The Fallacy of the "Golden Feature" in MRDBs: Data Modeling Versus Integrating New Anchor Data

Barbara P. Buttenfield[1], Charlie Frye[2]

[1]University of Colorado, UCB-260, Boulder CO 80305-0260 USA
babs@colorado.edu
[2]ESRI, 380 New York Street, Redlands CA 92373-8100 USA
cfrye@esri.com

KEYWORDS: MRDBs, scale-changing, multiple representations, cartographic base maps, data modeling

1. Introduction

> *Whether it is for the business of producing paper maps for sale, or whether it is for displaying maps on a screen to visualize the result of a query, we need computer systems that know how to represent the same geographical area at different scales.* Spaccapietra et al 2000: 57

The demand for operable Multi-Resolution Databases (MRDBs) has grown to the point of wide acceptance in most national mapping agencies (NMAs), and in many local and state government organizations concerned with modeling and mapping cartographic data at different scales. The procedures that support MRDBs involve acquiring topographic data, image data, and base-cartographic vector data at a very fine spatial resolution. The initial dataset forms the basis for deriving coarser resolution versions of the dataset, through data modeling (i.e., generalization and other geoprocessing operations). A large body of work has been published (largely but not exclusively by European researchers) describing obstacles to, and solutions for, automating various aspects of the generalization required for MRDB derivation. That work is well known among the participants of this ICA Generalization Commission workshop and will not be reviewed here.

A longstanding and widely accepted assumption (e.g., Muller et al 1995, Weibel and Dutton, 1999) of MRDB data modeling is that the ideal solution is to compile geometry information at the most precise resolution, and to derive versions at less precise resolutions. This approach works well up to a point, however the premise is flawed in cartographic practice. It does not hold up that an ultimately fine-resolution version of a cartographic feature can generate multiple representations across all scales and for all purposes, through item-level reduction or exaggeration, generalization, symbolization or some combination of these operations. We refer to this premise of a finest-resolution data version serving all scales and all purposes as the fallacy of the "golden feature".

Numerous reasons demonstrate that the "golden feature" concept is not always workable in practice. These reasons relate to discrepancies between map representations, database representations, and reality; the difficulty of compiling data that can support all scales and purposes; intransitivities in object and attribute semantic hierarchies; and the challenge of preserving geographic process in a generalized representation. The paper will discuss each argument in turn. We argue in favor of an alternate approach in which independently compiled data is introduced at intermediate resolutions to 'fill in the gaps' when mapping at multiple scales and map purposes. We call these *anchor data* because their introduction re-aligns the scale-changing process with the complex progression of feature geometry, content, and prominence in the landscape. A similar approach is used by Swiss cartographers (notably Cecconi et al, 2002) to derive intermediate data sets from a single compilation and avoid intensive computations, for example, in on-demand mapping.

Our project focuses on data modeling for topographic and reference base map cartography, and our goal is not only to derive representations from a single existing compilation, but to introduce new, independently compiled data into the MRDB. We acknowledge that extreme challenges accompany this approach, for example in maintaining efficient workflows, in establishing links between multiple representations, and in protecting data semantics and validity. Our work has not matured to the point of presenting a comprehensive solution, however early results of exploring the mechanics of scale-changing indicate that this approach bears further investigation. We outline our early results, and outline possible criteria for selecting resolutions for introducing anchor data sets to an MRDB.

2. The Fallacy of the Golden Feature

As described above, the "golden feature" concept is not workable in practice, for several reasons that complicate cartographic work practice. We highlight several of these reasons below.

2.1 Discrepancies between representation and reality

With the exception of settlement features (e.g., buildings) and selected transportation categories (e.g., dams, highway ramps, canals), the things that constitute database features don't exist as such in the landscape; rather they are defined in the context of a particular measurement framework (Morehouse, 1995). Measurement frameworks in turn are determined by measurement scale (units of measure), measurement granularity (minimum mapping unit) (http://en.mimi.hu/gis/mapping_unit.html) or detectable resolution (Tobler, 1987). For example, a simple feature such as a river channel is collected at 5 meter resolution as set of polygons, extending from river bank to bank at annual flood stage. Coarser resolution versions might be derived via medial axis transform or another type of centerline delineation. Cartographic problems can occur when polygons overlap. For example old oxbows might be overrun by main channels after a flood event, but both features must be maintained in the MRDB. The features themselves do not overlap in the long run (main channels will return to their original course within a few seasons), but the cartographic versions might do so. Polygon overlaps in the database will create special problems as scale changes.

2.2 Compiling a single feature representation for all mapping scales and purposes

Complex features are defined variously according to mapping purpose or measurement task. As a consequence, multiple representations can be difficult to derive from a single source. The classic example is a city represented in the MRDB as a convex hull marking its administrative boundary enclosing object networks of streets and object clusters of landmarks (fine resolution); as a mosaic demarcating incorporated and urbanized areas (medium resolution), or as a single coordinate pair marking a center of population or a network node (very coarse resolution). It would be difficult to obtain all possible representations from any single version, or to determine automatically how to link features for inclusion at multiple scales. In practice, it would present a computational nightmare to aggregate the urban complex from the sum of its finest resolution parts.

In some cases, it is impossible to derive alternative versions from an originally compiled representation, even at a single resolution. For example, a database of addresses maintained by a national postal service locates the geographic position of mailboxes, while the database of addresses required by a national emergency response service locates the position of residential entrances (front doors). In urban America the two are proximate; but in rural America, mailbox positions may differ by as much as 1 km or more from their respective houses. Moreover, it is impossible to infer either location from the other, even with access to orthophotos and vector street networks. Census demography databases such as the US Census Bureau's TIGER files do not store addresses as points at all. Instead, they associate address ranges with line features (street segments); and address matching is accomplished by interpolation. Targeted marketing databases such as created by grocery stores and department stores do not store addresses as points or lines, but as polygons based roughly

on 5- or 9-digit postal (zip) codes. The consequence is that a single, unified national address database is not maintained.

In specific applications, generic object definitions are difficult to model consistently. The archetypical example for USA geospatial databases is wetlands delineation, which is maintained in three agency versions (US Fish & Wildlife Service, Natural Resources Conservation Service, and National Wetlands Inventory) using the same three datasets (hydrography, vegetation and soils) but due to agency missions, with three separate data modeling protocols, three discrepant outcomes, and many (many!) consequent lawsuits.

2.3 Intransitive feature or attribute hierarchies

Most practicing cartographers understand that modeling data representations at coarser resolution is often problematic. Many spatial data sets are modeled hierarchically, in part because of administrative fiat (e.g., national, provincial/state, county/canton, etc.), but also because hierarchies form an organizational strategy that is intuitively straightforward to create (Lakoff, date) and to manipulate (Codd, 1970). However, a well-known limitation of hierarchical data structures is that searching and sorting is most efficient from parent to child nodes, and least efficient moving between parents. Moreover, object categories and attribute hierarchies are often intransitive in practice. For example, in the US Census Bureau demographic hierarchy (state-county-city-census tract- block group), cities are considered child nodes of counties. An intransitivity occurs in New York City, which is a parent to five counties. New York city contains the Five Boroughs (counties) of Queens, Brooklyn, Bronx, Manhattan, and Staten Island. Because of intransitivities, the attribute hierarchy cannot be modeled (generalized) automatically unless special cases can be anticipated.

2.4 Preserving geographic process

Mark (1991) argues that cartographic generalization is intended to preserve visual evidence of geographic process. A long thread of literature outside the fabric of cartography illuminates how geographic processes tend to become evident within specific ranges of scale (Steinhous, 1960, Perkal, 1966, Carpenter, 1981, Morrison and Morrison 1994). From a computational standpoint, it may be too intensive to tease out evidence of a process that is evident only at continental resolutions (e.g., isostatic rebound) from data captured at very local resolution (e.g., erosion). Furthermore, data sensitivities to scale are theme-based. Mark (1991) argues that across a range of map scales, the details of natural features (e.g., terrain data and hydrography) will change more often than for cultural data (roads, land cover, urban footprints). “More often” means that a larger number of critical breakpoints (Muller, 1991) or thresholds can be defined along the progression of scale where changes must be made either to symbols or to geometry (or to both). In part, this is due to the nature of cultural features. For example, roads are built to a fixed radius of curvature, based on the turning radius of automobiles, dictating a simpler geometry that prompts modification with scale change. However, it is also the case that certain naturally occurring surface processes (such as vegetation) will change less often across scales simply because they are dependent on, and surficial to, underlying coarser resolution processes (e.g, soil type and moisture, water table depth, or local climate). The resolution of the causative processes essentially determines the resolution at which changes in the dependent process become evident.

2.5 In what mapping situations does the golden feature model appear to work?

We note that in some mapping situations, it is possible to develop a single fine-grain data representation that applies across a wide range of mapping scales. For example in labeling diffuse regions such as canyons, oceans, physiographic areas or cultural regions, one cartographic method for automating label placement involves creating annotation polygons which are loosely coupled to the canyon or physiographic region (Frye, 2006) and then fitting a label within the polygon. This appears to mimic manual label placement (Imhof, 1975). Moreover, since the polygon shapes are

only loosely coupled to the physiographic shapes at a fine resolution, their shapes can be essentially "blown down" without further adjustment to permit labeling at finer resolutions. An example of this will be shown at the presentation.

Another mapping situation where a single fine-resolution 'golden feature' model may be operable occurs in data modeling cultural features whose shape is archetypal (buildings of a simple rectangular shape) or transportation features (on- and off-ramps). Essentially the archetype can be established and stored at a fine resolution. At smaller resolutions, the proportionately smaller archetype must be evaluated for spatial conflicts, and this problem has been investigated by cartographers at IGN and at Laval and will not be reviewed further here.

A third situation occurs when modeling a simple feature whose geometry collapses with scale change. Simple features include isolated points lines or polygons, such as hydrographic dams. Dams may be stored at fine resolutions as polygons and at coarser resolution as lines. In this case, one compiles two versions of the feature at the finest resolution and sets a layer query to select one representation or the other depending on map scale or purpose. The golden feature model works so long as the feature can be represented at all mapping scales as a simple object. In contrast, this approach won't work as well for a compound feature such as an industrial compound, urban area, or other feature that is represented as an object complex at fine resolution but as a simple object at coarser resolution.

2.6 The fallacy of the golden feature – so what?

It's understandable to ask why these problems cannot be treated as exceptions, and retain the Golden Feature approach as a data modeling norm. In response, we argue first that creating unique representations of every feature at every resolution and then setting layer queries will inevitably create a database that is so complex that feature representations will become ambiguous and difficult to query. We argue second that the reasons given in this section are likely not a comprehensive set, and it's not possible to predict all of the special cases (the exceptions) that might arise in practice. When deriving coarser resolution representations, versions will eventually (and perhaps quickly) come a point where additional newly compiled data must be introduced to the MRDB in order to re-align the database representation with reality, and to extend or push the MRDB beyond a limited range of usable resolution. We refer to any originally or newly compiled data as an "anchor database", and to the corresponding granularity as an "anchor resolution". Cecconi et al (2002) argue that creating a database at intermediate resolution can reduce workflow complexity by establishing outcomes for labor-intensive generalization tasks *a priori*, and we agree with their point. To be efficient, one must balance the effort to build and maintain intermediate data layers with the effort saved in subsequent mapping tasks. Extending their argument, we maintain that the creation of any database can be costly in terms of labor, skill, and/or computation; thus it is advantageous to minimize the number of anchor resolutions at which new data are called for.

Thus the management decision for many mapping agencies and organizations that produce and/or maintain base mapping data is to establish a balance between data flexibility and data cost. This decision has important financial impacts on smaller firms and local government agencies such as city planning offices. Three questions emerge from this line of thinking and these drive our work:

1) How to determine the usable limits of resolution for an initial compilation of a given feature in a given MRDB?
2) How to determine how many new anchor data layers can be incorporated to the MRDB? and
3) At what resolution(s) should this happen?

These are not small questions, and we do not presume to answer them in a single paper. Instead, we overview our recent and current efforts to explore scale-changing for base cartographic data modeling and mapping, and to propose a set of criteria for further exploration.

2 Existing approaches

The accepted approach to multi-resolution database creation in practice at most NMAs is based on the principle of compiling data at standardized resolutions based on product specifications and varying agency missions. NMAs adopt one of two approaches. In the first, a fine-resolution compilation of a Digital Landscape Model (DLM) is supported with image data and fine-grain vector data capture. The DLM forms a source for deriving coarser resolution databases (DCMs) that can be used for mapping at smaller scales. Examples of common national standard mapping scales are 1:5k, 1:10k, 1:25k, 1:100k, and 1:250k. The DLM-DCM approach is widely adopted in Europe, and well summarized by Kilpelainen (1997) and Meng (1997). The second approach involves independent compilation at each standard capture scale, leading to databases created and maintained in relative isolation. The USGS standard topographic series, mapped at 1:24k, 1:100k and 1:2 million forms a prime example of this approach, as summarized in Thompson (1988). Each approach brings its own challenges, and these are acknowledged widely in the community already.

Other approaches include Intergraph's efforts to automate map generalization in the 1990's, and LaserScan's projects including the LAMPS project and the AGENT project, both of which are well-documented in cartographic literature. Individual research labs at several universities have made excellent progress as well on conflict detection and resolution, shape generalization, preservation of topology during feature simplification, and other important problems. It is not our intention to downplay these efforts, but it is nonetheless outside the scope of this paper to review comprehensively all the advances in generalization and development of MRDBs. An excellent review by Meng (1997) can and should be updated, since many important advances have occurred in the recent decade.

Our project team forms a collaboration intersecting design cartography, data modeling, and cartographic database creation. Initiated in 2003 to identify and investigate impediments to multi-scale, multi-purpose base mapping in a relational database environment, we have explored several aspects of data modeling and map symbolization for cartographic base mapping at multiple scales. Early work (Buckley, 2005) specialized the existing ESRI geodatabase model to incorporate Valid Value Tables (VVTs) that established multiple representations by assigning specific feature codes to unique feature descriptions. VVTs were first applied to a statewide database for Texas (TNRIS). This early method was refined into Cartographic Feature Tables (CFTs) with more robust feature descriptions and a DLG database for Southern California was constructed at 1:25k and 1:2 million.

Brewer used this database to determine the range of scales across which symbol change alone could produce satisfactory base maps (Brewer and Buttenfield, 2006). Brewer's experiment indicated a cartographic breakpoint in the scale progression, between 1:250k and 1:700k, where symbology alone could not support multi-scale mapping. Buttenfield and Hultgren (2005) performed a semantic inlay at 1:250k of another (VMAP) database, mapping the VMAP thesaurus into DLG semantics. They discovered schema discrepancies must be handled individually in three situations: in the case of new feature geometries (collapse of dimensions or object complexes); when new feature codes emerged at a specific mapping scale or due to agency mission; and in the case of inconsistencies in feature or attribute hierarchies.

Frye (2006) argues for creation of 'informed' databases to alleviate many of the problems encountered in the schema experiments and other troublesome data modeling situations. "Today the best approach … is not to get bogged down in finding ways to automatically derive representations, but rather first start at a high level and understand the basic representational requirements that maps have and use that information

to drive an informed data capture methodology to produce or evolve data that are tailored to, or fit for use in efficiently producing those maps." *Informed* data capture or data modeling means that an organization's set of mapping requirements should be analyzed collectively and that these requirements should drive the data compilation process.

3. Possible criteria for choosing new anchor resolutions to be discussed:

Clearly, database creation takes time and effort, and as a result makes cartographic data production more expensive and more difficult. The obvious strategy is to minimize the number of original compilations needed for an organization's data products. Just as clearly, a single compilation cannot be fully adequate for all mapping scales and purposes and further, some data layers (feature categories) will be more sensitive to scale change than others. The realistic balance mandates creation of anchor databases for some but not all feature categories in some but not all database layers. To reiterate the questions posed at the outset of this paper, then, we pose three questions:

1. How to determine the usable limits of resolution for an initial fine-resolution compilation of a given feature in a given MRDB?
2. How to determine how many new anchor data layers must be incorporated to the MRDB? and
3. At what resolution(s) should new data be incorporated?

The first question must be addressed empirically, by mapping for multiple scales and multiple purposes by systematic experiment. Brewer and Buttenfield (2006) demonstrate one set of experiments working only with symbol change, for example.

We propose specific criteria to address the second and third questions forms the core of our presentation at this workshop and will present examples for as many as time permits. The criteria for selecting new anchor data and anchor resolutions for integrating these into an MRDB may be based on one of the following criteria:

To minimize uncertainty (e.g., when data no longer meet thresholds for preserving positional accuracy, or when logical consistency fails);

To minimize the amount of required manual intervention (e.g., to resolve spatial conflicts or problems of graphical context;

To minimize the work involved for integrating layers from independent sources (when two layers must conflate, it is often easier to generalize and then re-derive the co-dependent layer than to introduce new data and re-integrate with the independent layer. The example involves terrain and hydrography – it's often better to resample the DEM and re-construct hydrographic channels than to attempt to conflate independently generalized versions of the two layers; or

To minimize extensive computations (this is largely handled by LoDs (Cecconi et al 2002) and will not be covered in the presentation).

4. Summary

The presentation will demonstrate these criteria in practice, along with respective outcomes and possible pitfalls. Discussion will inquire as to relative benefits of each criterion. The research is in early stages, following experimentation with several aspects of implementing MRDBs for local (Ada County Idaho), state (Texas Natural Resouces) and federal (USGS DLG) agency databases. Long-range goals of this work are to generate guidelines for determining how many new anchors are required, and at what resolutions, and to investigate the interaction of symbol change with generalization in attempting to simplify cartographic workflows overall.

5. Acknowledgements

This research forms a part of the Multi-Scale, Multi-Purpose, Base Mapping Project, a collaboration between ESRI, University of Colorado and Penn State University. Funding by ESRI (2003B6345 and 2003B6347) is gratefully acknowledged. We also acknowledge database editing and data modeling by Kiyoshi Yamashita, graduate student and U. Colorado-Boulder, that supports many of the working examples given here. This paper reports an early version of ideas, many of which are still in formulation; comments and critique from Cindy Brewer, Aileen Buckley, and workshop reviewers have strengthened the work overall.

References

Buckely, A. 2004 Using Valid Value Tables in Geodatabase Design. *Cartographic Perspectives **48 Spring: 57-61.***

Brewer, C.A. and **B.P. Buttenfield** 2006 Mastering Map Scale: Formalizing Guidelines for Multi-Scale Map Design. *Proceedings AUTO-CARTO 2006, Vancouver Washington* (forthcoming), June 26-28, 2006.

Buttenfield, B.P. and **T. Hultgren** 2005, Managing Multiple Representations of "Base Carto" Features: A Data Modeling Approach. *Proceedings, AUTO-CARTO 2005*, Las Vegas, Nevada.

Carpenter, L. C. 1980 Computer Rendering of Fractal Curves and Surfaces. *Computer Graphics* July 1980: 9-15.

Cecconi, A., R. Weibel, and **M. Barrault**, 2002, Improving Automated Generalization for On-Demand Web Mapping by Multiscale Databases, *Proceedings of the Symposium on Geospatial Theory, Processing and Applications*, ISPRS/IGU/CIG, Ottawa, Canada.

Codd, E. F. 1970. A Relational Model of Data for Large Shared Data Banks. *Communications of the Association for Computing Machinery*, 13(6), 377-387.

Frye, C. 2006 A Geodatabase Design for a Multi-Purpose Multi-Scale GIS Base Map. *Proceedings,* AUTO-CARTO 2006. Vancouver Washington (forthcoming), June 26-28, 2006.

Imhof, E. 1975 Positioning Names on Maps. *The American Cartographer* **2**(2): 128-144.

Kilpelainen, T. 1997, *Mutliple Representation and Generalization of Geo-Databases for Topographic Maps*. Ph.D. Dissertation and Technical Publication of the Finnish Geodetic Institute No. 124, Helsinki University of Technology, Finland.

Mark, D. M. 1991 Object modeling and Phenomenon-Based Generalization. In: Buttenfield, B.P. and R.B. McMaster (eds.) *Map Generalization: Making Rules for Knowledge Representation*. London: Longman: 103-118.

Meng, L.Q. 1997 Automatic Generalization of Geographic Data. Technical Report of the Institute of Cartography University of Hannover, Germany. 76 pp.

Morehouse, S. 1995 GIS Based Map Compilation and Generalization. In : Müller, J.C. Lagrange, J.P. and R. Weibel (eds.), *GIS and Generalization: Methodology and Practice.* London: Taylor & Francis: 21-30.

Morrison, P. and Morrison, P. 1994 *Powers of Ten.* San Francisco: W. H. Freeman (revised edition).

Muller, J.C. 1991 Generalization. In Longley, Goodchild, Maguire and Rhind (Eds) *Geographical Information Systems: Principles, Techniques, Management and Applications.* London: Longman.

Müller, J.C., Lagrange, J.P. Weibel, R., and **Salgé, F.** 1995 Generalization: State of the Art and Issues. In : Müller, J.C. Lagrange, J.P. and R. Weibel (eds.), *GIS and Generalization: Methodology and Practice.* London: Taylor & Francis: 3-17.

Spaccapietra, S., Parent, C. and **Vangenot, C.** 2000 GIS Databases: From Multiscale to MultiRepresentation. In B.Y. Choueiry and T. Walsh (eds.), *SARA 2000*, LNAI 1864, Berlin: Springer-Verlag: 57-70.

Perkal, J. 1966 an Attempt at Objective Generalization. Trans. W. Jackowski. In Nystuen, N (ed.) Discussion Paper 10, Michigan Inter-University Community of Mathematical Geographers. Ann Arbor Michigan.

Steinhous, H. 1960 *Mathematical Snapshots*. London: Oxford University Press.

Thompson, M. M. 1988. *Maps for America: Cartographic Products of the United States Geological Survey and Others*. Washington DC: US Government Printing Office.

Tobler, W.R. 1987 Measuring Spatial Resolution. *Proceedings*, International Workshop on Geographical Information Systems, Beijing, P.R.C., 25-28 May, 1987: 42-47.

Weibel, R. and **G. Dutton** 1999, Generalising Spatial Data and Dealing with Multiple Representations. Chapter 10 in Longley, Goodchild, Maguire and Rhind (Eds) *Geographical Information Systems: Principles, Techniques, Management and Applications.* London: Longman: 125-155.