity measure may be used to quantify the significant differences between neighboring datasets. Using the information provided in Table 4.1, figure 4.6 demonstrates how neighboring participant heterogeneity may be calculated. In this, more complicated, example of the neighboring participant heterogeneity measure, an edgematch rating is determined based on the comprehensiveness of available data (i.e. whether or not all public roads are present) and the positional accuracy of data. The former dataset characteristic is considered most important since the absence of some road classes requires additional data sourcing. The edgematch rating is an ordinal variable measure derived via a logical query of the metadata provided by each state. Neighboring state values are summed to derive the cell value within an interaction matrix. The datasets presenting the least challenge to edge-matching will have the lowest edgematch rating. It is assumed that more comprehensive and positionally accurate data are the easiest to integrate. By varying a hypothetical, critical positional accuracy value, two neighboring participant heterogeneity measures are compared. In the first case, it is assumed that an automated edge-matching procedure performs sufficiently on datasets with a 15-meter or better positional accuracy. The right hand example assumes a 7.6-meter accuracy requirement for sufficient automated edge-matching performance.
Figure 4.6 – An example of how the neighboring participant heterogeneity measure is derived.
Ideally, empirical integration research from pilot studies would determine edgematch ratings that are related to the real workload particular dataset characteristics impose on the integration process. For example, if research shows that for every 5 meters loss in positional accuracy there is a 10% increase in integration time, then the edge-matching rating could be derived for each participant starting from a base value and increasing by 10% for every 5 meters loss in accuracy. A unit, perhaps hours per kilometer of integration, could then be assigned to the output of the participant heterogeneity measure. The inclusion of such empirical integration research would greatly increase the interpretability of this measure.

With edge-matching, it is important to weigh the edgematch value between neighboring entities by the spatial extent of their shared border. Therefore, the cell value within the neighboring participant heterogeneity matrix should be taken one step further than what is presented above. Figure 4.7 multiplies the cell values of the matrices in figure 4.7 by the shared border distance of each state pair. Assuming, for the purposes of this example, that the cell values in the interaction matrix of figure 4.6 are defined as minutes per kilometer of integration, the resulting neighboring participant heterogeneity value in figure 4.8 is then interpreted as the estimated number of hours required to edge-match the road datasets for these five states. Again, empirical research identifying the dataset characteristics that significantly contribute to the cost of integration is required before integration cost can be estimated from this measure.

3. The geometric integration complexity measure simply multiplies a participant’s mean positional accuracy by the percentage of land area it contributes to the project. Then, all area weighted accuracy values are summed to provide the GIC measure value. This concept is applied to the hypothetically integration project of the southeastern states example.
Figure 4.7 – Incorporating border length with the derived edgematch rating yields a truer, more meaningful neighboring participant heterogeneity measure.
Table 4.2 - State Area and Positional Accuracy

<table>
<thead>
<tr>
<th>State</th>
<th>Area (sq. km)</th>
<th>% Total Area</th>
<th>Positional Accuracy of dataset (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>135,775</td>
<td>20%</td>
<td>5 (1)</td>
</tr>
<tr>
<td>Georgia</td>
<td>153,952</td>
<td>22%</td>
<td>10.1 (2.22)</td>
</tr>
<tr>
<td>Florida</td>
<td>170,313</td>
<td>25%</td>
<td>12.2 (3.05)</td>
</tr>
<tr>
<td>North Carolina</td>
<td>139,396</td>
<td>20%</td>
<td>12.2 (2.44)</td>
</tr>
<tr>
<td>South Carolina</td>
<td>89,898</td>
<td>13%</td>
<td>12.2 (1.59)</td>
</tr>
<tr>
<td>Total</td>
<td>689,334</td>
<td>100%</td>
<td>GIC = 10.3 meters</td>
</tr>
</tbody>
</table>

Table 4.2 above lists the area, % of total area, and positional accuracy of each data provider. The area-weighted positional accuracy of each dataset is provided in parentheses within the far right column. The GIC value of 10.3 meters is the summation of the numbers in parentheses. This number can be compared with other combinations of datasets and with any critical values associated with positional accuracy. If, for example, a critical positional accuracy value for successful vertical integrations is 7.6 meters (the TIGER/MAF positional accuracy requirement), then this dataset population is not sufficiently accurate. The positional accuracy statistic can also be evaluated on an individual dataset basis. With the same 7.6 meter requirement, it is observed that 80% of the integration area would fail to meet this requirement given this provider population.

In addition to local variation in physical network representation, participant reliability issues add to the difficulty of integration. Both data and metadata must be continuously accessible to the integrating organization(s). Beyond being continuously available, dataset metadata (and optimally object-level metadata) must remain temporally accurate and sufficiently detailed, as it will be a key input to automated integration procedures.

When discrepancies between datasets are present, an acceptable representation must be derived for TNM representation. This will require communication, perhaps negotiation, with the local data providers. There is also the chance that subsequent modification of a local road
network object might have legal or political implications. Resistance to updating from the local data provider may result from such pressures.

These coordination efforts become more complex as the number of participants and facilitating stakeholders increase. Therefore, the **Number of participants** measure is an important relative indicator of integration complexity regarding institutional characteristics.

4. Maintenance – updating and change management

The procedures for maintaining an edge-matched database that meets the demands of TNM have yet to be developed. What is clear, however, is that updating both TIGER and TNM must be performed at the feature level. Through database transactioning procedures, only those roads segments with geometric or relevant attribution modifications will be flagged for potential update. New, or added, road network features will also be incorporated into the federated database, again, with the assistance of database transactioning procedures. The U.S. Census Bureau and its partners will also be employing image analysis techniques to more rapidly identify transportation land-use development. To summarize, maintaining a sufficiently accurate and integrated U.S. network, once created, will target only significant changes. The integration costs are heavily front-loaded. But the single most important measure in understanding the relative difficulty in managing data exchange and community requirements is the **Number of participants**. The greater the number of participants within this community, the more costly the coordination effort will be. Also important is the number of feature attributes that require maintenance and the frequency of such maintenance. This may be captured by the **global schema requirements** measure.
Summary

This chapter has outlined seven measures that may be used to assess the relative difficulty of data integration for TNM transportation data. These measures incorporate the major technical and institutional constraints to integration. The **Number of participants** is telling of the degree to which coordination efforts must be extended within the specific data-sharing community. The **Number of participant schemas** simply accounts for the diversity in data design among participants. The **Aggregate participant boundary length** measures the linear distance for which horizontal integration must be completed. Neighboring participant heterogeneity measures the spatial configuration complexity of schema types. The **Aggregate participant boundary length** and **Neighboring participant heterogeneity** measures may be combined resulting in an aggregate linear distance of partnership boundary for which significant horizontal integration effort may be required. Data quality from participants is incorporated into the **Geometric integration complexity**. The **Local-to-Federated compatibility** measure allows for measuring the effect of changing global schema requirements.

A high-level conceptual data flow of TNM integration process has outlined four major aspects: Data acquisition, federation, edge-matching, and maintenance. Each aspect of integration presents unique challenges. But all components require that careful attention be paid to TNM product requirements. In the development of a technical implementation plan for the transportation theme, the complexity of integration should be evaluated with respect to the various political scales of data participation.
Chapter 5

Evaluating The National Map Integration Strategy

Introduction

Chapter 5 evaluates three levels of data partnership strategy and presents a recommendation for TNM implementation. Transportation integration strategies comprised of local, state, or national datasets are compared and contrasted with the aid of the measures developed in the previous chapter. Each of these integration alternatives is based on generalities and assumptions of local, state, and federal data programs, respectively.

In the later part of the chapter, the organizational development and overall program management recommendations are made first. Second, a strategy for addressing the complexity of data integration is presented. This suggested strategy focuses on the need for understanding the requirements of TNM and the careful consideration for the integration constraints articulated in this thesis.

Three target levels of integration

The three levels of data integration described herein generally relate to the American system of political hierarchy. This is done for three reasons. First, most public sector transportation GIS data development programs reside within county, state, or federal agencies – under the control of their respective budgets. Second, the clear delineation of boundaries between each level of government as well as neighbors within each level of government
facilitates visualization of the concepts discussed. Third, the utilization of political boundaries as a surrogate for dataset boundaries can be made with the data integration levels described as such. Each alternative will be compared for each aspect, or phase, of the integration data flow as they have been presented in Chapter 4.

Specific definition of integration levels

County level integration

County, or local, level integration assumes a strategy which seeks to integrate datasets at the political scale of county or equivalent, exclusively. With this approach, participant data for TNM would be schematically and cartographically consistent among county government providers.

State level integration

Like county level integration, state level integration assumes state agencies to be the primary provider of transportation data. As such, only one transportation dataset is procured for each state, of which all data are assumed to be of a single cartographic quality and schema type.

National level integration

The national level integration strategy essentially relies upon an existing source of transportation data of national extent collected and maintained by a single agency. This dataset is completely horizontally integrated at the state and local levels and is represented with the same schematic and cartographic entity representation at the national level. Examples of national level datasets include those from commercial providers, such as NAVTEQ, GDT, and TeleAtlas.
TIGER Enhanced data is also considered a national level dataset if TNM unconditionally accepts the dataset, without additional update or attribute requirements.

The above implementation perspectives are compared and contrasted with respect to each aspect of the data integration process.

Evaluating each integration phase

1. Data Acquisition

There exist integrated datasets of the United States for a national level implementation strategy. Public sources, such as TIGER, require quality improvement to meet the requirements of TNM. Three commercial datasets provide national coverage of the U.S. These proprietary data are, typically, more attribute rich than public datasets. For example, commercial transportation datasets provide road segment attributes that enable sophisticated routing engines. Turn restrictions, traversal impedance values, and number of lanes are just a few such road segment attributes provided by commercial datasets. On the downside, these datasets are high priced, may be revoked at the pleasure of the provider, and do not currently meet the Census Bureau’s accuracy requirements for the 2010 Census (U.S. Census Bureau, 2000).

Most states maintain, to varying degrees of accuracy, vector road datasets. A 2004 survey of state governments found the following (GIS-T, 2004):

- 31 states have a fully operational transportation GIS function; 18 states are in a transportation network implementation phase; and 1 (West Virginia) is in a planning phase for road GIS program establishment.
- 15, 7 and 26 states have digital database representation of U.S. and State highways, State and County highways, or all public roads respectively.
• At least 40 of 50 states maintained a road database of a 1:24,000 scale or larger.

Only Alabama, Kentucky, and New Jersey indicated no statewide data-sharing program. Furthermore, all but Hawaii and West Virginia indicated the presence of full time GIS support for their transportation data (GIS-T, 2003). The complete results from the 2003 survey are available online at: http://www.gis-t.org/yr2003/2003_State_Summary.htm.

By contrast, county level datasets are inconsistently available and much more difficult to source. In Georgia, for example, only Cobb County transportation GIS data were readily available. Based on visual analysis of Missouri and Georgia data, county data may also be more variable than state level data regarding quality and schematic design.

2. Federation - Local-to-National attribute translation

In considering which approach might be best suited for federating schemas, obviously, the national level dataset, once constructed, presents the lowest cost. Therefore, only the state and the county level partnering strategies will be contrasted. As stated in Chapter 4, the Number of participants and Number of participant schemas are measures that can help estimate the level of complexity for data attribute federation. Other measures mentioned, associated with federated schema requirements, were global schema complexity and estimate local to global compatibility. Given the limited information available on the specific attribute requirement of TNM, these measures will not be included in the immediate analysis. Significant attention is paid to the need for requirements in the recommendation section of this chapter.

The Number of participants, or contributing agencies, indicates the organizational complexity of the program while the Number of participant schemas alludes to the complexity of federating information. Attention is first paid to the Number of participants. With each state contributing transportation data to TNM, there would be 50 participants. In addition, the state
offices designated to deal with geographic data are consistently staffed and most already participate in some sort of data sharing program (GIS-T, 2004). Consistently staffed, funded, and open programs contribute to the likelihood of successful state-national cooperation in data sharing. Contrast this with the number of U.S. counties: 3,142 (U.S. Census Bureau, 2004). County level organizations, as mentioned, are less likely to have transportation mapping programs. One could reasonably argue that county level agencies, with less tax base for program funding, are less likely than state agencies to spend time or have the skills to effectively implement and coordinate data integration programs with federal level agencies. Further, local agency personnel tend to have a more complete mental map of their jurisdiction, resulting in a relatively lower marginal benefit of a GIS mapping program.

The Number of participant schemas is, most likely, highly correlated with the Number of participants. Given the existence of organizations like the American Association of State Highway and Transportation Officials (AASHTO) and forums like the GIS for Transportation Symposium (GIS-T), state agencies seem to interact with one another more so than local government transportation departments. This interaction fosters shared approaches to data development, maintenance and storage. Therefore, not only is the number of schema types expected to be greater at the county level than the state level, but also the diversity of county GIS are expected to be greater as well.

3. Integration- Edge-matching

The extent of edge-matching required for integrating datasets is dependant on the length of the border between all participants to be edge-matched (Aggregate participant boundary length), the spatial configuration heterogeneity of dataset characteristics (Neighboring participant heterogeneity), and the accuracy of participants’ data (Geometric integration data).
complexity). First, an attempt is made to estimate complexity of integration with regard to the Aggregate participant boundary length.

An example of the difference in participant boundary length between state and county level integration projects, the aggregate coincident county boundary length and state boundary length four southeastern states of the United States (South Carolina, Georgia, Alabama, and Florida) are compared below. This measure excludes county and state boundaries that border the ocean, where no coincident boundary exists, and non-integrated borders, such as the border between Alabama and Mississippi. The county and state boundaries used are displayed in the maps below (Figures 5.1 and 5.2 respectively).

**Figure 5.1** – Map of county level aggregate participant boundary length.

**Figure 5.2** – Map of state level aggregate participant boundary length.
There are a total of 30,610 kilometers of coincident county boundary among these Southeastern states. Comparing this to 1,700 kilometers of coincident state boundary among these states reveals an 18:1 ratio of county to state edge-matching distance. This ratio varies with the shape of the integration area and the shape and number of the constituent county and state boundaries. For instance, a perfectly circular integration area would minimize the political boundaries excluded by the integration area periphery. In this Southeastern states example, for instance, most of the state boundary length of Florida is excluded. Nonetheless, this example demonstrates the magnitude of the difference in edge-mapping between county and state level network integration.

Complementing the length of participant boundary measure is the Neighboring participant heterogeneity measure. As stated earlier, this measure accounts for significant, definable classes of dataset characteristics. These classes may be measured as a categorical, ordinal or continuous variable and may incorporate significant accuracy and content attributes in its definition. The significance of database characteristics is defined, in part, by the data requirements for integration (i.e. What is considered seamless and consistent?). But significant partner characteristics are also defined by the capabilities of integration procedures, such as automatic spatial and attribute conflation routines. Even assuming that the rate of database characteristic difference is constant across political levels, the sheer volume of counties would yield a greater value for the Neighboring partner heterogeneity measure with county level integration than with state level data integration. This is made more evident if participant border length is included in the weighting of this measure, as is illustrated in figure 4.8.

Geometric integration complexity measures the relative complexity of topological integration ascribed specifically to the accuracy of the input datasets. This measure accounts for
the cost of poor data quality. Spatial accuracy also determines, to a large degree, the difficulty of vertical integration. Figure 5.3 demonstrates the difference in spatial data accuracy between St. Louis County the state of Missouri transportation data.

Figure 5.3 – Transportation and Orthophoto Overlay. Homeland Security Infrastructure Program (HSIP) 133 City Orthophoto Imagery (2000) with 2003 Missouri DOT road database (orange) and 2003 St. Louis County road database (yellow) overlayed.

In this instance, the least accurate of the county and state datasets is the St. Louis County dataset. Both datasets are derived from TIGER line transportation files. The Missouri Department of Transportation (DOT) data have been spatially enhanced from the original TIGER data, unlike the St. Louis County data. However, simple image analysis reveals that both datasets
are out of date. The area in the center-right of the image shows the expansion of a residential neighborhood. Neither the Missouri DOT nor the St. Louis County datasets contain these roads. In Georgia, the Cobb county dataset was more current than the state of Georgia dataset. Again, transportation data maintained at the county level for Georgia was sparse. The National Map requirements regarding vertical registration and spatial accuracy, comprehensiveness of coverage, and temporal accuracy can be used to guide the inputs for the geometric integration complexity measure. This measure is most useful when data exists for the vast majority of the geography to be integrated. Considering the lack of land area sufficiently mapped at the local level, the geometric integration complexity measure should be evaluated along side the percent of the geographic area sufficiently mapped. One would then conclude that state level data sourcing is preferable to local level sourcing, assuming all other factors are constant.

4. Maintenance

The Number of participants is a good measure of the complexity of the ongoing maintenance for TNM transportation. Based on this measure alone a state level partnership program will be less troublesome than a county level program. Of course, the data requirements play a role here as well; the more stringent the spatial-temporal requirements, the greater the data and institutional interoperability mandated. Agencies with full-time GIS staff are best equipped to develop the long-term partnerships required for developing cross-agency interoperability. More often than not, such agencies reside at the state level. That said, specific counties and most major municipalities might offer valuable resources.

In order to institute an incremental update system, the TNM transportation will require database transactioning capability. Coordination with a great number of smaller, relatively less
skilled staff at the local level would decrease the likelihood of a successful implementation. This is not to say that county and city government agencies should be neglected in data sourcing efforts. Nor should they be discouraged from contributing to directly to the federated system. The contribution of all levels of government and the private sector are considered in the final section of this thesis. The complicating factors of road data integration discussed in the preceding chapters and information on TNM program activities to date have been referenced in the development of the following recommendations.

Summary

Three levels of transportation data integration can be used to describe the target political scale of primary providers for TNM: County or Local, State, and National. A hypothetical integration project is used to evaluate each defined phase of integration in the light of the measures developed in Chapter 4. At the National level, a single transportation database would be maintained for the entire country. Little physical integration is needed yet this is a costly option that, most likely, will not meet data accuracy demands. At the State level, the primary data providers for TNM would be state agencies. State transportation information would be integrated, forming a national level composite database. A Local level integration focus calls on counties and municipalities to be the primary provider of transportation data.

The complexities of integration form both the institutional and technical perspectives suggest that the more local the focus for primary data providers, the more complicated the integration process with little, if any, additional content data improvement. The next chapter will present a recommendation that considers these complexities with regard for TNM management plan and vision.
Chapter 6
Policy Recommendation

Introduction

The following recommendations address fundamental issues of strategy development and strategic management. First, suggestions surrounding the existing difficulties of TNM organization and management are put forth. It is argued that proper organizational structure in support of TNM must be closely tied to NSDI in the form of an overarching authority. Furthermore, TNM thematic definitions and functionality must be based on a thorough requirements study. The second focus of this chapter presents a recommended approach to developing a National Map Transportation Database. This approach suggests the virtual integration of certified data providers from national, state, local and private organizations.

National program management

The impediments to data sharing and integration highlighted in this thesis, when considered with TNM and NSDI visions, present two fundamental project realizations:

1. Successful development of TNM is dependent upon an up front understanding of TNM stakeholder community needs and capabilities,

2. Long-term implementation must find a balance between mandated compliance with standards for participants and local control of data management.
To deal with these realities, the following program recommendations are articulated. First, the need for an overarching authority to manage TNM, and “sister” programs, should be established immediately. Project leadership should subsequently be defined. To facilitate cost estimation, allowing for the development of a reasoned plan, TNM content and functional requirements should be detailed and communicated. Further, the content for the transportation database should be derived from a distributed network of affiliated, standards-based national, state, and local mapping agencies. Each of these suggestions is now described in detail.

**Overarching authority**

Recognizing the importance of TNM and its relation to other federal geospatial initiatives is critical for the efficient management of spatial data. The participant and potential stakeholder community of TNM reach far beyond the bounds of USGS authority (NRC, 2003). In fact, there is considerable overlap with other federal initiatives to consolidate spatial data management. GeoSpatial One-Stop (GOS) and the Census TIGER Enhancement program are two other major geographic data collaborative efforts underway. While the USGS (2003a) implementation plan recognizes a need to coordinate efforts between GOS and TNM, the U.S. Bureau of Census and USGS data sourcing and data integration program development remain fairly distinct, in practice. For example, while Harris Corporation is trying to develop means for developing partnerships and integrating local transportation network geometry with Census geometry, the USGS is independently soliciting local data sharing and researching integration strategies for TNM. Further, the Census and USGS have independently maintained a database of state and local government organizations that share data with the Census and the USGS respectively. This has
resulted in both the USGS and Census soliciting the same information from a single, local government, illustrating unnecessary data sourcing duplication.

Serious consideration should be given to consolidating TNM activities with other geospatial initiatives, such as the GOS data portal development program and the TIGER Enhancement program. It is important for the administration to recognize that all federal level geospatial data collaborative should reside under the umbrella of an overarching NSDI implementation authority. As Ventura (1995) argues, an overarching body must be recognized and granted clear authority over the participating organizations. The overarching body would function as the NSDI implementation champion. The committee should be comprised of appointed officials that serve without any perceived bias toward any one agency (Johnson et al. 2001). This overarching body must hold executive authority, including task funding (Ventura, 1995), and accountability over all NSDI implementation projects.

Currently, the FGDC would be the government entity best suited for meeting the managerial demands of the NSDI mission. However, there exist some problems with the current FGDC structure and functioning. First, although carrying significant political weight, the FGDC exists by Executive Order, therefore, no funding is granted to it. Second, Congress does not appropriate directly for FGDC therefore no law exists to enforce its recommendations. The FGDC is only recognized as a “good” thing but can be, and is, ignored if and when budget constraints within an agency force cuts. Third, some FGDC members are not actively participating in its activities. Likewise, some federal agencies are not employing FGDC's standards. A National Academy of Public Administration panel found that FGDC's strategy for implementing NSDI is not reflected in agency strategic plans and annual performance plans developed under the Government Performance and Results Act (NAPA, 1999). It must be
acknowledged that many of these project issues have much to do with the lack of political mandate for TNM project.

The conclusions reached here, then, suggest that legislation should be passed that establishes an overarching authority for federal geospatial information consolidation. This authority should be armed with a budget drawn from the appropriate, existing agency budgets. Ultimately, it should be held accountable for successful execution of TNM and other NSDI initiative.

**Project leadership**

Following the identification of an overarching authority, it is critical that TNM leadership first establish a project management structure that defines and delegates priorities. Poor coordination and project leadership are the leading factors in the termination of data sharing programs (Nedvodic-Budic and Pinto, 2000). Currently, there is a lack of formal project structure within USGS for making efficient progress toward the ambitious vision of TNM. Formality will become increasingly important as USGS Geography expands the partnership program with other government entities (Johnson et al. 2001). This lack of formality has allowed for confusion in the responsibility of project personnel and the urgency of assigned tasks.

**Project requirements definition**

Once clear project authority is assigned and leadership is established, making the transition from the grand vision statement to a feasible implementation plan will require a thorough evaluation of the geospatial community needs, at all levels of government. Understanding the data, structures and procedures that are required by all classes of participants
prior to implementing hands-on development plans provides the best opportunity to employ the correct implementation strategy. As an example, the Australian state of Victoria successfully implemented a partnership program whereby the state provides the local governments with basic map data and the local governments provide the state with parcel data. After years of floundering, the first step toward success culminated in one of the largest requirements studies ever conducted by an Australian State (Jacoby et al. 2002). Currently, data standards teams for TNM have been formed and tasked with detailing the content standards for each of the eight geospatial themes. However, they have no documented data requirements from which to work nor do they have a mechanism for coordinating their efforts. In addition, the current edition of the implementation plan states that a full requirements study will be addressed during Stage II of TNM implementation. Stage II will not commence until fiscal year 2006, at the earliest (USGS, 2003e). Yet, the arguments in this thesis consistently illustrate the need for requirements before implementation cost assessments can be made.

With the notion of exchange theory in mind [an inter-organizational data-sharing program requires the perception of mutually reinforcing benefits (Fountain, 2001; Azad and Wiggins, 1995; Cook, 1977)], it follows that a comprehensive requirements study that includes representatives of each stakeholder agency at the federal level, and primary spatial data functions at each political scale is paramount. Without such a requirements study in Stage I of TNM implementation, how are data sourcing results judged to be sufficient? How are pilot implementations judged to be successful? This goes beyond inter-agency meetings and conference note taking as may be inferred from the current list of TNM accomplishments (USGS, 2003d, 41). Rather, a thorough requirements study mandates procedures for decision-making, accommodates significant stakeholders, and produces clear, precise organizational,
systems, and data content expectations. Beyond input from public agencies, a great deal of GIS expertise in spatial data collection, management, and distribution can be garnered from the private sector as well.

**The National Map content**

TIGER Enhanced data gathered with cooperation from state and city or county data, where useful, will constitute the road database. Commercial data providers should also be accessible through TNM and GOS portal to subscriber level users. Recognition by the USGS of the organizational impediments to data sharing constitutes their argument for the need of a zero-mandate policy with state and local governments. However, such an under-bound network of partners may produce a National Map product of such low quality that its cost may not be justified. The National Map overarching authority will need to define minimal standards for all road attribution determined to be essential to TNM beyond that provided by TIGER. This would be completed via the aforementioned requirements analysis. A conceptual model for a federally consolidated road database is presented below.

**Accomplishing a single national network**

The long-term solution to integrating national street centerline data is to develop a distributed database of national extent that meets the scale requirements of all parties. Jensen et al. (1998) call for the need to establish a centerline database accessible to all levels of government. Such a database could be maintained at the local level with no duplication of effort at the federal level. Centralized control of standards, informed by all stakeholders, would constitute the mechanism for information exchange and interoperability.
Figure 6.1 illustrates the suggested, long-term conceptual data flow for TNM transportation. Although such a program cannot be implemented immediately, the overarching authority of NSDI needs to develop a plan for achieving such a distributed non-redundant data maintenance program. The phased project approach should be used to determine the means for achieving this vision.

The schematic is annotated with a stage number. Generally, the flow of information is from left to right. Participating data contributors maintain data on the left, committing it to a centralized road data repository via a translation process, and a mapping service application, including TNM rendering, is represented on the right of the diagram.

**Stage 1: Data Partners**

The first stage of the data flow recognizes the participating data providers to TNM transportation theme. Under this design, each certified data providing agency will supply data to TNM using one of many available, standardized local system-schema designs. This “menu” of schema designs will derive from the Stage I requirements study, identifying the most common data structures among the data providers. Data will reside locally with each participating organization. Eventually, the majority of transportation data content within TNM primary network will come from state or local agencies, with the TNM overarching body coordinating relationships and managing standards. All data providers must adhere to these minimum content standards and capture conditions in order to ensure system integrity and maintain certification from TNM.

Census TIGER files, it is recommended, should provide the base network for transportation. Many local government transportation networks are based on TIGER files.
Figure 6.1 – This illustration diagrams a conceptual data flow for TNM. Data originates from primary data providers on the left and is translated to a federated database. The National Map road theme is generated from this database.
Therefore, many local GIS agencies are familiar with the format and history of TIGER data. In addition, the TIGER Enhancement project is actively soliciting local participation in improving the spatial accuracy of its network. Although no network maintenance plan beyond the 2010 Census has been publicly discussed, it is recommended that a decentralized, ongoing data development effort be initiated.

Another federal agency that may have unique road network requirements is The U.S. Forest Service. They need forest fire roads mapped. Since this demand is unique to them at the federal level, perhaps they will then maintain public land fire road geometry and attribution. *The objective is to have each agency fund the data development effort proportional to their respective use.* Data flows from each distributed database through a translation process and to a national road database. Local reference to neighboring partner data for edge-matching may be accomplished via local and regional virtual networks.

**Stage 2: Schema Translation**

With each participant storing transportation data within a known system and schema, automated data translation procedures customized to each schema “menu item” may be developed for importing all participant data to a centralized database. Data from each provider will be committed periodically, based on demand for each partner’s data, the frequency of network change, and schematic translation processing capacity constraints.
**Stage 3: Centralized National Road Database**

The road database will hold all geometry and primary key assignments for road segments and nodes, as well as minimal attribution for cartographic display. However, the majority of the network attribution will remain distributed. The road segment primary key will be a composite persistent identifier consisting of 1) the data provider’s own road segment persistent identifier and 2) the persistent identifier of the agency. This removes the chance of having duplicate primary key attribute values within The National Road Database. The Census Master Address File will be related to the road database similar to the manner in which it is currently related to the TIGER files. Network continuity validations, nation-wide quality assurance, or other analytical procedures may be developed and applied to the road database by USGS researchers or other authorized agencies interested in the continuity of The National Road Database. With transactioning attribution, only objects created or modified subsequent to the most recent local to global commit need to be reconciled to the National Road Database.

**Stage 4: Cartographic Rendering**

For the purpose of displaying the road database via TNM, USGS or government contractors will maintain a default cartographic template. Any interested party may develop custom templates for various purposes. The template will also incorporate other data sets to be displayed on TNM, including NSDI themes that are not common to TNM and views for restricted (non-public) access data. Maintaining the cartographic template, and the web mapping service (WMS) more broadly, may be a function the federal government should consider outsourcing to the private sector. There is a wealth of web mapping experience and success in private industry that should be tapped.
Stage 5: The National Map Product

TNM will provide the cartographic window to the Nation’s transportation data providers. Ultimately, all TNM data will be viewed through this WMS. Again, a commercial applications service provider may be contracted to host TNM site.

Thus far, only the primary, public transportation data have been recognized within this data flow. However, TNM will have multiple access levels, from unrestricted (public) access, to paid subscription access for commercial data, to classified levels for national security and emergency management access only. There will undoubtedly be a need for other, more robust, transportation networks to support sophisticated routing applications. Although a desire exists to have Census maintain or source routing attributes on its Enhanced TIGER network, the reality is that such an effort would be cost prohibitive. An application of the local-to-global compatibility measure presented in this thesis would undoubtedly bear this out. A network that supports a reasonably valuable routing application not only requires detailed attribution, such as travel direction, but also requires topological design and maintenance considerations that are beyond the framework of the TIGER data model.

The public sector would be wise to leave routing database development to commercial data providers. Organizations may be granted various levels of access to these networks via a TNM subscriber service. Of course, the commercial data venders such as NAVTEQ, or TeleAtlas would receive royalties for the use of their data. That said, the government should be able to take advantage of the extended capabilities of commercial road networks and existing, affiliated software, such as private routing engines and commercial geocoding applications. The
comparable spatial precision and addressing attribution of The National Road Database and commercial road databases would allow for the capability to reverse geocode event locations between networks quite accurately.

Stage 6: Other Data Theme Relations

Other data themes or datasets may be explicitly or implicitly related to the road database. The dotted line in figure 6.1 signifies that other data sets may have foreign keys to the road database or vice versa. Records within the structures theme model, for instance, may require a foreign key to a road segment identifier. Explicit linkages across themes must be considered in the design of TNM as a whole. This stresses further the importance of a coordinating requirements study and subsequent standards development effort among data themes. As discussed above, other transportation data may be indirectly related to the road database via network conflation applications. Spatial data sets should be available for view within a National Map access layer (public or restricted) and should also be registered with the Geo-Spatial One Stop portal (GOS).

Stage 7: Geospatial One-Stop Relation

It is recommended that TNM be coupled with GOS via a shared gazetteer. The National Map would serve as the data view of information within GOS. Of course, the core data themes within TNM are expected to be nationally seamless and compliant with explicitly defined standards, whereas the GOS information serves as a clearinghouse with relatively few participatory demands. But users should have the option of identifying participant data sets for
download via TNM interface or, conversely, by directly selecting the data sets listed within GOS for view in TNM. This allows users to relatively quickly assess the usability of a GOS listed dataset with the national base map (TNM). Secondly, it allows further opportunity to highlight inconsistency in entity representations between datasets.

TNM transportation team should consider the alternative described in these seven stages. It is unlikely, however, that this approach would be adopted without political elevation of TNM. Nonetheless, some of the benefits of this recommended approach are listed below:

1. Primary key attribution of road segments is resolved with the above approach. Unique identification of road segments and nodes within a data structure will be a requirement of American National Standards Institute’s (ANSI) transportation standard (FGDC, 2003).

2. Maintenance of a global schema is greatly simplified when a moderate level of schematic standardization is applied to the participant network. There is no longer a need to constantly monitor each participant’s data for changes in schema design or capture conditions. With the present USGS approach, the technical data structure may stay the same, however the local data capture conditions may change. Without a communication mechanism to TNM, such a change would go unnoticed. For example, a local government may decide to no longer map alleys, driveways, and unpaved roads due to local budgetary constraints. The database structure may remain the same but the maintenance procedures that dramatically affect content would change. Such procedural changes may be invisible to the USGS without a coordinating effort from the local data provider. Based on the assessment of local and state datasets, it is recommended that minimal schematic design and capture conditions be adhered to by all participants, allowing for both a temporally and spatially consistent road database for the long run.
3. Inconsistent, unstandardized content may require rapidly evolving integration procedures that are customized to individual data providers. Minimum standards reduce the level of customization required for data integration. In the long-run, data contributors should be encouraged to maintain datasets in a way that minimizes the neighboring participant heterogeneity measure presented in this thesis.

4. Partnership interoperability is encouraged by the suggested National Road Database design. Existing software developed around TIGER and commercial datasets will be utilized and expanded. With some level of standardization, applications developed by one municipality could be more easily marketed to other municipalities – considerably decreasing the total public cost of system development. Conversely, the current heterogeneity of state and local database schemas, capture conditions, and content quality may make it difficult to perform spatial or network analysis across partnership datasets - effectively limiting the potential benefits of the national transportation data inventory.

5. Ideally, the proposed strategy would establish neither an overbound nor underbound network organization (Golembieski, 2003 pers. com.). Without any standards being established at the participant level, massive confusion and inconsistent quality will characterize TNM maintenance procedures. At the same time, it is recognized that too much top down control will increase the participatory cost of local governments, perpetuating redundant, public GIS development. The standards development process must recognize the diversity of local data management systems and refrain from mandating how to maintain spatial data. Rather, standards should be data focused, merely addressing basic schema design and minimum database content.
Why not use commercial transportation data as the primary provider?

A great deal of consideration was given to the use of commercial data as a primary road data source. After all, data providers such as NAVTEQ maintain spatially precise databases with rich attribution and nation-wide coverage. There would be no need for horizontal integration if a single data provider were selected to provide road data for the entire nation. But the competitive market is the motivating factor that drives commercial data providers to create such robust data sets. With the adoption of a single commercial provider for roads, and its subsequent integration with other TNM themes and applications, the competitive pressure for the selected commercial data provider would diminish over time. Ultimately, the lack of competition would impact the quality of service provided by the commercial licensor. Consideration for the use of commercial road data as a primary provider in the short-term should not be ruled out, however.

Future research

Although this thesis outlines the major complicating factors of road data integration for TNM, empirical research regarding integration costs need to be produced. One source of empirical data may be The National Map pilot projects by way of documenting the cost associated with the different phases and tasks of road data integration. Research which highlights the relationship between significant dataset characteristics (particularly characteristics that can be mined from metadata) and integration cost would provide an empirical foundation to many of the measures presented in this thesis. Data integration procedures also need to be developed. More sophisticated and effective data conflation routines engender lower integration costs.
Conclusions

Program management

Paramount to TNM implementation is an administration level recognition of the need for an overarching authority for consolidating the Nation’s geospatial data. The vision of a national base map offered by the USGS merits the full attention of the geospatial community. However, the successful implementation of this national mapping program is beyond the control of the USGS. Procuring data, developing appropriate data maintenance strategies, and coordinating geographic information sharing partnerships encompasses more than twelve federal agencies (USGS, 2001), forty federal functions (NAPA, 1999), and countless state, local and private organizations. Further, TNM vision is very closely related to other federal geospatial consolidation programs, such as the U.S. Census MAF/TIGER Enhancement Program. A viable long-term approach to TNM development will require the direction of an objective, accountable, overarching body that will establish geospatial product requirements and develop a national geospatial consolidation plan. Perhaps such an authority might only be constituted by a congressional mandate.

The current project management structure in the research and development of TNM is lacking within the USGS. Premature activity regarding system design and development abounds as TNM vision has begun large implementation efforts prior to developing sound implementation plan. A requirements assessment for TNM stakeholders is far from complete and with no apparent beginning.

The long-term issues of integration and interoperability must be addressed at the outset of TNM development. With data requirements as a guide for implementation, an overarching authority must consider the costs and benefits of various alternatives to road data strategies. A
heuristic integration complexity model is recommended to assist in the cost assessment of each strategy. The total estimated cost of each strategy should be weighed against the benefits of their respective results.

**Road Integration Strategy**

A viable, long-term approach to TNM road data includes the development of a distributed road database. A combination of state, local, and national data contributors would maintain such a database. Each contributor should be responsible for maintaining road data within its geographic jurisdiction while conforming to minimal standards that provide for national base data consistency. Minimal standards should be derived from thorough requirements analysis and practical considerations. The measures developed in this thesis can assist in this endeavor by providing for the quantitative comparison of complexity and risk among alternative strategies.

The existing variability in geospatial data content, quality, and design highlight the need to more carefully consider technical integration constraints. The content provider variability, however, should be defined relative to the data federation constraints and minimal content requirements of TNM. Measures such as the **number of participant schemas** and the **global schema requirements** can quantify the detail of federated schema demands. While the **neighboring participant heterogeneity**, **geometric integration complexity**, and **local-to-global compatibility** measures help to quantify the relationship between the data provider characteristics and the federation requirements.

Previous research indicates that organizational issues may present the greatest challenge to TNM success. The **number of participants** is then, perhaps, the most telling predictor of data sharing complexity. The more primary stakeholders there are and the more data providers there
are, the less likely all data demands will be met. With more players in the game, it is usually more difficult to agree on the rules and compensate equitably.

While more empirical research is needed to fine-tune these complexity measures and determine TNM requirements, a few general conclusions can yet be made.

1. The MAF/TIGER enhanced dataset should provide for the initial, comprehensive TNM road base. Requirements extending beyond those of this Census enhancement program will, most likely, be met at a high marginal cost. Additional requirements will make necessary additional sourcing and data discrepancy resolution.

2. While geographic “pockets” of local level data may be valuable, TNM should focus on state level agencies as primary data providers for information needed beyond that provided by TIGER. State agencies have more comprehensive data coverage and almost all have full time GIS personnel than local agencies. The relatively fewer number of agencies and shorter aggregate border length of state jurisdictions support this conclusion.

3. The long-term goal of a nation-wide, distributed national road database will require the presence of an overarching authority which consolidates the demands of major federal stakeholders under one decision-making body. This would alleviate many of the competing political incentives which are currently hindering the NSDI.

TNM can be implemented successfully with more thoughtful consideration of the broad geospatial community. Careful attention must be paid to public spatial data demands across governing levels, not merely among federal agencies. Its place within the NSDI must be clearly
defined and managed at the appropriate executive level. If properly implemented, TNM will serve The Nation well through more effective governance, improved geospatial research, and an unprecedented window to the world for the public.
References


Biggs, Ed. 2003. *Building an enterprise geographic information system, the five year GIS implementation plan*. Cobb County Government, Cobb County, GA.


Golembieski, R. T. 2003. Personal communication. Professor of Public Administration, University of Georgia, Athens, GA.


University of Georgia, Information Technology Outreach Services. 1998. DLG-F data dictionary, transportation.


