DATA BASE MANAGEMENT FOR GEOGRAPHIC INFORMATION SYSTEMS

BY: Michael G. Pavlides
Greenhorne & O'Mara, Inc.

Remote Sensing has permitted scientists to view "old" data - or data that have been examined previously by other techniques - from a variety of new and different perspectives. It has also permitted scientists to view and interpret "new" data - or data that have not previously been acquired and analyzed. The analytical processes involved (e.g., spectral signature analyses), as well as the technology itself from a data acquisition point of view are data comprehensive.

The technology of Geographic Information Systems (GISs) has automated the once manual technique of overlaying data for analysis and interpretation. When manual techniques were employed only the most important data overlays (or "data bases") were used since analysts could not possibly assimilate all available data or review all possible data combinations due to the traditional constraints of time, budget, accuracy and human factors. In other words, when performing overlay analyses manually, analysts were required, out of necessity, to be "selective" and choose only those data bases that provided the most comprehensive amount of relevant information for the intended objectives of the analyses. With the advent of increased data storage, computing speed, assimilation capacity, and analytical methods now available in GISs, a shift in this philosophy has occurred. The emphasis now is to add more data to required analyses in order to refine and make the results more accurate. Selective use of data is no longer the modus operandi. This use of GISs is also data comprehensive.

The integration of remotely sensed data with GISs, particularly in the field of energy resource management, has consequently resulted in the formidable problem of managing an extraordinary amount of data. "Data Base Management" (DBM) has become a generally accepted phrase utilized to describe a variety of data storage and manipulation functions, and, in the context of this paper, DBM refers to the "planned" management" or more specifically, the procedure for "order of entry" of the massive variety and quantity of data bases into Geographic Information Systems. A logical approach to determine the data base order of entry is presented herein utilizing management techniques and Consideration Factors (CFs).

Insight into the problem of the magnitude and variety of data bases as related to GISs can be further derived from a discussion of a classification of GISs. In an editorial featured in the July/August 1982 Computer Graphics News entitled "The Future of Geographic Information Systems", Dr. Robert Aangeenbrug classified five (5) types of GISs:

GIS #1. Natural Resources Inventory Systems: Used to perform monitoring and evaluation functions of resource data (e.g., overlays of ecologically-sensitive habitats) for regulation of regionwide activities.

GIS #2. Urban Systems: Serves a dual function as a Land Record Management file (for tax purposes, for example) and a related engineering design data file (e.g., topographic data).

GIS #3. Planning And Evaluation Systems: Used generally to provide thematic display of the entire realm of geographic data (such as socioeconomic information) most frequently by general or relative spatial relationships for use by planners and policymakers (Dr. Aangeenbrug cites the Decision Information Display System as an example).

GIS #4. Management, Command And Control Systems: Used for "strategic" planning determinations by industry and military planners. According to Dr. Aangeenbrug this GIS is similar to GIS #3 with the exception of actual program structure. However, another difference lies in the analytical inference capabilities of these systems to derive systematic relationships from examination of combinations of data bases. Whereas GIS #3 may display the number of unemployed individuals by county in a state, GIS #4 may relate the unemployed individuals to their previous annual income to allow the analysts to conclude what level of jobs are being lost with most frequency.

GIS #5. Citizen/Scientist Systems: Provides the user, access to informational data bases through common telecommunication carriers such as home computers and television sets.

Dr. Aangeenbrug has, of course, developed a broad categorization of GISs as a function of applications rather than internal program structure. It is interesting that, broadly speaking, each of these application functions represent a relatively distinct type of data set.

GIS #1 deals with scientific data or the physical and/or environmental (i.e., "real") characteristics of eco-terrain units (e.g., the polygonal area of a specific soil type). Since the boundaries of most eco-terrain unit data are interpreted and subject to natural, ongoing change rather than "absolutely fixed" in space or time, relative (rather than true geographic) spatial relationships between data subsets are ordinarily sufficient for purposes of GIS analyses. GIS #2 deals with "engineered or measured" data, which are important from an accurate (rather than relative) geodetic spatial relationship (e.g., property tax maps with meets and bounds, planimetric maps based on state coordinate grid systems). GIS #3 deals with "descriptive data" relating general information of the geounits, most often social and economic characteristics (e.g., the number of voters in a county). GIS #4 deals with developing a data base of "data interrelationships". Although not stated by Dr. Aangeenbrug, this GIS directly, or indirectly through user interpretation, is used to derive "interrelationship data bases" that are directly relative to intended user objectives, from analyses of inputted data bases (e.g., development of income statistics related to levels of

energy consumption in a county). GIS #5 deals purely with "informational data" or the display of data that may or may not be geounit-oriented. GIS #5, by definition, might include all data in GIS #1 through GIS #4 and any other information of a general interest to potential users (e.g., the stock market history of a given corporation, the number of cancer patients in a county). In summary:

GIS #	System Name	Data Set Type
1	Natural Resource Inventory	Real
2	Urban	Measured
3	Planning & Evaluation	Descriptive
4	Management, Command & Control	Interrelational
5	Citizen/Scientist	Informational

All of these data set types are further subdivided into data subsets. Subsets for the "Real" data set type are, for example, structural geology, fault and folds, soil types, vegetation and so on. If we consider each possible data set type and subset as a possible data base, the enormity of the data entry problem becomes self-evident.

There is no doubt that GISs exist, or are in process of development with program structures capable of handling multiple GIS application functions and the corresponding data set types inferred by those applications. With the advent of these more sophisticated GISs that are capable of handling more data sets and subsets, the GIS manager is confronted with the significant problem of ordering data base entry to the GIS. This is further complicated by (1) limitations of digitizing budgets, (2) primary user requirements, and (3) by traditional system management problems (monotony of digitizing causing quality problems, high personnel turnover, and so forth).

GIS managers have attempted to solve this problem by immediate entry of data bases required to satisfy user needs without affording the GIS data management concept a more holistic approach. The objective of data base entry ordering is to (1) maximize the efficiency and productivity of the digitizing operation, (2) utilize the available digitizing budget to the maximum extent (i.e., input the most amount of data for the given budget), and (3) satisfy primary user demand for data. The point of this paper is to proffer the concept that the types of distinct data sets represented by GIS #1 through GIS #5 and their subsets should be viewed as an entire set and ordered for input by a management technique. By applying a number of CFs to each data set (and/or subset) the data base sets or subsets can be ranked. These rankings can be related by some type of decision process (e.g., weighting schemes, decision matrices and so forth) to obtain the final entry priority or order for input to the GIS of the data bases. A discussion of applicable decision methods is beyond the scope of this paper but can be found in standard textbooks.

Each of the following CFs should be applied to each individual data set (or subset) intended for GIS use:

CF #1. Data Use

Of primary consideration is the ordering or ranking of data sets by relative importance to the primary objective of the GIS and the time and need requirements of the primary user. However, expected use of data set combinations should be also examined. For example, after the user has ranked each data set in order of importance (e.g., geology - #1, vegetation - #2, habitat - #3, wetlands - #4, and so forth), the GIS manager should have combinations of data sets similarly ranked by their most expected use, which is a function of the primary user's intended analyses (e.g., geology/wetlands - #1, vegetation/wetlands - #2). This combination ranking, when compared to the original individual ranking, may significantly influence the data entry priority of the data sets. In the example, wetlands data were not considered a high individual priority data set (#4), but in combination it became important for priority data entry because the primary and secondary analytical analyses of geology/wetlands and vegetation/wetlands could not possibly proceed without the wetlands data set.

CF #2. Multiple Uses/Users

In general, any data sets that have multiple uses or can be used by multiple users have an intrinsically higher "added" value than data sets restricted to a single use or user. When such is the case, digitizing budgets can often be expanded due to funding participation from multiple user sources.

CF #3. Unit Digitizing Cost

All data sets should be given a rank in relation to their digitizing cost. Digitizing cost is a direct function of the attributes of the data and mechanics of digitizing the data. The most important data attributes are data denseness and complexity and data reliability and accuracy. Although too elaborate a topic to discuss herein, the most important element pertaining to the mechanics of digitizing is the legibility of data to be digitized (i.e., the overall condition of the source material). Data from good source material can be digitized quicker at a lower unit cost. Relative to data attributes, the least complex data to digitize will have the lowest unit cost and permit the most data entry into the GIS for a given budget. However, the above data and digitizing attributes must be examined in concert with their relationship to unit cost by the GIS manager. If the accuracy of a particular data set is known to be questionable, it is assumed that its intended use will, consequently, be severely restricted or used only with extreme caution. In such a case, regardless of the ease (or low unit cost) of digitizing the data set, the value of the entire data set is in doubt, thereby rendering the cost of the digitizing a possible wasted value that could have been applied to another, more reliable, but, perhaps, more complex data set.

CF #4. Data Set Interrelationships

All data sets should be ranked for "stand-alone" usefulness independent of the other data sets. Data sets that are not usable without the availability of other data sets (except perhaps for data sets creating specifically-used visual displays) should be given low rankings. To derive use from such data sets requires the digitization of multiple data sets, thereby increasing unit digitizing time and costs. For example, digitizing manhole cover locations (although of a relatively low cost) can be useless without inputting the entire sewer system map; however, the total cost of digitizing these two data sets may be more appropriately applied to inputting an entire topographic data set as an alternative. Data sets that can be used to infer, imply, or check other data sets by computerized algorithms or interactive user involvement should be given high priority rankings and made more use of to reduce manual editing requirements and streamline quality control of the entire digitization process.

CF #5. Data Sales

Data sets should be assessed and ranked according to their potential for sale to sources beyond the user agency. With the advent of wider use of CAD/CAM systems by private and public organizations, data sharing and sales are already experiencing greater demand. Data sets with the highest potential for revenue generation should be given high ranking and priority for digitization. Added revenue can be used to finance further digitization of the other data sets.

CF #6. GIS Requirements

The GIS manager must rank data sets considering any limitations imposed by the actual GIS (hardware and software) being utilized and the objective of the GIS (i.e., considering Aangeenbrug's GIS categories #1 through #5). Data entry is often constrained by system limitations. For example, some GIS software does not handle small "islands" extremely well, and in such cases, data sets with numerous "islands" should be given a lower priority for digitizing than other "island-free" data sets. Hardware, particularly scanning versus manual digitizers, may also play an important part in ordering data sets as well.

CF #7. Other Data Sources

Occasionally, specific data sets desired for a GIS may be available from a variety of sources but often are at different scales or have other undesirable characteristics. The varying characteristics often compromise the degree of accuracy obtainable

relative to the desired result, and this must be assessed by the GIS manager. An option that is not in widespread use, is the application of computer algorithms to modify existing data sets available from other sources. This modification would be performed in lieu of users digitizing their own original data sets. For example, topographic data tapes available from the Federal Government can be obtained and processed via an interpolation algorithm to refine contour intervals. To a county government, this may be a less expensive alternative than digitizing their own original, larger scale topographic sheets on a countywide basis if the data itself is only intended for use in a Category #3 GIS (i.e., Planning/Evaluation GIS). Data sets are not ranked for this CF, they are simply examined for acquisition from other sources at reduced cost.

The above CFs were not discussed in any order of intended importance nor are actual numerical value rankings (high/low) in each CF suggested. Furthermore, it is acknowledged that there are other CFs or sub-CFs that can be defined and added to the total analysis process. The development of appropriate CFs and entire decision process for management of GIS data base entry is considered to be one that should be individually designed and customized by the GIS manager. The point of this paper, however, was to introduce the application of a systematic management methodology to provide a logical decision process to govern data base entry. The "GIS Data Base Entry Management" concept is a planning tool of major significance that directly influences the amount of data capable of being entered into large-scale GISs within a given budget. With the massive amounts of data to be entered into GISs (which are required to make GISs useful and pay for initial costs), effective data base entry management may become the "Value Engineering" of the entire field.