AN APPROACH TO EVALUATION AND BENCHMARK TESTING OF CARTOGRAPHIC DATA PRODUCTION SYSTEMS

David Selden
Battelle Columbus Laboratories
2030 M Street, N.W., Washington, D.C. 20036

BIOGRAPHICAL SKETCH

David Selden is a Research Cartographer with the Battelle Columbus Laboratories, Washington Operations. His current research activities include a study of raster-to-vector conversion processes and development of an analog-to-vector conversion cartographic benchmark testing package for the United States Defense Mapping Agency. He is also a task leader for the design and production of maps of U.S. cancer mortality rates and trends for the U.S. Environmental Protection Agency (EPA). In addition he has recently completed a preliminary investigation of geographic information systems for the EPA Office of Toxic Substances. Mr. Selden holds a B.A. in Geography from Clark University and a M.S. in Cartography from the University of Wisconsin-Madison.

ABSTRACT

The Battelle Columbus Laboratories has developed an approach to evaluating state-of-the-art automated cartographic data/ map production systems. Evaluation is performed within the framework of an analog-to-vector (A/V) conversion model, which encompasses the overall cartographic data production process (i.e., source preparation, digitization, feature tagging, spatial coding, and data management). Attention is focused on the unique characteristics of cartographic manuscripts and the limitations of data capture systems. It also addresses editing functions, process speed, procedural trade-offs and production bottlenecks. These are all analyzed in the context of pertinent organizational requirements. A set of benchmark testing materials and procedures have been developed to assist in the evaluation of A/V systems. The paper discusses the A/V conversion model and the development of the benchmark testing package and methodologies.

INTRODUCTION

The need to transform large quantities of analog paper map products into digital cartographic data continues to confront many organizations. The development of sophisticated automated graphic data capture systems has made a significant impact on the map conversion process. Raster scanners and automatic line following devices are replacing manual digitizers. Data editing facilities, built upon state-of-the-art color computer graphics workstations, combine interactive, computer-assisted and automatic functions in raster and/or vector modes to provide advanced capabilities. New algorithms and hardware processors improve raster-to-vector conversion rates. Automatic feature recognition/extraction and spatial/topological encoding software is emerging from research laboratories to support human operators in building feature coded digital cartographic vector data files. These technological advances appear to bode well for the efficient conversion of remaining stores of paper maps and related graphic manuscripts. Significant challenges and difficulties remain, however.

Organizations in search of state-of-the-art capabilities must judge the type and brand of data capture system which will best serve their special needs. Organizations in possession of such capabilities, or less modern facilities, must continually reassess the most effective utilization of their existing systems in light of changing present and future requirements. Those in the market for new systems and those already in production both need to ask similar questions.

The ability to ask meaningful, penetrating questions about automated cartographic data capture systems depends on a solid understanding of current and developing technologies. It also presumes a sound knowledge of map products, digital data specifications, organizational qualities, and the overall cartographic map conversion/data production process. This final layer of knowledge should form the basis for evaluating state-of-the-art automated cartographic data capture systems. Commercial data capture systems are built to convert many different types of analog graphic materials to digital data and are not necessarily specifically designed for cartographic applications. Therefore, it is incumbent upon mapping organizations to integrate such systems into their production process. Conversely, a system's specifications and limitations will often dictate the details of a map/data production process. These often cross-cutting requirements must be reconciled at every juncture in the map conversion process. Thus, Battelle Columbus Labs has developed the analog map to digital cartographic data conversion model (A/V) as a conceptual framework for evaluation of systems in production or those in the marketplace.

THE ANALOG-TO-VECTOR (A/V) CONVERSION MODEL

The conversion of paper map products into digital cartographic data is accomplished through a series of procedures which generically are common to most organizations involved in this activity. The A/V conversion model (Figure 1) identifies the five major components of this process: Preparation, Digitization, Tagging, Spatial Coding, and Data Storage. A sixth optional component, Raster-to-Vector Conversion, is also included where raster scanning has been employed for initial data capture. The remainder of this section will discuss each individual component and highlight critical aspects of the interrelationship between the process and automated data capture systems.

Analog-To-Vector Conversion--Preparation

Preparation of analog cartographic manuscripts for conversion to digital representation addresses two categories of information: analog graphic and analog feature attribute. Analog graphic is any cartographic "paper" product containing point, line and area elements, derived from either photogrammetric compilations (e.g., color pencil on mylar), four color printed maps, color/feature separation film positives or color proofs. Analog feature attribute preparation refers to the process of preparing feature code overlays which annotate codes for each type of cartographic element to be digitized and processed. Preparation of cartographic manuscripts consists of four phases: (1) Data Selection, 2) Manuscript Reformat, 3) Manuscript Enhancement, and 4) Manuscript Review Edit.

ANALOG-TO-VECTOR CONVERSION MODEL

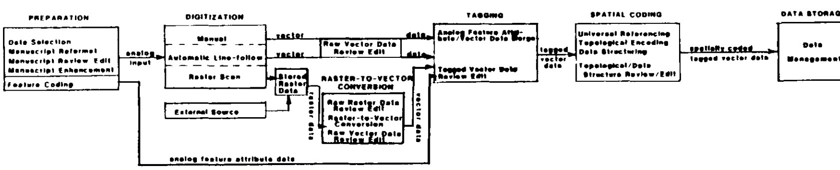

Figure 1. Analog-to-Vector Conversion Model

<u>Data Selection</u>. The selection of specific data categories from cartographic manuscripts is the first step in the preparation stage. In many cases only select categories are required from an original input (e.g., only contours from a topographic quadrangle). Alternatively, all information on a compilation manuscript may be submitted for digitization. The results of data selection greatly influence the proceeding steps in preparation and also challenge the capabilities of data capture systems to varying degrees. Its particular impact depends on the data categories chosen, their cartographic geometries (i.e., non-intersecting lines, closed polygons, merging networks, crossing networks, point features, symbolized elements, and labeling - see Figure 2), material type (e.g., mylar, paper), data representation type (e.g., ink, scribing, color pencil), material age, data density, data quality (e.g., broken lines), and material-data/system(s) compatibility (e.g., color input for a black and white scanner).

<u>Manuscript Reformat</u>. This process concerns the transformation of cartographic manuscipts prior to digitization. In some cases, analog manuscripts may require recompilation, photo-reduction, panelling or subdivision into multiple pieces due to size (i.e., larger than scanning format) or image density (resulting in too much data for a digitizing system to efficiently process in a single pass). In many instances, manuscript reformatting reflects an initial incompatibility between input cartographic documents and a particular digitization system.

Manuscript Enhancement. The enhancement of analog cartographic manuscripts is done to provide additional data coverage (e.g., intermediate contour lines in sparsely covered areas), assist in topological encoding (e.g., highlighting bridge intersection in a unique color to identify a node location), and generally improve the efficiency of the analog-to-vector conversion (A/V) process.

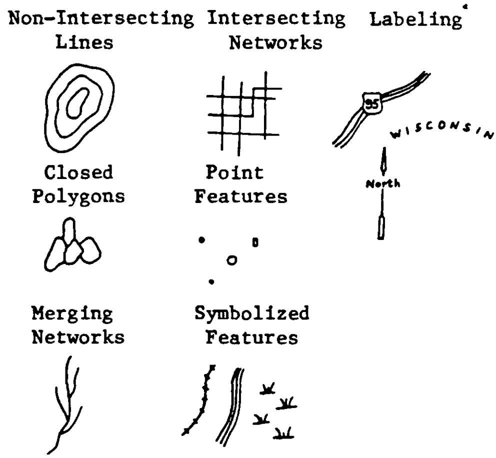

Figure 2. Generic Cartographic Data Types

Manuscript Review/Edit. The final step in the preparation stage is to review the input manuscript prior to digitization to assure data quality (e.g., no gaps in "continuous" lines or merging of "discontinuous" lines such as contours), completeness, and efficient utilization of resources (e.g., analog manuscripts with sparse data coverage may be directed to manual digitization instead of raster scanning).

Analog-To-Vector Conversion--Digitization

The type of digitizing method will dictate to some degree the extent and manner of manuscript preparation. Three basic types of digitization have been identified: raster scanning, automatic line following and manual.

Raster Scanning. This automatic process captures cartographic elements and converts them to digital representation by means of electro-optical, solid state, or laser "cameras" which traverse the manuscript (mounted on a rotating drum or stationary flatbed) in scan lines at prescribed resolutions. Some scanners are color sensitive, capable of capturing and storing up to twelve colors simultaneously, thus offering flexible analog input requirements. Others are solely single color scanners or require multiple pass scanning (through color filters) to achieve full color scanning capabilities. Scanning resolution is variable on some systems and fixed on others. Variable resolution scanners permit the selection of the appropriate setting for maximum "acuity" and minimal data storage. Output of raster scanning typically requires data editing and conversion to centerline vector representation.

Automatic Line Following. This automatic process (although also operated in "manual" and semi-automated modes) captures the x,y coordinate positions of cartographic elements through the use of video "cameras" or deflected laser beams. Two important advantages of this approach is the capability of concurrent feature tagging and the elimination of raster-to-vector conversion processing (although data filtering, editing, and reformatting usually occurs as a post-process).

Manual Digitization. The manual capturing of x,y coordinate positions of cartographic elements with hand-held electronic cursors still maintains certain advantages over automatic procedures. Primarily, it is more effective to capture map sheets of sparse data coverage in this manner. Additionally, concurrent feature tagging and no raster-to-vector conversion requirement make this method advantageous in certain cases.

Analog-To-Vector Conversion--Raster-To-Vector Conversion

This is the process of converting digital cartographic data from a raster (scan line or grid) data structure to a x,y vector, centerline (or edge) representation. This is only required as a result of raster scanning for initial data capture and reflects typical vector data structure requirements for many digital cartographic data bases or geospatial analysis packages. The A/V conversion model identifies three components of raster-to-vector conversion: raw raster data review/edit, raster-to-vector conversion and raw vector data review/edit. Individual data capture systems implement raster-to-vector conversion in unique ways, through alternative algorithms, special hardware processors, or combining or by-passing entire components (e.g., no raw raster data review/edit).

Raw Raster Data Review Edit. This is the process where raw raster data (i.e., direct output of raster scanning) is reviewed for data quality, completeness, and anomalies or errors (e.g., gaps, stubs, spikes, sticks, streaks, coalescence, misalignment) commonly associated with the scanning process or attributable to the analog cartographic manuscript. Three approaches to raw data review and editing have been identified. Interactive computer graphics techniques permit human operators to visually search through a file, locate problems and correct them via light pens, cursors, trackballs and other assorted electronic interfaces to the graphics plane. A second alternative utilizes computer-assisted software functions which automatically identify errors and facilitate their correction through interactive computer graphics techniques. A third option relies on fully automatic error identification and correction software. Typically in this case, certain tolerances (e.g., identify and close gaps equal to or smaller than three pixels by two pixels) are set by an operator prior to initiation of a routine.

One evaluation criterion for data capture systems is the availability of raw raster editing facilities. The provision of one, two or three forms of raster editing support is a secondary consideration. Finally, the overall effectiveness of any of these functions dictates their true value. A commerical systems lack of raster data editing capabilities in any form may reflect a serious handicap. In recent years, a number of vendors of data capture systems have developed sophisticated capabilities in this area. Certain functions such as "snow" removal (i.e., unwanted, miscellaneous data elements) may be more effectively accomplished in a raster environment. The sole provision of interactive computer graphics for raster data editing limits an organization to costly labor intensive activities in this phase of the A/V conversion process. Computer-assisted and automatic editing functions promise lower manpower requirements and improved editing performance over manual editing techniques. However, this can be very misleading. Successful resolution of data quality problems and the degree to which automated functions create new errors or inconsistencies (e.g., gap closure on contour data can result in the unintended connection of close contour lines) requires scrutiny.

Raster-To-Vector Conversion. This concerns the actual conversion of raster cartographic data to centerline vector representation. One standard approach implements this data conversion in three steps: skeletonization, line extraction, and topology reconstruction. Skeletonization is the reduction of raster elements to one unit of resolution through a process of peeling, ballooning or centerline calculation. A number of data anomalies have been shown to result from the skeletonization process including generation of stubs, gaps, and unthinned data elements. Subsequent correction of these anomalies in raster and/or vector mode is required and should be anticipated for systems applying skeletonization algorithms. Line extraction is the identification of unique line segments and is accomplished in two basic ways, line following or scan line. The line following method tracks vectors until encountering intersections at which point nodes are identified. Scan line processing extracts individual line vectors as they are encountered one scan line at a time. This latter approach requires that topology reconstruction (i.e., the identification of point, line and area geometric, spatial relationships) be explicitly performed as a post-process.

A second approach to raster-to-vector conversion in the CAD/CAM, engineering drawing market has recently been reported.* In this case, boundary or edge chains are generated from line drawings in real time concurrently with raster scanning via a special hardware processor. A subsequent process generates centerline vector data via software calculation of centerline x,y coordinate locations.

The evaluation of raster-to-vector conversion revolves around two basic issues, speed and quality. Processing large amounts of data typically associated with cartographic input often results in production bottlenecks in raster- to-vector conversion. This is complicated by the variability of cartographic geometries and the capability of alternative

* A Defense Mapping Agency (DMA) Raster-To-Vector Analysis, Battelle Columbus Labs - currently unpublished.

conversion algorithms to efficiently process them. The quality of the data resulting from raster-to-vector conversion is of equal concern. Fast conversion with poor quality output is of little value. The extensive efforts and time required for error detection and correction may seriously counteract the perceived advantages of raster scanning data capture (i.e., fast raw data capture with lessened dependence on human operators). Unfortunately, most commercial vendors reveal little or no information about their raster-to-vector software (for proprietary reasons) thus leaving potential customers without critical knowledge.

Raw Vector Data Review/Edit. This process directly parallels raw raster data review/edit. Interactive computer graphics, computer-assisted software and automatic error detection and correction software functions can all be implemented in vector mode. Some commercial systems rely almost solely on the vector editing functions with minimal or no raster editing capabilities.

It should be noted that both manual and automatic line-following digitizing systems also require and usually offer raw vector data review/edit capabilities. Errors such as gaps, spikes, and linear misalignments are quite common as a result of human or machine deficiencies or existing analog manuscript anomalies. The need for the full array of error detection and correction tools similar to those described for raster scanning systems is equally important for vector digitizing systems.

There are two notable trends in the development of editing capabilities. The provision of three levels of editing as previously described is becoming an industry standard. This reflects an editing philosophy based on the "80-20" rule which sets as a goal the detection and correction of eighty percent of all errors via automatic functions and the remaining twenty percent through a combination of interactive computer graphics and computer-assisted routines. (No evidence has been presented which demonstrates attainment of this goal by any commercial vendor. The degree to which a vendor can legitimately claim successful editing automation should be of great concern to prospective customers.) The second trend worthy of note is the development of "parallel" raster and vector data editing capabilities, often within a single computer graphics environment. Both raster and vector data can be edited at a single "station" and in increasing cases vector data can be overlaid on its raster counterpart. This is particularly effective in reviewing certain kinds of anomalies including linear misalignments of centerline vector data.

Analog-To-Vector Conversion--Tagging

This is the process of logically and physically associating vector cartographic data with analog feature attribute information in a digital computer file.

Analog Feature Attribute/Vector Data Merge. This process links vector data with analog attribute information (codes) by means of interactive computer graphics, computer-assisted software, or fully automatic methods. The reported development of automatic contour tagging routines (in commercial vendor research laboratories) is ushering in a new era of automatic feature extraction and tagging capabilities. This presages diminishing labor intensive feature tagging activities which presently require significant organizational personnel and time resources. The evaluation of cartographic data capture systems should consider capabilities in this area and more importantly ascertain a company's commitment to on-going research and development. The successful development of automatic feature tagging capabilities in raster scanning systems will emerge as a significant factor in comparing raster and vector systems.

Tagged Vector Data Review/Edit. This is a quality assurance step which assures the correctness and completeness of the tagging procedure. An array of techniques are available which support this activity. Among them are interactive computer graphics, graphic plotting, and software checking routines.

Analog-To-Vector Conversion--Spatial Coding

Spatial coding is the logical and physical definition of a cartographic data file's internal and external spatial relationships. Three types of spatial relationships are defined: universal referencing, topological encoding, and data structuring.

Universal Referencing. This describes the conversion of a cartographic data file stored in table (cartesian) to a universal reference system (e.g., Universal Transverse Mercator or Latitude/Longitude). Most state of the art cartographic data capture systems provide universal referencing capabilities.

Topological Encoding. This process defines the internal spatial structure of a digital cartographic file by explicitly delineating the adjacency relationships of points, lines and areas. In recent years a number of standard government topological data structures have emerged (e.g., U.S. Geological Survey's DLG-3 or GIRAS and Bureau of the Census GBF/DIME or TIGER). Vendors of commercial cartographic data capture systems currently support production of various types of topological structures and some are reported to be developing software to support the government structures referred to above. As data interchange increases in importance, the need for standard topological data formats will also grow.

Data Structuring. Data structuring is the transformation of a raw data file into a specified data format. All commercial data capture systems maintain their own internal data formats and often support system independent formats to permit conversion to other systems and data structures. Many vendors will develop custom software to support a client's unique data structure requirements.

Analog-To-Vector Conversion--Data Management

Data management is the final component in the analog-to-vector conversion model. It concerns the processing, storage, manipulation and retrieval of digital cartographic data. All data capture systems provide some form of data file management capabilities. Some support attribute storage and retrieval. Generally, these management capabilities are designed to support editing functions and are not intended for comprehensive digital cartographic data and spatial data management applications.

THE DEVELOPMENT OF BENCHMARK TESTING MATERIALS AND METHODOLOGIES

The discussion of the analog-to-vector conversion model and the interrelationship between cartographic data production and the technology of data capture systems highlights critical production concerns and significant areas of ongoing development. The intention of the previous discussion was to develop a comprehensive understanding of the overall production process in the hope that such knowledge will assist organizations in formulating evaluation criteria of data capture systems which meet their special needs. An integral part of the evaluation process is the actual benchmark testing of systems to produce empirical evidence of system performance. Battelle has developed a conceptual approach to creation of benchmark testing packages for cartographic data production systems.

The benchmark testing capability described below is limited in technical scope to basic analog-to-vector conversion processing technology. Specifically, it addresses the procedures and problems associated with capture and conversion of elemental cartographic geometries found on typical analog cartographic base maps. The geometries consist of point, line and area cartographic elements possessing limited symbology (i.e., line thickness variation and limited dashing, only). This does not imply an ignorance or denial of the emerging importance of advanced data capture system technology (e.g., pattern/feature/character recognition and tagging) and the challenges of processing complex cartographic input. It does emphasize, however, the fundamental role and inherent complexities of processing elemental cartographic data.

Benchmark Testing Materials

Three types of testing materials have been developed: standard products, synthetic cartographic geometries, and a quality/editing capability test sheet. Together these materials provide a comprehensive assessment of a data capture system's level of performance with different types of input and processing.

Standard Products. A representative set of typical map products which are produced, stored or required by an organization form the basis for a standard products benchmark testing materials package. Selection criteria should be based on representative types of maps, data densities, kinds of materials, method of production (i.e., ink, scribing, etc.) and overall quality variations. The obvious value of such a package is to test a data capture system's capability in assimilating and processing the range of input materials commonly held by a mapping organization. Every organization maintains a unique array of analog map products, thus the creation of a custom set of standard product materials is required.

Synthetic Cartographic Geometries. These benchmark testing materials consist of twelve film positive sheets containing three abstractions of cartographic geometric patterns found on the products of a mapping organization. Each pattern is reproduced four times (thus twelve sheets) with increasing amounts of data. The primary utility of the synthetic cartographic geometry benchmark testing materials is to support analysis of the impact of geometric formation and data density on raster-to-vector conversion throughput. Given a known quantity of linear inches it is possible to calculate raster-to-vector conversion rates for different types of geometries, data densities and alternative algorithms or systems. Thus, without understanding the inner workings of particular algorithms, a prospective customer can empirically test a data capture system for raster-to-vector conversion performance.

Quality/Editing Capability Test Sheet. Given the importance of data quality and the amount of resources currently invested by mapping organizations in quality assurance activities, in addition to commercial development of advanced editing capabilities, there is an obvious requirement to evaluate and test system functions in this area. Figure 3 portrays a small section of the test sheet. It essentially consists of a series of "perfect" cartographic geometries (i.e., abstractions of cartographic geometric patterns with no discernible anomalies or flaws) in the first column. Arrayed to the right each "perfect" geometric sample are replications of the pattern with purposefully created errors (e.g., gaps) with known dimensions. The combination of perfect geometry and geometric degradation samples are repeated two additional times in different lineweights.

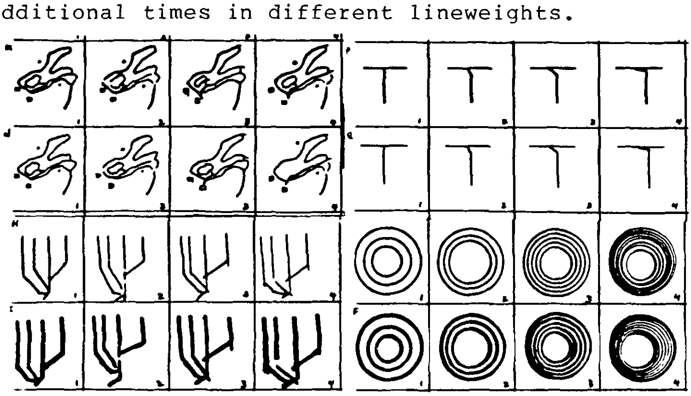

Figure 3. Quality/Editing Capability Test Sheet (portion of sample prototype)

The overall purpose of the quality/editing capability test sheet is twofold. First, as the analog sheet is processed (i.e., digitized, edited, converted, etc.) evidence of system introduced errors can be collected by observation of the "perfect" geometric samples. All errors, anomalies and geometric degradations will be readily apparent and easily identified in the first column of the test sheet after each operation. Secondly, extensive testing of computer-assisted and automatic editing functions is made possible. Tolerances can be set to verify the sensitivity of algorithms to small variations. For example, with known gaps of two, three, five, and seven thousandths of an inch (or a pixel equivalent) in various kinds of geometric patterns, will gap detection and correction routines succeed in resolving only those gaps at the prescribed tolerances? How will different geometric patterns with equal gap sizes be treated? What rate of automatic editing success will be observed. This benchmark testing sheet will provide a prospective customer of any commercial data capture system with the means to more closely, and concretely evaluate system editing functions and overall quality related issues.

Benchmark Testing Methods and Evaluation Criteria

Benchmark testing methods fall into several activities: raster scanning (or alternative digitization methods), automatic raster editing, raster-to-vector conversion, automatic vector editing, and plotting on an automatic vector plotter. Evaluation criteria is based on individual process times, combined process times, virtual image quality assessment, digital plot/analog input "overlay" quality analysis, system integration/user friendliness evaluation, and statistical tests of timing data.

CONCLUSION

A comprehensive understanding of the overall analog map to digital cartographic data conversion process (A/V) is critical to an effective evaluation of cartographic data production systems. Although certain components may appear to wield the greatest influence over production performance (e.g., large format color raster scanners and fast raster-to-vector conversion algorithms), it is the total production process as implemented by a particular system which must be considered. Additionally, the unique characteristics of personnel and products associated with a mapping organization should be factored into the evaluation equation. It is the integration of organizational requirements and system capabilities which ultimately dictates the degree to which production goals are achieved. Thus, development of effective benchmark testing tools and the evaluation of cartographic data production systems depends as much on documentation of "in-house" procedures, capabilities and requirements as it does on an indepth understanding of current technologies.

REFERENCES

Selden, David and Went, Burton, A Defense Mapping Agency (DMA) Raster-to-Vector Analysis, 1984 (currently unpublished)