23^{rd} International Cartographic Congress
Moscow, Russia 4-10 August 2007
http://icaci.org/documents/ICC_proceedings/ICC2007/html/Proceedings.htm

RE-THINKING BEST PRACTICES IN CARTOGRAPHIC DATA CAPTURE AND DATA MODELING

Barbara P. Buttenfield, University of Colorado-Boulder (USA) babs@colorado.edu

Charlie Frye, ESRI (USA) cfrye@esri.com

Introduction

Data capture of base cartographic features continues as a major activity in national mapping agencies, regional and local governments, and field science offices. Data modeling to create map products at a single scale and for a single purpose is both costly and labor-intensive. As a consequence, most organizations capture data once with the intention of modifying it for use in multiple products produced for display across a range of mapping scales and for reference, general purpose or special purpose mapping. In ICC member nations, National Mapping Agencies (NMAs) capture topographic data, image data, and base-cartographic vector data at the finest spatial resolution (e.g.,1 meter or sub-meter ortho-imagery), for display at standardized mapping scales (e.g., 1;10,000, 1:25,000, 100,000, 250,000, or 1: 1 million), according to the mission and mandate of the agencies, e.g., defense and security, or natural resource management, or archival of natural landforms.

It is widely acknowledged that a single data representation cannot serve all audiences, all types of use, or all mapping scales without some type of modeling, alternative symbolizations, or both. Managers are aware of this, but do not at present structure data capture projects with aforethought to the software processes needed to transform the data for specific uses. Regardless of precision or accuracy, limits are imposed upon data by the very act of capture. A large-scale topographic data set may provide a record of rock formations, of stream channel widths, or of glacier positions; but neglect high-risk avalanche paths, flood plain boundaries or wetland delineations. Whatever is not explicitly captured will not appear in the Digital Landscape Model (DLM) "raw" database, nor in any databases derived from it. Methods used to capture data will inevitably constrain its suitability for multiple scales and purposes. Thus practices underlying data capture define and delimit its fitness for use.

One priority when capturing “general purpose” cartographic base data must be to minimize the loss of content and characteristics that could in turn limit data’s applicability to fewer purposes across a smaller range of mapping scales. For example, Figure 1 shows captured stream channels for an area of New England. Arrows identify the visible “To” nodes. One tributary flows (correctly) north from a ridgeline, into another stream channel that flows (incorrectly) back uphill into the same ridge. The inconsistency could lead to problems mapping hydrology or flow accumulation. Best practice would guide data capture such that topologic arcs flow consistently downhill.

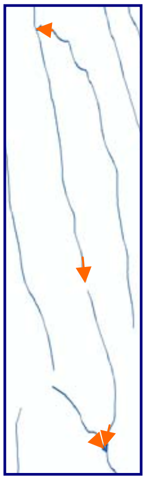

Figure 1

Another priority is to specify parameters of use, to “inform” data capture and modeling. To the extent that data is captured for particular purposes that are loosely or tightly coupled to a particular resolution and mapping scale, specifying intended use *in advance of data capture* will likely reduce or ideally minimize subsequent processing problems. Robust metrics establishing data’s suitability for a given cartographic purpose have not yet been formalized, nor published in the literature. We cannot itemize a comprehensive list of metrics in the space of a single paper. Rather, our intention is to set forth a rationale for such formalization, to maintain a manageable workflow during data capture and modeling.

We begin the paper reviewing an article published fifty years ago this year, promoting a practice of planning at the time of data capture for multiple data applications and mapping scales. Drawing from a theoretical framework of project management in business, we discuss basic factors constraining cartographic data capture and data modeling. We present examples from recent work to demonstrate where “informed” data capture can augment general purpose data capture for multi-scale mapping, and reduce or obviate intensive data modeling. Each examples shows guidelines for best practice to capture data in preparation for general use. Finally we discuss implications for data sharing that incorporate best practice reports delineating fitness for use, which can inform data partners about how to model shared data to a different scale or purpose.

The Multiple Use Concept in Cartography

Sherman and Tobler (1957) published a short note on capturing cartographic base data with an eye toward applying the data to multiple uses. That article is largely forgotten in the literature on cartographic project management, and to be honest it is steeped in the legacies of paper map printing, photographic production, and a blind faith that "…changes of scale are easily made…" (p 7). The article is nonetheless prophetic in the context of arguments that when data capture is planned to maximize the number of uses to which it can be applied, overall production costs and problems can be minimized, in most cases.

Arguing that the aggregate labor of all tasks in any cartographic projection project is considerable, they draw the term "multiple use" from engineering construction, using the analogy of building a hydroelectric dam:

> … to serve purposes of flood control, river navigability, irrigation and similar benefits, as well as electric power. Applying the multiple use concept satisfies the need for an optimum allocation of the diverse but finite social and physical resources available. The same concept can be applied to cartography with similar results. (Sherman and Tobler 1957, p.5)

The concept is in fact an advised work practice, where data are compiled from multiple sources, integrating the inputs from each source. Following initial capture, individual map elements are separated and archived for later "selective reassembly depending on purpose" (p.6); and this aspect actually parallels the separation of map elements into feature classes (hydrography, transportation, settlement, etc.). Map separations can be photographically combined, selectively masked, re-scaled (within limits) and integrated with new data, at a much reduced labor cost than initiating new data compilations for each mapping project. "The multiple use concept enables considerable reduction of input and is extremely useful in guiding the development and continuation of a mapping program" (p.5).

The validity of the multiple use concept is firmly situated in planning stages of data catpure. It is obviously not possible to know in advance all the possible uses and scales to which a compiled data set may be applied, and one cannot protect against unsuitable use of any data. Nonetheless, planning in advance for suitable scale changes and mapping purposes can eliminate duplication of effort and make compiled data easier to disseminate. State and federal agencies are targeted user communities in the article, since they operate within time and fiscal constraints of the public domain. We argue that applying the multiple use concept to strategize the capture of digital cartographic data can reduce subsequent processing, can preserve the quality of data products, and can sustain the goal of capturing data for general uses across a wide range of mapping scales.

The Iron Triangle – Fast, Good, or Cheap

For many years, the emphasis of the management science profession has been “an on-time, on-budget, on-quality delivery promise” (Norrie, 2006: 48). This is known as the triple constraint or the Iron Triangle, and it is embedded into the fabric of the profession (PMI, 2004). The constraints of funding are obvious, as is time: clearly, project managers want to bring in data products on schedule and within budget. The third vertex of the traditional Iron Triangle (“quality”) is a term that is taken to mean several things by various authors, for example, available expertise and skill sets of the project staff, processing (manufacturing) capabilities that are available, quality assessment or technical performance measures (Morris and Hough, 1987; Turner, 1999; Kerzner, 2006).

The Iron Triangle is dynamic, varying in response to the particular demands of each project. In the analogy, the length of each triangle side reflects the amount of funding, time, or quality that can be achieved. The length of any individual side can vary, but the sum of all three lengths remains constant (Saarinen, 1990). Constraints interact, and when one is prioritized another must compensate. For example, if a short production time forms the highest priority, then a good quality product can be produced only at a higher cost. The constraints of the Iron Triangle are summarized perhaps by a personal anecdote. Our colleague Jim Robb, the former cartographic manager at University of Colorado, had a sign over his office door for many years that all clients saw as they entered the lab: “Fast, Good, or Cheap – Pick Two”.

Production managers are trained to work within the Iron Triangle when planning projects. They might ask how distorted can the Iron Triangle become and still sustain successful completion of a project. In recent years, the management literature has posed a different question, whether three constraints are sufficient to describe all aspects of project management (Atkinson, 1999; Norrie and Walker, 2004). Failure to look beyond ‘the Iron Triangle’ may cause poor management decisions, and compromise the project as a whole. Norrie (2006: 78) states: “Project management practitioners need to be willing to move beyond the Iron Triangle and make sure that the project work they undertake is truly strategic", and modifies the Triangle to create a Pyramid (Figure 2):

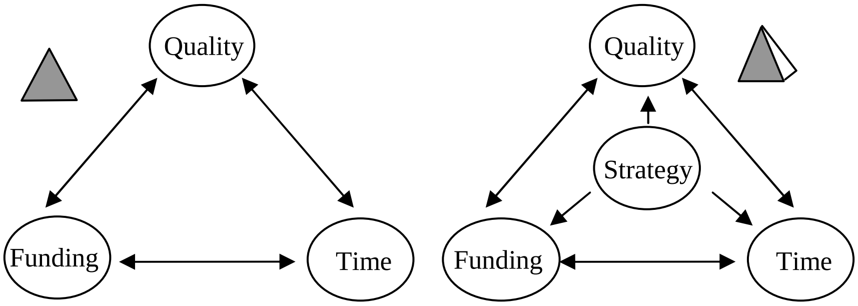

Figure 2

Adding a fourth constraint to management practice (redrawn from Norrie 2006).

Let us return bring this analogy into the context of cartographic data capture and modeling. One might readily understand how balancing the first three constraints can distort the Iron Triangle. For example, when capturing hydrographic features, quality is highest when stream channels and centerlines are both captured, which takes extra time. When capturing buildings, quality is highest when both the footprint and the entry points are captured, which requires additional skill in image interpretation. Adding a fourth (strategic) constraint introduces some complexity to project management, but the extra dimension also offers flexibility: a modified strategy can rebalance an otherwise distorted management situation, as for example if time is short and costs must be preserved, quality could be nearly maintained by capturing only stream channels, and later computing stream centerlines as a data modeling step.

Linking the parameters of the Iron Triangle with Sherman and Tobler's Multiple Use concept, one could argue conversely as follows: managers should recognize that capturing data intended for multiple scales and mapping purposes will require a strategy that differs from capturing data for a single purpose or scale. Furthermore, a more ambitious strategy will require additional time, funding, staff skills, or a combination of these constraints. Knowing this in advance of data capture may save labor, time and resources later in data modeling stages of a project.

Particularly when the goal is 'general purpose' topographic data capture and modeling, the constraint called "strategy" in the Iron Triangle analogy refers to flexibility of data use, or in the cartographic jargon, “fitness for use”, which means capturing data with an eye toward general use. The management strategy might be described as follows: do not capture for special purposes, but rather opt for multiple uses, multiple scales, and capture features that can be used to derive other features, within existing constraints.

Identifying the requirements for data that maximize its fitness for use at multiple scales or purposes in advance of data capture serves two important management goals. First, it balances the planned workflow by anticipating where labor and time must be concentrated in stages of data production. For example, recognizing that building entry points must be recorded in the database adds time to the initial data capture, but may reduce subsequent processing time (for example, geocoding building addresses to infer an entry point), which may free staff time for other processing tasks. Second, a codified set of requirements can be exchanged with state and local agencies or with private sector groups to whom data capture may be outsourced. In this way, consistency in the final data product can be maintained even when the workflow is distributed.

We turn now to case studies in a recent cartographic data capture and data modeling exercise where an absence of pre-defined requirements for suitable use affected data capture practices so as to actually hamper subsequent data modeling, somewhat limiting our ability to generate smaller scale datasets from the initial capture version. We do not intend to criticize those county offices who contribute to the spatial data infrastructure.; rather, we present these case studies as “lessons learned”, to help such contributors work more efficiently and effectively.

Case Studies from Ada County Idaho

The project objective was to build a multi-scale, multi-purpose topographic database for a county government office in Idaho, a state in the Northwest region of the USA. Beginning with data they had captured at 1:5,000 (5k), we modeled the data to generate another database for base mapping at 1:25,000 (25k). We have also run mapping experiments for multiple map purposes, and those are reported on elsewhere in this conference proceedings (Brewer et al 2007). Figure 3 shows an example 5k map product and a 25k product, to give a sense of captured and targeted content, and detectable resolution. In the 5k community map product on the left, settlement features are displayed by building footprint; larger buildings and landmarks are labeled; and major and minor roads are displayed as labeled polygons. In the 25k city map product on the right, residential features are shown as regularized rectangles, while larger building footprints are preserved. Only major roads are labeled. Hydrographic features at both scales include major rivers, minor streams, canals and irrigation ditches, and intermittent stream channels, shown as polygons or line features depending on size.

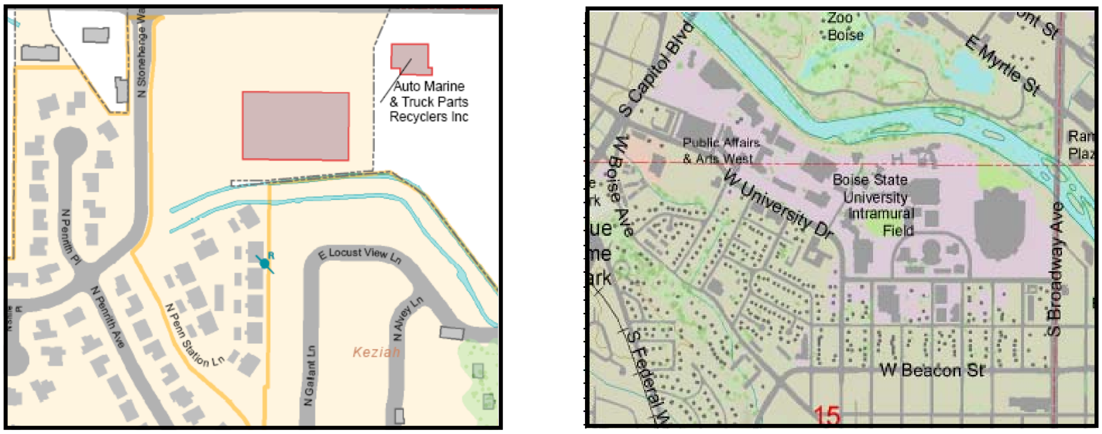

Figure 3

Sample Map Products from the Ada County, Idaho multi-scale database.

Ada County is an arid region, and water is quite important to business and to quality of life. Thus, much of our attention has been paid to re-thinking best practices for capture and modeling of hydrography. Our first case study compares hydrographic polygons with centerlines (Figure 4). The image shows that the stream channel passes under the road. The polygon was captured as a straight path terminating at the edge of the road. The initial centerline (red in the figure) was captured independently of the polygon and as a consequence does not coincide, in fact it falls outside the polygon channel altogether. We modeled a new centerline (blue) to correct this problem. It’s difficult to identify all such discrepancies by automated methods, and manual data modeling adds time and cost to the data modeling process. Re-thinking best practice

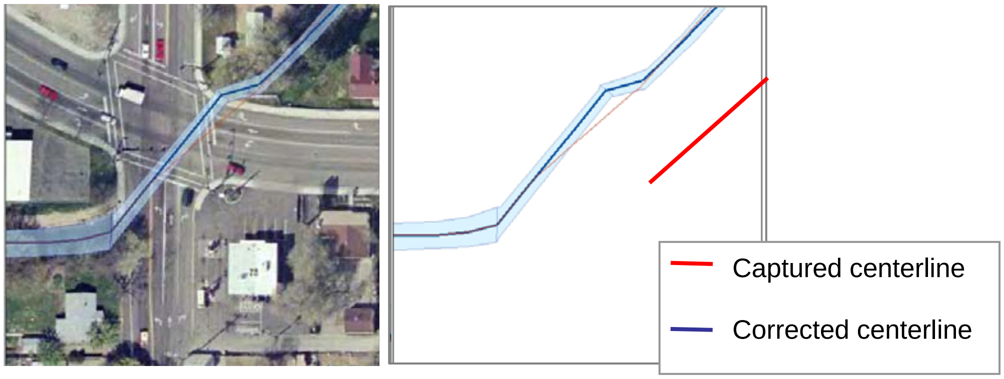

Figure 4

Problems arise when centerlines are captured solely from imagery.

based on this case would have codified an initial requirement that stream centerlines should be captured so as to lie inside stream polygons, ensuring that the data capture would be based not only on imagery but also on previously collected hydrography.

The next case study follows from the first. When capturing centerlines of each feature informed by polygon capture, capture requirements should specify that linear segments extend as far, and not farther than the polygons (Figure 5a). When captured centerlines are overlaid on the Ada County polygons, it is evident that multiple

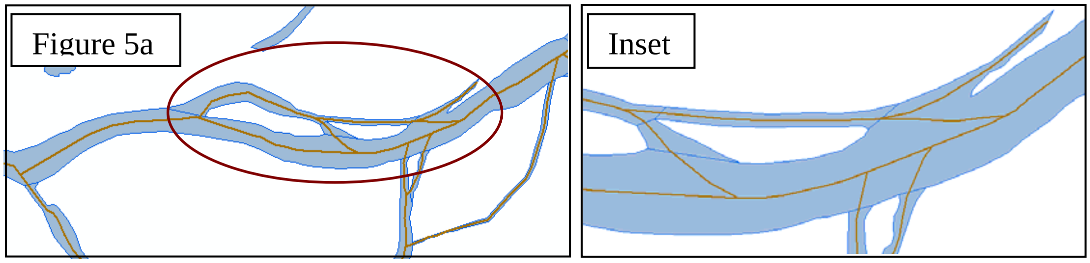

Figure 5

Multiple centerlines within polygon boundaries.

centerlines have been captured within individual polygons. A traditional selection and elimination process will convert all channels to centerlines, whereas these data are captured to support a process where polygons that represent channels are eliminated and replaced with centerlines. Re-thinking best practices based on this case would codify requirements for centerlines in smaller channels by capturing hydrographic polygons on the basis of width, and to attribute widths to centerlines so that selection and elimination can remain flexible.

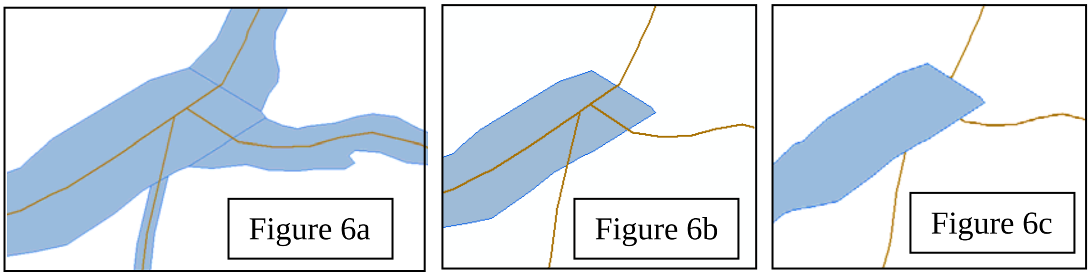

Figure 6

Lack of coincidence between representations creates modeling and display problems.

In Figure 6a, centerline nodes do not coincide precisely with the polygon boundaries. This can create geographic nonsense when narrower channels are displayed by centerlines. In Figure 6b, the centerline confluence appears inside the hydro polygon. Awkward graphic transitions are not entirely satisfied by modifying drawing order (Figure 6c). The break between the main channel polygon and the centerline feature running north from it leaves the impression that a dam or some other human barrier is present. Re-thinking best practice in this case would specify that centerline confluences and polygon endpoints must coincide, and that no more than one centerline should be captured within a single polygon. The data modeling required to correct this case involves intensive effort, integrating a combination of generalization (rounding off the shape of the polygon terminus) and symbolization (changing the centerline symbol width).

A third case study highlights the best practice of correcting stream topology at the time of data capture, to protect hydrologic continuity during selection and elimination. This is especially important when data are captured with an intention to generate additional representations at smaller mapping scales. In Figure 7, stream channels were captured as line features, but segments do not begin and end at hydro area boundaries.

Selection queries on stream width have damaged topology. One modeling solution is to insert stream order as a new attribute, and run selection queries on both stream width and stream order. Another solution is to run an identity overlay against the captured data to reselect lost segments, then rebuild topology. Coding stream segments by name and then stream name by notoriety would be another strategy, thus either culturally or hydrologically important water ways could be preserved during generalization processes independently of any statistical characteristics specified in the generalization workflow.

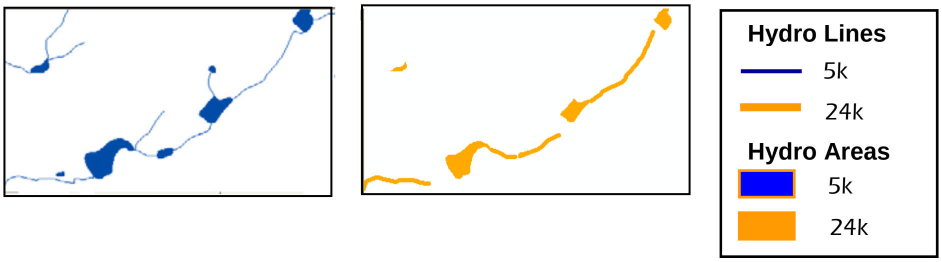

Figure 7

Topology may be damaged by selection queries for smaller scale mapping.

Summary and Discussion

The goal of the paper is to re-think best practices for cartographic data capture and data modeling. Informing data capture by pre-specifying strategies that achieve the highest suitability of captured data for modeling at multiple scales and for multiple purposes can insure that fitness for use is preserved, as well as by reducing subsequent modeling to correct cartographic nonsense and geographic inconsistencies. The practice of informed data capture relies on the multiple constraints planned for in management science discussions of the Iron Triangle, namely time, funding, quality, and more recently, a fourth (strategic) constraint to stabilize management of workflows.

Sherman and Tobler's (1957) Multiple Use concept advises introducing a degree of redundancy by capturing and deriving multiple representations of cartographic data that serve multiple purposes. Alluding implicitly to the Iron Tirangle framework, they claim that planning which anticipates the multiple use concept "permits maximum flexibility without loss in quality" (p.5). Clearly in digital data production environments, redundancy creates problems for data updates; and linking multiple representations is complicated. Managers work to achieve a balance between flexibility and parsimony. Compiling data is more expensive than modeling data, for any product.

The rationale for specifying intended uses at intended target scales prior to data capture is twofold. First it "frontloads" work effort at the data capture stage rather than during data modeling, when inconsistencies become much harder to correct. Limitations imposed at data capture cannot be overcome if the semantics of intended use have not been codified. Metrics about what is captured can be added (e.g., geometry and topology, feature crosswalks, etc.), but this becomes more difficult after the capture process is complete. Second, formalizing metrics on fitness for use can help to educate local and regional data producers about efficient and effective data capture strategies. Strategies that insure high suitability for multiple scales and purposes and also insure consistency and high quality can only hasten the advance of national and global geospatial data infrastructures.

We emphasize that we cannot itemize all possible strategies, nor (on the basis of a single multi-scale, multi-purpose mapping project) propose a standard for codifying the metrics for suitability. We call upon national mapping communities to initiate lists specific to their own agencies, to codify and to exchange them across local and state levels, in order to improve best practices at all levels of cartographic data capture and modeling.

Acknowledgement. This research is funded by ESRI Professional Services Agreement No. 2003B6347. We acknowledge student research assistance from Kiyoshi Yamashita and Anna Harmon at University of Colorado.

References

Atkinson, R. 1999 Project management: Cost, Time and Quality, Two Best Guesses and a Phenomenon. *International Journal of Project Management* 17: 337-342.

Brewer, C.A., Buttenfield, B.P., Frye, C. and Acosta, J 2007 Scalemaster: Multi-Scale Mapmaking from Multiple Database Resolutions and for Multiple Map Purposes. *Proceedings 23rd International Cartographic Conference*, Moscow, Russia, Aug 2007.

Kerzner, H. 2006 *Project Management: A Systems Approach to Planning, Scheduling and Controlling*. 9th ed. New York: Wiley.

Morris, P.W.G. and Hough, G.H. 1987 *The Anatomy of Major Projects*. New York: Wiley.

Norrie, J.L. 2006 *Improving Results of Project Portfolio Management in the Public Sector Using a Balanced Strategic Scoring Model*. Doctoral Dissertation, School of Property, Construction and Project Management, Design and Social Context, Royal Melbourne Institute of Technology (RMIT). http://dhtw.tce.rmit.edu.au/James%20Norrie/Norrie-Thesis.pdf (28 May 2007).

Norrie, J.L. and Walker, D.H.T. 2004 A Balanced Scorecard Approach to Project Management Leadership. *Project Management Journal* 35(4): 47-57.

PMI 2004 *A guide to the project management body of knowledge*. Newtown Square, PA: Project Management Institute.

Saarinen, T. 1990 Systems Development Methodology and Project Management. *Information and Management* 19: 183-193.

Sherman, J.C. and Tobler, W.R. (1957) Multiple Use Concept in Cartography. *Professional Geographer* 9(5): 5-7.

Turner, J. 1999 *The Handbook of Project Based Management.* 2nd edition. New York: McGraw-Hill.