Data Layer Integration for *The National Map* of the United States

E. Lynn Usery
usery@usgs.gov
Michael P. Finn
mfinn@usgs.gov
Michael Starbuck
mstarbuck@usgs.gov
U.S. Geological Survey
1400 Independence Road
Rolla, MO 65401

The integration of geographic data layers in multiple raster and vector formats, from many different organizations and at a variety of resolutions and scales, is a significant problem for *The National Map* of the United States being developed by the U.S. Geological Survey. Our research has examined data integration from a layer-based approach for five of *The National Map* data layers: digital orthoimages, elevation, land cover, hydrography, and transportation. An empirical approach has included visual assessment by a set of respondents with statistical analysis to establish the meaning of various types of integration. A separate theoretical approach with established hypotheses tested against actual data sets has resulted in an automated procedure for integration of specific layers and is being tested. The empirical analysis has established resolution bounds on meanings of integration with raster datasets and distance bounds for vector data. The theoretical approach has used a combination of theories on cartographic transformation and generalization, such as Töpfer's radical law, and additional research concerning optimum viewing scales for digital images to establish a set of guiding principles for integrating data of different resolutions.

Key words: data integration, *The National Map*, federated GIS data, cartographic theory

INTRODUCTION

The U.S. Geological Survey (USGS) has begun a new program for supporting the needs of the nation for topographic mapping in the twenty-first century. That program is referred to as *The National Map* and involves a vision of:

> information current, seamless national digital data coverage to avoid problems now caused by map boundaries, higher resolution and positional accuracy to better support user requirements, thorough data integration to improve the internal consistency of the data, and dramatically increased reliance on partnerships and commercially available data. (USGS 2002)

This vision includes the development and maintenance of eight data layers: transportation, hydrography, boundaries, structures, elevation, land cover, orthographic images, and geographic names. The data will be available over the World Wide Web (WWW) and accessible for both direct viewing on the Web and for download by users. Data will be comprised of the best available source, and the USGS will depend on state, local, tribal, and other government organizations and private industry to supply data.

Initial submission, July 31, 2008; final acceptance, October 1, 2008.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

The USGS will become a data producer only in cases where no other data are available.

The problem of using data from such a variety of sources is positional and thematic integration of the various resolutions and accuracies of data. Data must be positionally, sometimes referred to as horizontally, integrated to provide the seamless nationwide coverage as specified and thematically, sometimes referred to as vertically, integrated among the different themes to provide internal attribute consistency (Gösseln and Sester 2005). A large part of the data integration problem lies in matching points or features between data sets with different ontologies, data models, resolutions, and accuracies. A variety of methods to achieve feature matching have been developed for multiple vector data sets including an iterative closest point algorithm by Gösseln and Sester (2005), a statistical approach using measures from information theory by Walter and Fritsch (1999), and a data modeling approach in used by the Institut Geographique National in France (Devogele, Parent, and Spaccapietra 1998). For integration of vector and raster datasets, research has been focused on transportation and image datasets (Chen, Knoblock, Shahabi 2006; Wu, Carceroni, Fang, Zelinka, and Kirmse 2007). With the federated database approach (Sheth and Larson 1990; Devogele, Parent, and Spaccapietra 1998), *The National Map* has significant vertical and horizontal data integration problems, and the USGS continues research to develop procedures to accomplish this integration (Finn, Usery, Starbuck, Weaver, and Jaromack 2004). It is the purpose of this paper to document some of our progress to date and to better define the exact nature of the data integration problems. Specifically, Section 2 addresses the basic meaning of the term *data integration* in raster, vector, and combined geometric domains. Section 3 details our basic approach, data, and study areas. Section 4 documents an empirical study to determine the visual meaning of data integration. In section 5, the basis of a theory for integration is presented. In Section 6, we document an automated approach for vector and raster integration based on transportation and orthographic images. Section 7 provides further discussion with our conclusions for a theory of integration based on the concepts of scale and resolution ratios, optimum viewing scales, and image fusion presented in Section 8.

"The problem of using data from such a variety of sources is positional and thematic integration of the various resolutions and accuracies of data."

1.0 Data Integration Definition and Visualization of the Problem

The concept of an integrated dataset of various layers is based on the approach used in the standard five-color lithographic topographic map, which the USGS has produced for decades and provided to its customer base. In the same way that all features of different types on the lithographic map are co-registered and integrated into a single document, digital data sets need to register and integrate in a similar fashion. A major difference is that the USGS produced all the data for the topographic map and could force resolution and accuracy limits to maintain an integrated product. In the current environment of *The National Map*, data are provided by a variety of sources and at a variety of resolutions and accuracies. Forcing consistency is no small achievement, and simply establishing the meaning of an integrated dataset poses difficulties. For example, Figure 1a shows transportation and an orthographic image in an area west of St. Louis, Missouri. The image is a color orthophotograph from Nunn-Lugar-Domenici 133 priority cities of the Homeland Security Infrastructure Program (Vernon, Jr. 2004) with 0.33m (1 foot) pixel size, which approximates the resolution. The transportation file is from the Missouri Department of Transportation (MODOT) and provides one of the most accurate

"The concept of an integrated dataset of various layers is based on the approach used in the standard five-color lithographic topographic map, which the USGS has produced for decades and provided to its customer base."

"What does it mean to be integrated?"

sources for this area. Note the mismatch between roads as shown on the image and roads from the vector data file. Is this an integrated dataset? We provide a second example in Figure 1b using the same area and the same orthophotograph, but with Census Topologically Integrated Geographic Encoding and Referencing (TIGER) line files for a transportation source. The base source of the TIGER data is the USGS 1:100,000-scale topographic maps. As is evident in this example, the TIGER data are not integrated well with the image. Note that in both cases we really have not integrated the datasets; we have merely provided an overlay of the roads on the image. A final example is shown in Figure 2 including hydrography data overlaid on the same image base. In Figure 2a, the hydrography source is the USGS National Hydrography Dataset (NHD) while Figure 2b shows hydrography from St. Louis County. The St. Louis County data are certainly better and actually show the streams as double lines, but these data still do not match the image exactly. What does it mean to be integrated?

We take the position that integration means the datasets match geometrically, topologically, that is, have the same spatial relationships in the

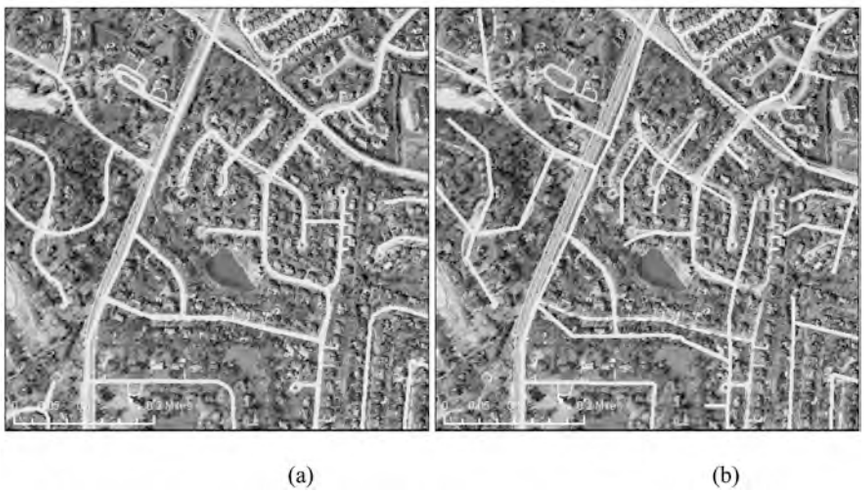

Figure 1. MODOT transportation overlaid on an orthographic image is shown in (a) while Census TIGER transportation overlaid on the same image is shown in (b). (see page 59 for color version)

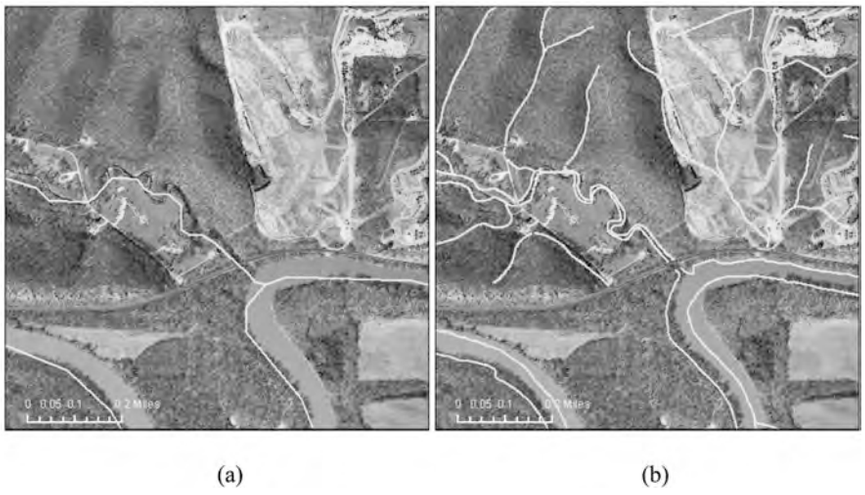

Figure 2. Shown in (a) is hydrography from USGS NHD whereas (b) shows hydrography from St. Louis County. (see page 59 for color version)

data as those that exist in the real world, and have a correspondence of attributes. Thus, from the point of view of position, to be integrated, the vectors from the transportation and hydrography files in Figures 1-2 need to follow or match the corresponding features in the images. Further, if we have such a match we can fuse the vectors into the image without loss of information since the vectors will align. From a thematic integration viewpoint, two maps must share exact attribution so an extension of a feature from one horizontal partition to another remains the same feature with the same attributes.

Positional and thematic integration of vector and raster data are discussed above, but what does it mean to have two integrated raster datasets? For example, from *The National Map*, we use the USGS National Elevation Dataset (NED). This dataset includes data at 1, 1/3, and 1/9 arc-sec resolution (approximately 30, 10, and 3 m, respectively). The orthographic images for urban areas are 0.33 m resolution. If we integrate the elevation data, perhaps in the form of a shaded-relief presentation, with the image, we combine approximately 8,100, 900, and 81 image pixels, respectively, to match one elevation pixel of the resolutions of NED (Figure 3). How do we know when two raster datasets are integrated? We can base successful integration on the geometric frame of reference, but visually does it matter? In the case of a lake, the elevations should be flat; and, with flowing streams, the water should flow downhill, but can we really determine that with large resolution differences? One of the goals of our work has been to try to define the limitations, based on resolution and accuracy, at which datasets can be realistically integrated.

"From a thematic integration viewpoint, two maps must share exact attribution so an extension of a feature from one horizontal partition to another remains the same feature with the same attributes."

We have a similar problem if we discuss integration of two vector datasets, which, in *The National Map*, are layers for transportation, hydrography, boundaries, and structure outlines. For transportation and hydrography, positional integration should yield locations of bridges, culverts, and other structures. Resolution issues abound here as well, but accuracy appears to be a larger issue as shown in Figure 4 where the stream follows the road centerline.

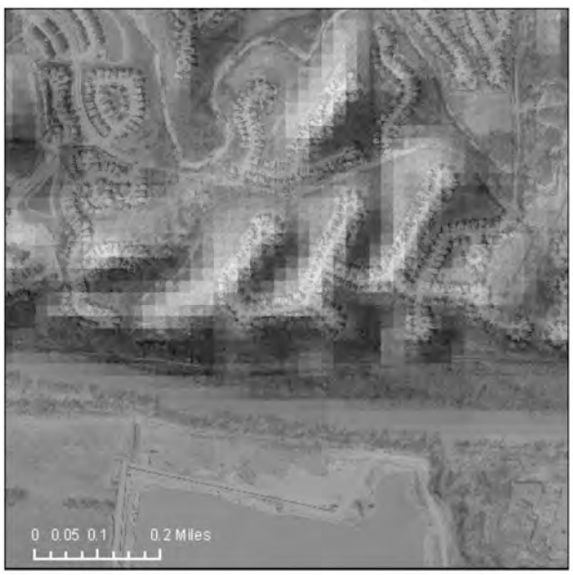

Figure 3. An orthographic image with 0.33 m pixel overlaid with elevation data with 30 m pixels. (see page 60 for color version)

Figure 4. Vector data for roads (red) and streams (blue) overlaid for the same areas. Note the area in the purple circle where the stream follows the road centerline. (see page 60 for color version)

3.0 Approach and Study Areas

Our approach includes an empirical exploratory analysis to establish a meaning, both visually and numerically, for data integration; theoretical development and proposition generation for data integration feasibility based on resolution and accuracy; and algorithmic development of procedures to shift features from one dataset to match a second to accomplish data integration. We selected five datasets and two test sites. The data include transportation, hydrography, land cover, elevation, and orthographic images (Table 1). We selected test sites over St. Louis, Missouri, and Atlanta, Georgia, based on the availability of the five data layers for testing. We used the available data for the test sites, which at the time was limited to 30 m resolution elevation and land cover. Accuracy in the table is from the accuracy specified for the dataset or that in the metadata.

"The empirical testing was accomplished by overlaying one dataset on another, producing printed versions of the overlaid datasets, and conducting a visual analysis using a set of respondents to judge the effectiveness of the integration (match) between features in the two datasets."

The empirical testing was accomplished by overlaying one dataset on another, producing printed versions of the overlaid datasets, and conducting a visual analysis using a set of respondents to judge the effectiveness of the integration (match) between features in the two datasets. The proposition development is based on concepts from cartographic theory, including the radical law (Töpfer and Pillewizer 1966), known limits of generalization methods, the relation between raster resolution and map scale (Tobler 1988), an empirical analysis of viewing scale (Fleming, Jordan, and Madden 2005), and an examination of the results from image fusion methods for remotely sensed images of varying resolution (Ling 2006; Ling, Usery, Ehlers, and Madden 2006; Ling, Ehlers, Usery, and Madden 2007).

Data	Source	Type	Resolution	Accuracy (from metadata)
Elevation	NED	Raster	30 m	2–10 m
Hydrography	NHD	Vector	1:24,000 source scale	13 m
Orthoimages	133 Urban Areas	Raster	0.33 m	0.33 m
Land Cover	NLCD	Raster	30 m	1 ha
Transportation	Variable	Vector		Variable

Table 1. Data Layers Used in the Empirical Study.

The algorithmic development has followed the work of Chen, Knoblock, and Shahabi (2006) and Chen, Knoblock, and Shahabi (2008) and attempts to force a vector transportation network to fit a corresponding image.

4.0 Empirical Testing

For the five datasets in Table 1, we produced plots of all pair wise combinations at 1:24,000 and 1:12,000 scales. We selected group of four skilled cartographic professionals to judge whether the two datasets were integrated. The goal was not to achieve a statistical test of individuals, but rather to establish the requirements of integration as viewed by cartographic professionals. For statistical analysis, each respondent judged forty locations on a two-quadrangle test area for two different sites, Manchester and Kirkwood in St. Louis, Missouri, and Chamblee and Norcross in Atlanta, Georgia. Cartographic professionals were selected because they possess significant experience in working with geospatial datasets and, at the time, were working on interactive data integration using GIS software to move vector transportation lines to match the image data.

"These ratings were provided for three aspects of integration: position, shape, and temporality."

We used a scale of 1 to 5, where 1 means no correspondence between the two datasets, 3 is moderate correspondence or integration, and 5 is perfectly integrated, meaning no visual discrepancy of position between the two sources. The numbers 2 and 4 provided intermediate values in the scaling (Table 2). This 5-point scale is similar to the Likert scale used in psychometric testing (Trochim 2001). These ratings were provided for three aspects of integration: position, shape, and temporality. Position is a measure of distance separating the same feature on the two sources. Shape assesses the correspondence of shapes but not necessarily directional alignment. Temporality is a judgment of whether the same feature exists on both sources. The respondents were shown examples of overlaid datasets meeting these measurements, including a standard that had been manually edited to force a match to the high-resolution orthoimages and produce a 5 scale value. Table 3 presents a summary of the results of the forty locations from the two test areas. The scores are a composite of the three measured aspects.

"Our preliminary interpretations are that the results generally follow expectations regarding data resolution."

Our preliminary interpretations are that the results generally follow expectations regarding data resolution. Orthoimages with a 0.33 m resolution did well, especially when compared with MODOT vector transportation, which is a high-resolution vector source. The Ortho/TIGER results can be explained by the poor spatial registration due to the small 1:100,000-scale source of the original TIGER data and the generalization

Scale Value	Data Integration Interpretation
1	No correspondence between the two datasets
2	Somewhat integrated
3	Moderate correspondence or integration
4	Highly integrated
5	Perfectly integrated with visual discrepancy of position, shape or existence

Table 2. Rating Scale for Visual Assessment of Data Integration Scale for Position, Shape, and Temporality.

Paired Data Sources*	12K Average Score (1–5)	24K Average Score (1–5)
NLCD – NHD	1.2	1.1
NLCD – MODOT	1.0	Not evaluated
StierLC –NHD	1.2	1.2
Ortho – NHD	2.3	2.8
Ortho – TIGER	1.3	1.5
Ortho – MODOT	3.6	3.6
NED – NHD	1.0	1.3
NED – MODOT	1.0	1.0
StierLC – MODOT	3.0	3.4

*NLCD – National Land Cover Dataset, NHD – National Hydrography Dataset, MODOT – Missouri Department of Transportation, StierLC – High resolution land cover, Ortho – Color Ortho image 1 ft.

Table 3. Summary Results of Visual Interpretation of Integration.

necessary for that scale. The NED, with a 30 m resolution, was hard to visually assess compared with the other data layers, plus it is difficult to determine what to actually use to assess quality of feature registration. In general, the raster-to-raster overlays were not evaluated since there is no obvious basis for visual assessment. For the forty locations on two sites, the standard deviations from the averages in the table were generally 0.5 or less. There were outliers, for example, the Ortho-MODOT integration had a standard deviation of 4.6 for the shape measure at the 1:12,000 scale, but only 0.5 at the 1:24,000 scale. The Ortho-GADOT comparison for geometry at 1:24,000 scale yielded a standard deviation of 4.4, indicating significant variance among the individuals and more specifically among the forty locations on the two-quadrangle area. Other than these outliers, all other comparisons showed standard deviations of less than 1.0 and in most cases less than 0.5.

"For the forty locations on two sites, the standard deviations from the averages in the table were generally 0.5 or less."

The plotting of data at 1:12,000 versus 1:24,000 scale made little difference since the level of integration is dependent on source scale and

resolution and not on display scale. However, at some sufficiently small scale, all data sets will appear integrated since the small scale will override positional discrepancies and line weights will obscure actual lack of integration.

To quantify the meaning of the visual, empirical study, we made measurements of displacements of the roads from the MODOT dataset with respect to the orthoimages. Since these were the "best" data integration from a visual interpretation, it is logical to measure the discrepancies to establish a quantitative basis of what it means to be integrated. Using a sample of 38 points of the largest discrepancies on the test area, an average measurement of 6.2 m was obtained. Note that these are the largest areas of deviation and, at 1:24,000 scale, are well within the National Map Accuracy Standards (NMAS) accuracy specification of 13 m. Thus, apparently within 6 m or so, two data sets portrayed at 1:24,000 scale are perceived to be integrated.

"Using a sample of 38 points of the largest discrepancies on the test area, an average measurement of 6.2 m was obtained."

5.0 Cartographic Basis for a Theory of Data Integration

In order for geospatial datasets to be integrated, a basic compatibility of scale, resolution, and accuracy of spatial position and thematic attribution must exist. The three basic cartographic transformations of Keates (1982) provide a starting point to develop a theory of data integration. The first transformation is map projection which transforms from the sphere (or ellipsoid) to a plane representation. This transformation is mathematically rigorous, deterministic, correctable, and reversible. Thus, if we have two datasets that differ only in projection, they can be integrated through a mathematical transformation. Similarly, the second transformation from three-dimensions to two-dimensions, the planimetric transformation, is also mathematical, deterministic, correctable, and reversible. Again, if two datasets differ only in three-dimensional versus two-dimensional representation, they can be integrated through a mathematical transformation. The last transformation is generalization, which is non-mathematical, scale- dependent, subjective, and not correctable nor reversible. Thus, two datasets generalized at different levels may not be integratable unless they are close enough in scale and resolution to make integration possible. An example of the results of generalization and its intractability for integration is shown in Figure 5. The question remains how close is close enough.

To address that question from a theoretical perspective, we used cartographic theory, particularly, the radical law of Töpfer and Pillewizer (1966) with basic concepts of generalization and abstraction, and developed the following working proposition. If data meet NMAS or the National Standard for Spatial Data Accuracy (NSSDA), then integration can be automated based on the scale ratios as follows:

"The three basic cartographic transformations of Keates (1982) provide a starting point to develop a theory of data integration."

- If linear ratios of scale denominators are >= 0.5, then integration is possible through mathematical transformations (12,000 / 24,000 = 0.5) and adjustments.
- For ratios < 0.5, generalization results in incompatible differences (12,000 / 48,000 = 0.25) and data integration cannot be achieved through transformation, but will require manual/interactive adjustments of spatial data elements.

We use our empirical study with respondents to verify this working postulate, but further have developed an automated procedure, based on the work of Chen, Shahabi, Knoblock (2003) for integrating vector trans-

Figure 5. Example of generalization problem for data integration. The blue line is the generalized stream as represented in the Census TIGER data, which was developed from the USGS 1:100,000-scale topographic map; the red line represents the true stream course without generalization. (see page 61 for color version)

portation with orthographic images when the scale and resolution ratios are in the appropriate range.

6.0 An Approach for Vector and Raster Integration

In trying to expand on the methodology for integrating vector data with orthoimages, the USGS provided a small grant to the University of Southern California (USC) Information Sciences Institute to fund, in part, continuing work on an automated road integration approach. The approach, described in Chen, Shahabi, Knoblock (2003), Chen, Knoblock, Shahabi (2006), and Chen, Shahabi, and Knoblock (2008) requires identifying nodes (intersections) in the vector data, using the nodes to identify candidate locations of intersections in the image data, classification of road and non-road pixels in a buffer around the vector nodes, pattern patching of a vector template based on road widths and intersection angles around the node, elimination of poorly matched points, then computing a transformation between the locations of the vector intersections and the identified locations of the intersections in the image and applying the transformation to the vector data to force a fit with the image positions. This approach can be contrasted to the approach in Wu, Carceroni, Fang, Zelinka, and Kirmse (2007), of Google, Inc., in which the orthographic images are warped to register to the vector data. We chose to transform the vector data to match the images since our image data has much higher resolution and accuracy than the vector data that came from map sources.

"Once the intersections are identified in the images, the conflation techniques described by Saalfeld (1993) are used to match geometry of vector roads and orthoimages."

Once the intersections are identified in the images, the conflation techniques described by Saalfeld (1993) are used to match geometry of vector roads and orthoimages. The USGS research team has replicated this procedure to use in testing integration methods for *The National Map*. The USGS-developed software for testing and feasibility analysis of automated road integration includes the following steps:

1. Locate nodes (intersections) in vector data;
2. Using road width, create a geometrically accurate buffer around nodes and create within each buffer an image template of road segments;
3. Overlay the buffer template onto the original raster images;
4. Perform pattern matching, using correlation analysis, to identify the best match to the template;
5. Repeat steps 3 and 4 for all nodes in the vector data;
6. Based on distance and direction, filter poorly identified intersections;
7. Apply a rubber-sheeting transformation to correct the vector roads to match the image locations (for example, see Saalfeld 1985).

The results of the automated road integration were assessed by both qualitative and quantitative methods. Qualitatively, we compared the output of the automated approach with the ideal result created by manual editing of the vectors to force a match to the high-resolution orthoimages and produce a 5 scale value as was the standard for the empirical analysis above. Figure 6 shows an enlarged portion of this ideal standard. Figure 7 shows a case where the automated procedure improves the alignment for the road vector data. The visual assessment shows that this algorithm improved the alignment in most cases but, unfortunately, there were some cases where the algorithm caused degradation to the alignment.

"In most cases displacements were reduced to less than 1 m."

Quantitatively, measurements of discrepancies between the road vectors and the image positions of the roads were made as with the visual empirical study. In most cases displacements were reduced to less than 1 m. Whereas the MODOT roads are initially of high quality, the application of the automated procedure enhances the positions of the vector roads with respect to the locations of the roads in the images.

The semi-automatic process consists of a manual part and an automated part. The manual processing requires an operator to "train" the image

Figure 6. A vector transportation dataset was manually edited to match the orthoimages. The display of the manually edited data over the orthoimage became the standard against which qualitative evaluations were based. (see page 61 for color version)

Figure 7. MODOT and orthoimage integration after implementation of the automated procedure showing improvement in alignment for integration (red: MODOT; green: automatically processed roads). (see page 62 for color version)

by denoting areas of **roads** and areas of **non-roads**. The automated process consists of two aspects. First, there is the automated processing of the entire image to classify the roads and non-roads based on the input training datasets. Second, there is the automated process of finding the intersections based on the classified roads and then relocating the vector nodes. All of these processes are performed on one image tile.

The manual training of the roads and the automated classification of the images are required to be executed once per project (on a single image tile). The automated find intersection/relocate process must be executed on each image tile in a project.

"Thus, it becomes apparent that this semi-automated process can be a real time saver in integrating vector road data with orthoimages."

To manually correct all roads on an image tile as was done for our ideal standard (see the enlarged portion displayed in Figure 6), took on average approximately 16 man-hours. The manual training of an image tile in our semi-automated process takes an average of approximately 2 man-hours. The automated image classification of an image tile takes an average of approximately 0.25 man-hours. In addition, the automated find intersection/relocate process per image tile takes an average of approximately 0.25 man-hours. Thus, it becomes apparent that this semi-automated process can be a real time saver in integrating vector road data with orthoimages.

For an example, for one image tile the manual process would take 16 hours (1 * 16h); whereas, the semi-automated process would take 2.5 hours (2.0h + 0.25h + 1 * 0.25h)–a savings of 13.5 hours. Further, for a typical 7.5 minute USGS quadrangle, which is comprised of 20 image tiles, the time savings would be greater than 300 man hours (20 * 16 h = 320 h versus 2.0h + 0.25h + 20 * 0.25h = 6.25h). Obviously, the time savings increase exponentially, as would be the case when doing the two study areas of St. Louis and Atlanta. St. Louis (as defined by the 133 Urban Areas project) consists of 50 standard quads yielding a savings in excess of 15,000 hours.

In further qualitative analysis, we examined the type of control point filtering and the magnitude of the filtering on the output results. We looked at plots of a portion of the St. Louis area with 50 percent of the control points filtered using two different methods: a distance filter and a vector median filter. The distance filter eliminates control points identified by the algorithm solely on the difference of the distance, i.e., considering only magnitude, whereas the vector median filter calculates a median vector of all control points and filters those points with the greatest difference between the control point and this vector, thus considering both direction and magnitude. A visual assessment of these plots appears to indicate that the vector median filter is preferable, but at this point this conclusion is tenuous. In addition, we compared the different percentages of points removed for the vector median filter method sequentially between 10 percent and 90 percent incrementing by 10 percent, and found that there is a more noticeable difference between 10 percent and 50 percent of points removed than between 50 percent and 90 percent of points removed.

"The approach documented here provides a design for general vector/raster integration based specifically on integrating vector road data with high-resolution orthoimages."

The approach documented here provides a design for general vector/raster integration based specifically on integrating vector road data with high-resolution orthoimages. The approach can be effectively used with data that are integrated at a level of 3 from the empirical analysis. That is, our results show that MODOT data, with an empirical integration value of 3.6, while reasonably matched to the images in original form, can be improved to produce an acceptable final integrated product. TIGER data, with an empirical integration value of 1.3-1.5, cannot be transformed to produce an acceptable integrated product with the orthographic images. While this design may be able to support a variety of geospatial data and image sources, further testing is required. For example, integrating vector road data with land cover will not work with this design since the road intersections do not appear in the land cover data.

7.0 Towards a Theory of Data Integration

Our project goal is to develop theory that can be used to implement an automatic method to support data integration based on available information about resolution and accuracy in metadata. This development is based on concepts from cartographic theory, known limits of generalization methods, an empirical analysis of viewing scale (Fleming, Jordan, Madden, Usery, and Welch 2005), and an examination of the results from image fusion methods for remotely sensed images of varying resolution (Ling 2006; Ling, Ehlers, Usery, and Madden 2006). Our working proposition is that if scale denominators of source maps for vector data are within a factor of two, then the datasets can be integrated. If the factors are greater than two, then it may be possible to integrate the datasets, but significant processing and human interaction may be involved. For raster data, our working hypothesis is similar, but is based on a resolution ratio of two. This hypothesis is supported by research on raster resolution and map scale equivalents by Tobler (1988). For example, for a map of 1:24,000 scale, the equivalent raster resolution is 12 m, 24,000 divided by 1,000 to determine detectability, then divided by 2 to determine resolution in meters. The hypothesis is also supported by research on viewing scales by Fleming, Jordan, Madden, Usery, and Welch (2007), which provides optimum viewing scales based on the resolution of raster image data. The optimum viewing scale of 1:24,000 corresponds to a raster resolution of 12 m. The hypothesis is contravened by ongoing work by Ling, Usery, Ehlers, and Madden (2007), which shows image fusion of satellite sources can be accomplished at resolution ratios of 1 to 30, and by Lüscher, Burghardt,

"This development is based on concepts from cartographic theory, known limits of generalization methods, an empirical analysis of viewing scale, and an examination of the results from image fusion methods for remotely sensed images of varying resolution."

and Weibel (2007) showing road matching at ratios of 1 to 8. The capability to fuse images of such resolution differences results from the continuous nature of the pixel values and does not hold for mixing other types of data, such as vectors, with images. Those large ratios also introduce artifacts, however, and the exact resolution ratio for true integration and image fusion without artifacts is in the process of being established.

"We have defined the nature of the integration problem and, drawing from cartographic theory, have begun to set limits on the ranges of scales and resolutions of data that may be effectively integrated."

8.0 Conclusions

The integration of the various data layers for the *The National Map* of the USGS is a significant scientific and technical problem. Problems include the basic definition of data integration and the cartographic practices of generalization that prohibit recovery of the original information that could be integrated. We have defined the nature of the integration problem and, drawing from cartographic theory, have begun to set limits on the ranges of scales and resolutions of data that may be effectively integrated. The theory points to limits of a factor of 2 in terms of map scale denominators or resolution ratios that permit effective integration. This practical limit is supported through an empirical response survey, research on viewing scales for image data, and an automated procedure for integrating roads with orthographic images.

REFERENCES

Chen, Ching-Chien, Craig A. Knoblock, and Cyrus Shahabi. 2008. Automatically and accurately conflating raster maps with orthoimagery. *GeoInformatica* 12(3): 377-410.

——, Craig A. Knoblock, and Cyrus Shahabi. 2006. Automatically conflating road vector data with orthoimagery. *GeoInformatica* 20(4): 495-530.

——, Cyrus Shahabi, and Craig A. Knoblock. 2003. *Automatically conflating road vector data with high resolution orthoimagery.* Report to U.S. Geological Survey on Grant No. 03CRSA0631. Los Angeles: University of Southern California.

Devogele, T., C. Parent, and S. Spaccapietra. 1998. On spatial database integration. *International Journal of Geographical Information Science* 12(4): 335-352.

Finn, Michael P., E. Lynn Usery, Michael Starbuck, Bryan Weaver, and Gregory M. Jaromack. 2004. *Integration of The National Map*. Abstract presented at the XXth Congress of the International Society of Photogrammetry and Remote Sensing, Istanbul, Turkey, July. Available Online: <http://carto-research.er.usgs.gov/data_integration/pdf/integrationsISPRS.pdf> (Accessed April 28, 2005)

Fleming, S., T. Jordan, M. Madden, E. L. Usery, and R. Welch, R. 2008. GIS applications for military operations in coastal zones. *ISPRS Journal of Photogrammetry and Remote Sensing ., doi:10.1016/j.isprsjprs.2008.10.004.*

Gösseln, G.V., and M. Sester. 2004. Integration of geoscientific data sets and the German digital map using a matching approach. In *Proceedings of the XXth Congress of the International Society for Photogrammetry and Remote Sensing.* Istanbul, Turkey. Commission IV, XXXV-B4, 1247.

Keates, J.S. 1982. *Understanding maps*, New York: John Wiley and Sons.

Ling, Yangrong. 2006. *Fusion of high-resolution satellite images*. PhD diss. University of Georgia.

——, Manfred Ehlers., E. Lynn Usery, and Marguerite Madden. 2006. FFT-enhanced IHS transform methods for fusing commercial high resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing* 61(6): 381-392.

——, Manfred Ehlers, E. Lynn Usery, and Marguerite Madden. 2007. Effects of spatial resolution ratio in image fusion. *International Journal of Remote Sensing* 29(7): 2157-2167.

Lüscher, P., Burghardt, D. & Weibel, R. 2007. Matching road data of greatly different scales. *Proceedings 23nd International Cartographic Conference*, Moscow, 5-9 August 2007. CD-ROM.

Saalfeld, Alan. 1985. A fast rubber-sheeting transformation using simplicial coordinates. *The American Cartographer* 12(2): 169-173.

——. 1993. *Conflation: Automated map compilation*. Baltimore: University of Maryland Computer Vision Laboratory, Center for Automation Research.

Sheth, Amit P., and James A. Larson. 1990. Federated database systems for managing distributed, heterogeneous, and autonomous databases, *ACM Computing Surveys*, 22(3): 183-236.

Tobler, Waldo. 1987. Resolution, resampling, and all that. In *Building databases for global science,* ed. Helen Mounsey, and Roger Tomlinson, 129-137. London: Taylor and Francis.

Töpfer, F., and Pillewizer, W. (1966). The principles of map selection. *The Cartographic Journal*, 3:10-16.

Trochim, William M. K., and James P. Donnelly. 2006. *The research methods knowledge base.* 3rd Ed. Mason, OH: Atomic Dog Publishing.

U.S. Geological Survey. 2002. *The National Map: Topographic mapping for the 21st century*. Available Online: <http://nationalmap.gov/nmreports.html> (Accessed April 25, 2005)

Vernon, D.E., Jr. 2004. Geospatial technologies in homeland security. *Earth Observation Magazine* 13(1).

Walter, V. and Fritsch, D. 1999. Matching spatial datasets: A statistical approach, *International Journal of Geographical Information Science*, 13(5): 445-473.

Wu, Xiaqing, Rodrigo Carceroni, Hui Fang, Steve Zelinka, and Andrew Kirmse. 2007. Automatic alignment of large-scale aerial rasters to road maps. In *Proceedings of the 15th annual ACM international symposium on advances in geographic information systems.* Seattle, Washington.