

INTEGRATION AND ACCESS OF MULTI-SOURCE VECTOR DATA

Dan Edwards and Justin Simpson

Topographic Engineering Center¹
7701 Telegraph Road, Alexandria, VA
22315 USA

Commission IV, WG IV/2

KEY WORDS: Linking, Fusion, Multi-source, Common Schema, Best Map.

ABSTRACT:

When vector representations of the terrain from different sources are overlaid, the representations do not agree. The disparity can be due to scale, resolution, compilation standards, source accuracy, registration, sensor characteristics, currency, temporality, or errors. Whatever the case, the data do not agree and the user is unsure which data best represents the terrain.

This paper concerns two managerial tools that are useful in integrating and validating data from multiple sources. The first tool, called "linking", integrates the data automatically, while preserving disparities. A second tool, called "Best Map", assembles the data into a unified whole.

Linking is an automated process that forms a correspondence between feature elements in two overlapping vector data sets; these feature elements must correspond spatially and thematically. Linking simply identifies two representations of the same element on the terrain without judging which representation is better. A fundamental tenet of linking is that we do not know the truth.

The Best Map is an application that automatically assembles linked data sets into a unified representation. The Best Map application neither destroys the original data nor the links. The schemas of the two data sets are combined into one Common Schema that makes full use of the separate attribution.

1. INTRODUCTION

People, wishing to integrate vector information, discover a disturbing reality: multi-source data over the same geographical area are disparate, both spatially and thematically. The disparity can be due to scale, resolution, compilation standards, operator license, source accuracy, registration, sensor characteristics, currency, temporality, or errors. Whatever the case, the data do not agree and the user is unsure which data best represents the terrain.

It is reasonable to expect that a vector data set is internally consistent, in the sense that a line is suitably placed with respect to other lines in the data set, given the limitations mentioned above. Normally there is spatial inconsistency between multi-source data sets over the same geographic area, which raises concerns about spatial validity. Thematic validity comes into question when objects, which seem to be the same, have different attribute values or even different attributes, due to differences in the schemas.

Concerns about consistency and validity lead to other questions. To what extent do overlapping data sets represent the same physical entities? What is the true location of a spatial object? What are its attributes? How does one store, manage and integrate multi-source information?

Multi-source data needs special tools to manage, automatically, the disparities that are endemic to such data. This paper concerns two managerial tools that are useful in the integration and validation data from multiple sources. We shall discuss some of the problems that arise and give our automated solution, called "linking", where disparities can be preserved yet integrated. We shall discuss a second tool, called "Best Map" in which linked data can be viewed as a unified whole. Linking is built around a fundamental tenet: we do not know the truth, without further information or source imagery.

Nevertheless, this linking process generates new information when linked data elements reinforce, augment, or contradict each other. Linked data elements represent differing analysts' views of the same entity. When viewed in the context of the collection specifications, independent interpretations can provide a spectrum of information, ranging from mutually reinforcing to contradictory. Linking is a comparison of data and that comparison can strengthen or weaken our confidence in the underlying data, providing a more realistic corroboration in either case. Map information is inherently pluralistic – we do not know the truth, but we can alter our certainty through linking.

2. WHAT IS LINKING?

When vector data from two sources are overlaid, the vector representations of the terrain do not agree. Linking is an

¹ The work described here reflects work over a number of years and across several programs. Intergraph Corporation has written prototype linking software. Intergraph and Swiftsure Spatial Systems Inc. have written prototype best map software. These software developments have helped us test and refine the ideas presented here.

automated process that forms a correspondence between feature elements in two overlapping vector data sets; these feature elements must correspond both spatially and thematically. Linking does not attempt to eliminate differences. Linking simply identifies two representations of the same element on the terrain without judging which representation is better. A fundamental tenet is that we do not know the truth, in the absence of further evidence. See figure 1.

For example, higher resolution data cannot automatically be assumed to be “better data” than lower resolution information. Figures 2 and 3 illustrate this point. Figure 2 contains a collection of features loaded from different sources. In this

example, an urban area is covered by road feature data from two sources, one of a higher resolution and detail than the other. Some of the information available on the 1:50,000 scale source (black) is unique and not available on the 1:250,000 scale source (gray). But the reverse is also true in a nearby region of these same collective data sets (Figure 3). The assumption that one data set is better than the other is obviously incorrect. The discrepancies may be due to different collection specifications, temporal disparity, sensor differences, or obscuration in the source imagery.

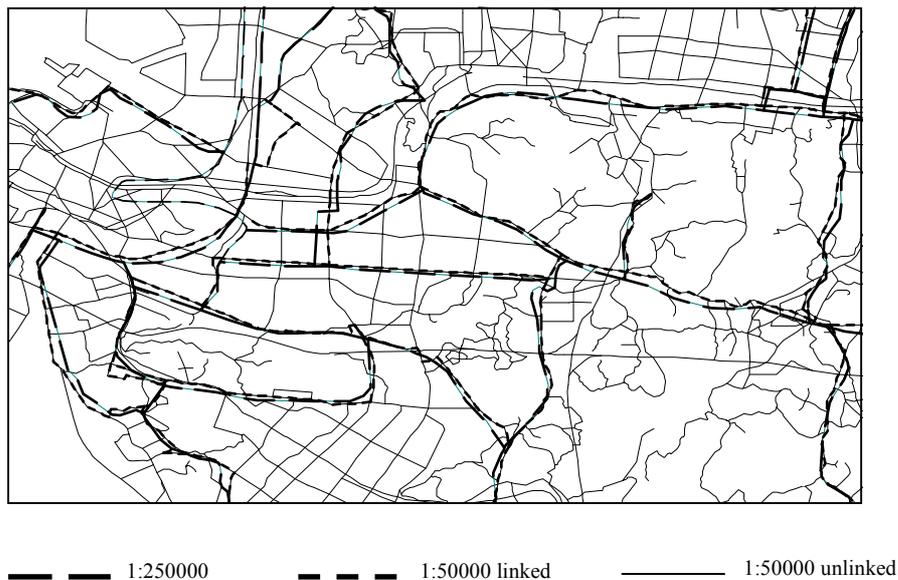


Figure 1

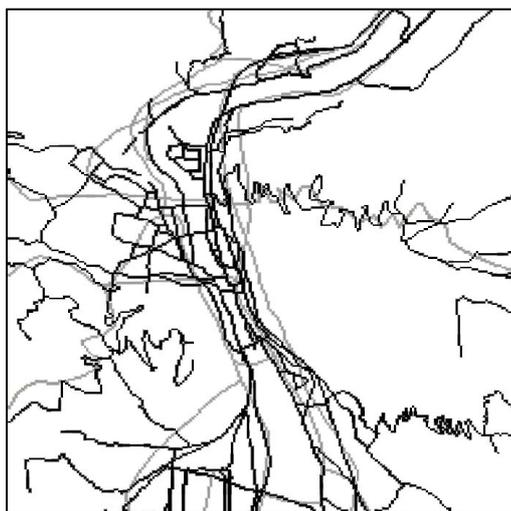


Figure 2

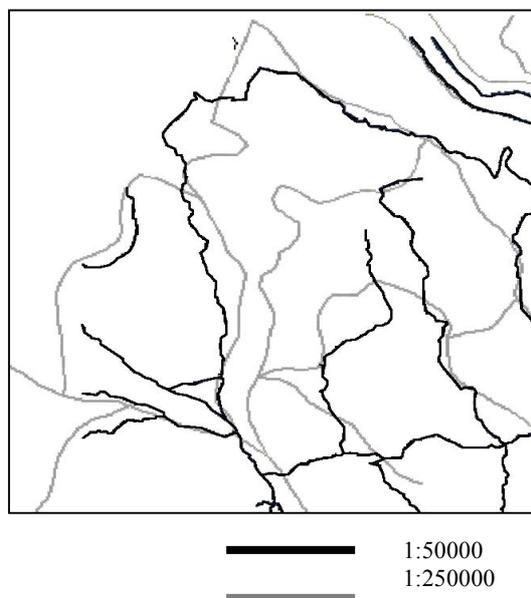


Figure 3

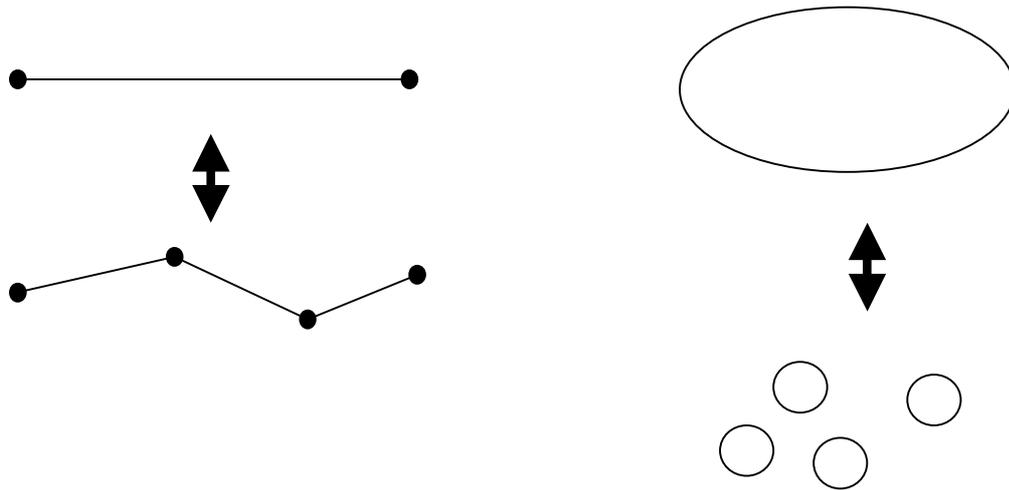


Figure 4. Linking between feature elements

Comprehensive “truth” is not attained from a single source. The linking tool, within a pluralistic system, integrates different descriptions of the same geographic area. Data are stored “as-is”, as received from the provider, preserving separate spatial and semantic detail. From a data perspective, this resulting integrated view has many topographic feature elements. Each feature represents the “best” version of the spatial component available in the pluralistic database. User applications can then be run on these integrated “best” elements.

Feature elements can represent points, linear entities, or areal entities, and the linking correspondence is not constrained by dimensionality. For example, an areal river in one source can be linked to a linear river in another source. A point bridge can be linked to a linear or areal bridge. An areal city can be linked to a point city. The entities in one source that correspond to entities in another source can be single feature elements or groups of feature elements.

Nevertheless, linking generates new information when linked data elements reinforce, augment, or contradict each other. Linked data elements represent two analysts’ views of the same entity. When viewed in the context of the collection specifications, two independent interpretations can provide a spectrum of information, ranging from mutually reinforcing to contradictory. The new information provides a more realistic confidence in the feature elements. (See Section 3.5.)

Although linking is a non-destructive process that preserves the original data sets, linking facilitates the assembly of a unified view of the terrain that represents the intelligent fusion of the disparate information. (See Section 5.) Since any number of data sets can be linked, one can assemble an integrated view of all available data over an area of interest.

3. BENEFITS FROM LINKING

The previous section on the reasons to link provides a background against which we can appreciate the benefits of linking. In this section we shall discuss various benefits that result from linking.

3.1 Potential to Use All Available Information

As discussed above, no single source contains the perfect representation of the terrain for all applications. This pluralistic reality suggests that a goal of mapping is to integrate data from various sources in a way that captures the unique advantages of each source. Commercial data, used outside of the traditional mapping community, may have rich thematic content that can enhance data of a higher spatial fidelity. As the commercial importance of geospatial data increases, all of tomorrow’s data sets, both commercial and governmental, could become a rich collective source of information for the mapping community.

3.2 Automated Use of Future Information

New vector data can be linked to existing data – automatically integrating new vector information with the old. Minor discrepancies are resolved and only significant differences are highlighted so that human intervention is invoked only when problems are indeed major. This integration of new data through linking avoids conflation and its inherent pitfalls. (See Section 3.4.) Moreover, linking greatly reduces the need for human editing and the resulting destruction of information.

3.3 Dealing With an Uncertain Reality – We Do Not Know the Truth

We do not know the truth – at least when comparing disparate maps. We have seen examples where low-resolution data contained roads that were not present in higher resolution data. Lower resolution data can be collected with a richer set of attribution than higher resolution data. When maps differ, massive human intervention is needed to explain the differences, and sometimes even human intervention does not suffice. Even with the aid of source imagery, it can be difficult to resolve differences between maps.

Such issues as scale, resolution, spatial accuracy, currency, and thematic accuracy are important aids to be considered when resolving differences between maps. However, it is risky to decide that one feature is correct while another is incorrect based upon automated criteria.

3.4 Not Destructive

Since we do not know the truth, we are reluctant to destroy one of two conflicting data sets, in favor of the other. Linking preserves the original data sets. Linking simply identifies corresponding terrain elements in two data sets without making a judgment that one element is to be preserved and the other deleted. Linking strengthens our confidence in the underlying data, when they agree; it provides a forum for comparison, when they are similar; and it notifies us when they disagree.

3.5 Creation of Knowledge

Linking yields new knowledge – knowledge that things in different data sets are the same, which reinforces the validity of both. However, linking can result in additional knowledge. Linking data sets can result in a synthesis that is greater than the sum of the component parts.

For example, a set of high-resolution roads that are linked to a low-resolution road is endowed with an additional attribute: that each high-resolution road is a candidate for a major road. Attributes can be transferred from one road to its linked counterpart. As we shall see in the discussion of Best Map, linking two road networks provides additional knowledge of how to integrate the networks – of how to pass from one network to the other.

3.6 Opportunity to Compare Sources – Strength in Numbers

Linking is a comparison of data and that comparison can strengthen or weaken our confidence in the underlying data. Linking is a confirmation that separate analysts agree on a representation of the terrain, increasing our confidence that each representation is correct. When linking indicates that two data sets are contradictory, linking each data set to a third might give support to one of the conflicting views. Multi-source map information is pluralistic – we do not know the truth, but we can alter our certainty through linking.

4. PLURALISM²

Overlapping vector data provides independent views of the same underlying physical reality – views that are sometimes incongruous or sometimes reinforcing. Among these multiple views, we do not know which view, if any, is correct or even best. The nature of multi-source map data is pluralistic.

Linked feature elements are constrained by the linking algorithm to be similar both in space and in a core set of thematic attribute values. In this section, we shall examine how linking can alter our perception of this pluralistic state of affairs, both inside and outside of the core areas of agreement. The following table suggests the kinds of relationships that can arise in linked information.

4.1 Thematic Pluralism

In this section we compare the thematic content of objects that are spatially similar or identical, as established through linking. Before data sets are linked the user defines core sets of

attribution in each data set that must match before features are linked. Since the linking software does not consider attributes outside of this core, these non-core attributes may or may not agree for linked features. For example, “road” should link to “highway” but other attribution may agree, disagree, or provide additional and unique information. We shall examine the relationships that can exist between linked features and their attributes whose values were not constrained by linking.

4.1.1 Thematic Identity

Linked features are thematically identical when their feature codes and a set of attribute values are identical. Thematic identity provides strong corroboration that the attribution is valid.

It is useful to recall a difference between thematic and spatial agreement. One would expect spatial differences between linked features to be small, with the discrepancies being due to scale, resolution, generalization, registration and the like. The technology to derive accurate spatial information is mature. In contrast, supplying attributes to objects in imagery is an imprecise craft, making identical thematic agreement a happy event.

4.1.2 Thematic Similarity

Terrain elements are linked when they are spatially similar and are similar in a core set of attribute values. In this section we are considering linked terrain elements that have thematic similarity outside of this core set.

Thematic similarity increases confidence when, for example, a road, whose transportation use is “road” in the schema of Source 1, links to a road, whose transportation use is “highway” in the schema of Source 2.

Comparing similar attributes within a schema to similar attributes in different schema can increase or decrease uncertainty, depending upon the context. Translation between schemas can be so imprecise that a meaningful comparison of features is difficult. For example, a vegetation feature in Source 1 can have the attribute “without trees” while the same feature in Source 2 could be labeled “paddy”. These two descriptions are hardly an inspiring confirmation; possibly, the attributes are as close as the two specifications allow. This is a case in which one feature could inherit the attribution of the other, in the sense that the feature without trees is possibly a paddy and that the paddy is possibly without trees.

An example of decreased certainty arises when a forest with predominant tree height of 10-15 meters is linked to a forest with predominant tree height of 15-20 meters. In the absence of other evidence, such as temporality or accuracy, the certainty or each attribute decreases while we become more certain that the predominant height is between 10 and 20 meters.

4.1.3 Thematic Augmentation

If linked features have some identical or similar attribute values, then the unique attributes of one can be transferred as possible attributes of the other, making the fullest use of information from both sources. This is particularly useful when one source contains much richer thematic detail than the other.

² We thank Ms. Gail Kucera for helping us develop our ideas on pluralism and Best Map. Ms. Kucera suggested using “pluralism” in this context.

Table 1: Relationships in Information Pluralism

	Identity	Similarity	Augmentation	Contradiction
Spatial Pluralism	Identical spatial coordinates	Equivalent spatial representation (linked)	Additional spatial detail	Conflicting spatial information
Thematic Pluralism	Identical semantics (feature code & attributes)	Equivalent semantic representation (Thesaurus)	Additional semantic detail	Semantic disagreement

Linking also provides the opportunity to augment the data with new knowledge – knowledge not explicitly present in either source. For example, if a 1:50000 road is linked to a 1:10000 road, then the higher resolution road is endowed with the additional attribute that is possibly a major road. Using this new information is discussed below in Section 6, Best Path.

4.1.4 Thematic Contradiction

The eyebrow of uncertainty is raised when linked features have attributes that are contradictory. Linking has identified a conflict that needs resolution – either by a human or by another linked source. In practice, thematic contradiction uncovered by linking has proved to be an important tool in automated error detection. If timely resolution is impossible at least the user is given a warning to use the data with care.

4.2 Spatial Pluralism

Linked objects are constrained by the linking algorithm to be spatially similar. Moreover, in the basic model, we expect that a base feature is linked to secondary feature elements. Nevertheless, comparing the spatial content of linked objects can provide new information.

4.2.1 Spatial identity

Given the difficulty that a single operator experiences in placing the cursor twice on the same location, it is unlikely that two feature elements have identical vertices. If two feature elements have identical vertices, the coincidence would be so striking to suggest that the feature elements have a common digital origin. If all the vertices of one feature element are present in a second linked feature element then there is a strong possibility that the first feature element is a generalization of the second.

Spatial identity, coupled with agreement of core attributes is a strong indication that the feature elements represent the same entity and have a common genesis.

4.2.2 Spatial similarity

This is the case where objects are spatially similar while agreeing on a core set of thematic attributes. This is the sense

of agreement that the linking software is designed to uncover. To say that linked objects are spatially similar is simply to say that they are linked.

The linking of two objects provides a confirmation that the positioning and core attributes are correct, within the constraints of both the product accuracies and of their respective schemas. In addition, new information can arise from attribute transfer where attributes from one object are transferred to the other.

4.2.3 Spatial augmentation

Transferring spatial information from one source to another can violate the internal consistency of the recipient data set. Ideally, spatial transfer of data is best done manually using editing tools in a photogrammetric environment. However, such spatial transfer represents a temptation that linking is designed to diminish, not increase. We link so that we do not have to manually adjudicate differences between data sets.

While spatial augmentation has its perils, it can have benefits as well. For example, consider a low-resolution road network that is linked to a high-resolution road network. The linked high-resolution roads represent a view of the low-resolution roads but at a higher resolution. The linked high-resolution roads can be generalized to obtain another view of the low-resolution network. It is quite possible that the generalized roads violate the internal consistency of the high-resolution data.

4.2.4 Spatial contradiction

A spatial contradiction arises when a base feature element can not be linked. This is a strong indication that something is wrong, either in our expectation, in the data, or in the linking algorithm. Human intervention is needed to resolve this contradiction – possibly with the aid of additional information or imagery.

Figures 5 and 6 illustrate spatial contradictions that were uncovered by linking. In Figure 5, the gray high-resolution roads should link to low-resolution roads, but linking fails in the center area due to a gap in the high-resolution data. Figure 6 shows that the gap in the high-resolution data is a bridge.



Linked low-resolution
 High-resolution

Figure 5



Unlinked low-resolution
 Bridge

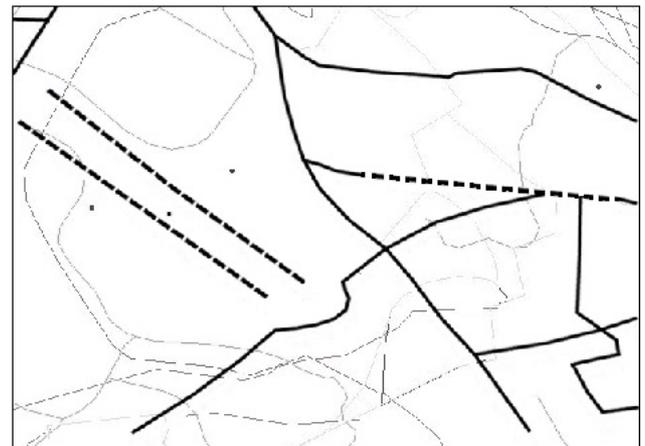
Figure 6

Roads in the low-resolution data are represented as both roads and bridges in the high-resolution data. Linking roads to roads produces a linking failure, while linking roads to roads plus bridges succeeds.

The unlinked low-resolution road at the top of Figures 5 and 6 is another example of a spatial contradiction. It seems that this road should have linked to a high-resolution road. Perhaps this failure is due to a topological problem in the high-resolution data, or to the features being too far apart, or to a flaw in the linking software.

The high-resolution roads exhibit curious behavior near the center of Figure 6, where the unlinked low-resolution road (black dashed) joins the linked low-resolution road (black). Which resolution, high or low, better represents reality? Without further information or source data, we do not know.

Figure 7 illustrates a similar problem. The road network contains a gap because a road has been mislabeled as a runway. Here linking would fail since one does not expect that roads in one source will be represented as runways in a second source that is shown above.



Road Runway

Figure 7.

5. BEST MAP

Linking establishes a correspondence between feature elements across two data sets. The Best Map is an application that enables one to view linked data sets, assembling them in an intelligent manner into a unified representation, capable of supporting digital spatial analysis. As we shall see, the Best Map application destroys neither the original data nor the links. The output of the Best Map software can be either stored or deleted, as the user wishes, without affecting the integrity of the source data.

5.1 Current implementation of linear Best Map

Since the larger issues are still under investigation, this discussion of Best Map centers on software developed independently by Swiftsure and Intergraph, under the direction of TEC. These applications are designed for the case of a low-resolution road network linking to a high-resolution road network. We are still investigating the representation of area, point, or other linear data sets. The rules of similar-scale Best Map are still under investigation, as well.

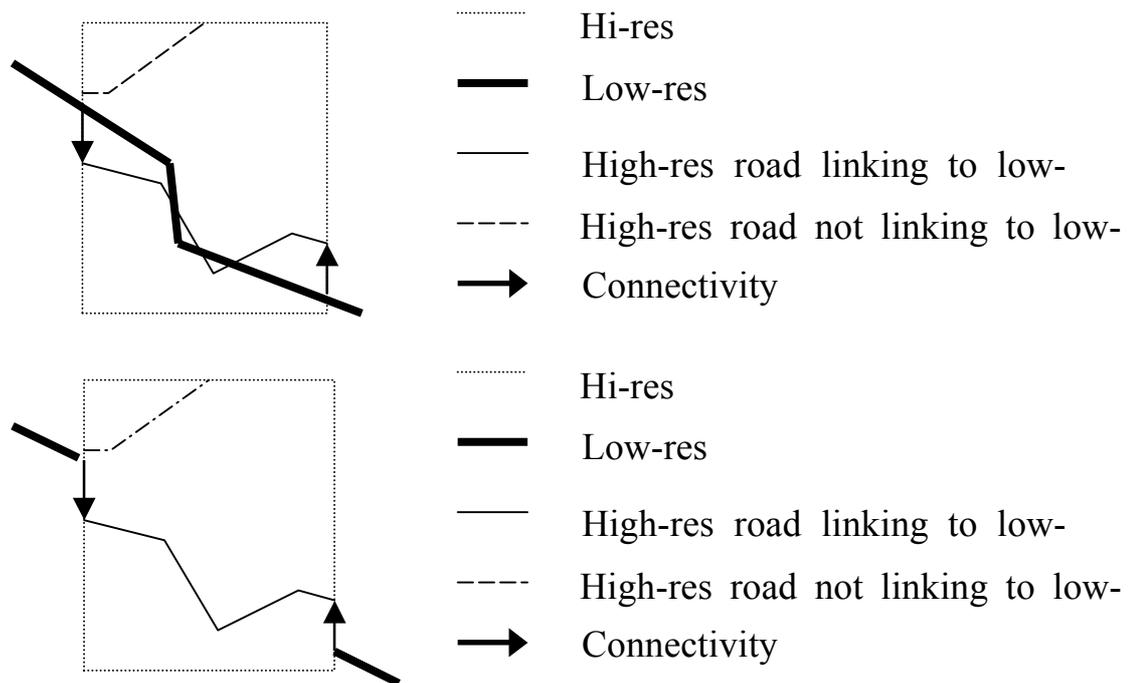


Figure 8 Construction of Best Map

5.1.1 Spatial rules

In constructing the best representation of the data this software uses spatial and thematic rules. The current spatial rules for creating the Best Map are summarized:

- 1 in high-resolution areas, the high-resolution data take precedence
- 2 in outside high-resolution areas, the low-resolution data are the best
- 3 place connectivity vectors at the boundary
- 4 retain unlinked low-resolution features that penetrate the high-resolution area.

Include all high-resolution roads in the Best Map. Outside the area of high-resolution, the low-resolution data are best, by default. Along the boundary between these data sets, place connectivity vectors to join a low-resolution road to the high-resolution road to which it links. These connectivity vectors join base and secondary roads that are linked in the interior of the high-resolution area – not simply joining roads that are closest at the secondary boundary. This ensures connectivity of major roads. Linking is based upon the behavior of the base and secondary roads within the secondary area, rather than upon the nearest neighbor at the boundary. See Figure 8.

5.1.2 Thematic rules – Common Schema Formation

Quite possibly, no one schema will encompass the attributes of two linked data sets. Mapping these schemas into an existing schema usually results in a loss of information. Creating a new schema results in non-standard information. The schema of the Best Map is, notionally, the union of the schemas of the source data sets, called the common schema. This non-standard schema is unique to the Best Map, but it has the advantage that

no information is lost. This is based on the principles, described above, of thematic similarity and augmentation.

In forming the common schema, a base feature class serves as a core feature class in the common schema. The common schema feature class has all the attributes of the core feature class plus the attributes of all the secondary features classes that are permitted to link to the core feature class. Common attributes, if any, among these core and secondary features classes are used only once in the common schema.

In addition, there are new attributes in the common schema feature class. There is a new attribute indicating that the Best Map feature coincides spatially with a secondary feature element that is linked to a base feature element. Another new attribute indicates that a low-resolution road is unlinked. The connectivity vectors form a feature class in the common schema.

6 BEST PATH

Best Path is an example of a practical application of linking and Best Map. A person might want to travel from a rural area, mapped at a low resolution, to an urban area, mapped at a high resolution. These maps can be linked and a unified Best Map created, enabling a Best Path to be calculated.

The first step is to obtain one continuous, unifying map that combines differing scales, differing attributes and differing spatial representations. This was discussed in the section on Best Map. The second step is to traverse the Best Map, making use of its unique qualities. This is the Best Path application whose first version was developed by Dr. Dianne Richardson of the Canada Centre for Remote Sensing, under a contract with Swiftsure Spatial Systems Inc. Helonics Inc. and Intergraph developed later versions of this software.

As with many applications that traverse networks, after the user selects starting and ending points, the software computes a Best Path between them based upon such things as distance, geometry, and the attributes of the roads. This software is able to give preference to the high-resolution roads that link to low-resolution roads by using the new attribute of “linked” that is added by the Best Map application. Consequently, the software can be biased toward main roads in high-resolution areas as preferred avenues of entry or exit.

Best Path is one example of the many types of spatial analysis that can be performed on a multi-source Best Map. Since connectivity and spatial relationships have been preserved during the formulation of Best Map, any processing tool of spatial analysis can operate on both multi-source data and single source data.

7. SUMMARY

Sources of spatial information – whether sensors, maps, or other – vary widely. The spatial information itself differs in scale, accuracy, and representation. There are varied uses and applications of spatial information. All of these inevitably call for multiple data sources, rather than a monolithic database.

Multi-source data are disparate. We believe that this disparity need not increase uncertainty. Rather, multi-source data gives us the opportunity to increase our confidence by allowing pluralistic representations to become mutually reinforcing, within the context of their respective collection specifications. Differences, which cannot be resolved, can identify possible errors.

We believe that the solution to maintaining a national core geospatial database is to exploit the similarity, duplication, and uniqueness in multiple data sets in order to augment current data and to derive new information. Research in feature linking involves developing a set of tools that enable these capabilities. Research in Best Map involves developing a set of tools that can quickly extract the best data from multiple data sets and assemble this into a coherent picture for a given application and for a given instant in time. To date, our results are exploratory and tentative. Further research is needed.

ACKNOWLEDGEMENTS

Intergraph Corporation has written linking software. Intergraph and a team of contractors, headed by Swiftsure Spatial Systems Inc., have written Best Map software, under the direction of TEC. These software developments have helped us test and refine the ideas presented here. The authors wish to thank the contractors and their colleagues at TEC for the many lively conversations that have contributed to this work.