CONSIDERATIONS IN THE DESIGN AND MAINTENANCE OF A DIGITAL GEOGRAPHIC LIBRARY

Hugh Keegan and Peter Aronson
Environmental Systems Research Institute
380 New York Street
Redlands, California 92373
(714) 793-2853

ABSTRACT

A digital geographic library is a structure for organizing large digital geographic data bases for efficient storage and retrieval. Its successful implementation requires considerable planning and possibly significant file manipulation. The issues associated with the effective design and organization of a digital geographic library are described.

THE NEED FOR DIGITAL GEOGRAPHIC LIBRARIES

Increasing numbers of government, private and academic organizations are assembling large digital geographic data bases. In these efforts, primary attention has been given to data collection, and efficient data automation and manipulation. Secondary attention has been given, when considered at all, to the compatibility and long-term management of digital geographic products. As a consequence, data is either functionally lost from the system because it cannot be retrieved or it is used in a manner which is inconsistent with its attribute and intended use.

Frequently insufficient attention is given to a systematic organization of geographic data based on issues such as thematic content, spatial coverage, precision, resolution, age , scale, classification systems and end use. These factors contribute to inadvertent misuse of geographic data and result in a digital geographic data base that is considerably less useful and reliable than it could be.

In addition, the volume of data assembled to support most organization´s requirements far exceeds its on-line data storage capacity. Data transfer to magnetic tape is typically performed only when storage requirements compel it, and frequently in a haphazard manner.

Documentation of file contents are generally maintained mentally by the tape´s creator. The association between tape documentation, file names, and file contents become increasingly obscure over time. The result is that data becomes unretrievable.

DIGITAL GEOGRAPHIC LIBRARY GOALS

A digital geographic library (map library) rectifies these problems through the imposition of a consistent format for organizing data by location and attribute (for a complete discussion refer to Aronson and Morehouse, 1983; Morehouse, 1985). It provides the user with search, retrieval, and perusal capabilities of both on-line and archived data sets based on both thematic and spatial criteria. The map library gives the user access to map data as well as information about the maps themselves. With a map library, a user is able to browse through digital map library catalogs and locate map data by subject as well as by location. He can determine the map's scale at automation and the tolerances used in its processing. He can determine the maps' age and resolution. He is also able to determine whether a map is in circulation or has been archived to tape. He can leave instructions for the map librarian to retrieve archived material for use. The user is able to make copies of digital maps from unrestricted areas of the map library for his own use. He is not, however, allowed to recopy these maps back into the library after use. Instead, he can nominate maps for addition to the library.

The actions of the user are restricted so that the integrity of the map library is always assured. Decisions about what maps to include in the map library are the responsibility of the library administrator(s).

The map library should enhance the user's access while maintaining data base integrity.

A MAP LIBRARY DATA MODEL

A number of data models can be applied to the organization of a digital geographic library; however, for the purpose of this discussion, the model used in the ARC/INFO MAP LIBRARY (Aronson and Morehouse, 1983), will be used.

In the MAP LIBRARY model, geographic data bases are organized simultaneously in two dimensions -- by subject or content into layers and by location into tiles (see Figure 1).

Tiles. The geographic area represented by the map library is divided into a set of non-overlapping tiles. Although tiles are generally rectangular (e.g., 30 deg. squares), they may be any shape (e.g., counties or forest administration units). Tiles are the digital analogue for the map sheet of a conventional map series. All geographic information in the map library is partitioned by this tile framework.

Layers. A layer is a coverage type within a library. All data in the same layer have the same coverage features (e.g., points, lines, polygons) and feature attributes. Examples of layers are land sections, roads, soil types, and wells. It is useful to think of a layer as a coverage which spans several tiles.

Map Sections. Once a layer has been subdivided into tiles, it consists of a set of individual units called map sections.

In terms of a map library, a geographic data base may be organized as a set of tiles and layers. The tiles are defined in a special INDEX coverage, where each polygon in the INDEX coverage represents a single tile in the library. Layers are defined by the feature classes present (points, lines or polygons) and the thematic items associated with each feature class.

The specification of the tile structure and layer content for a map library which optimally fulfills the requirements of a given user community requires thorough documentation and planning.

MAP LIBRARY CONCEPTS

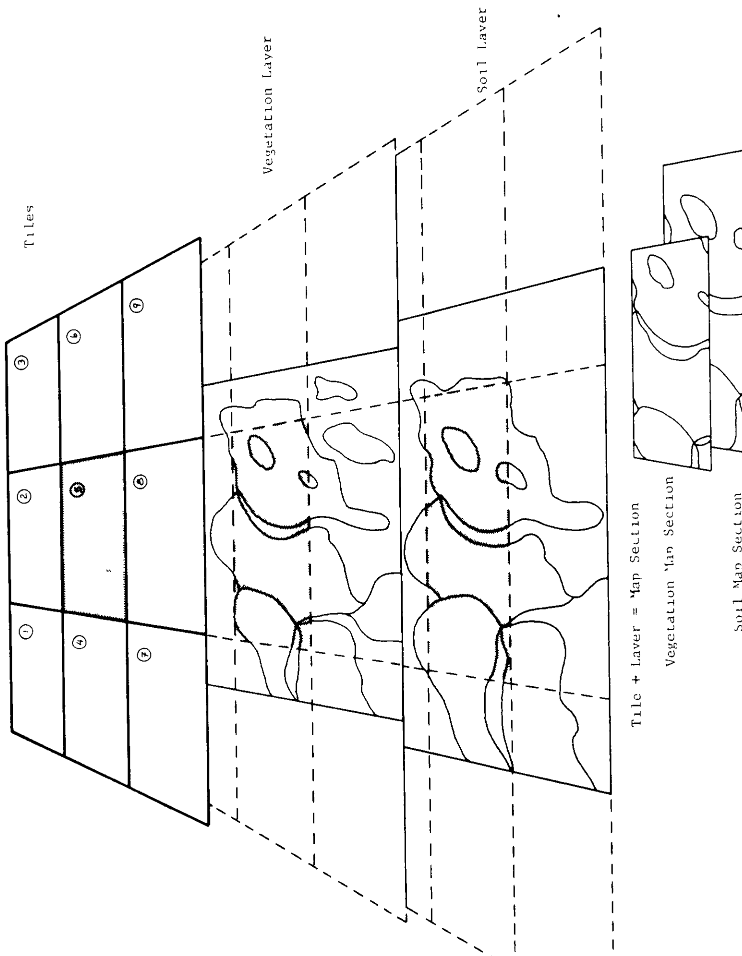

Figure 1

THE MAP LIBRARY DESIGN PROCESS

The principle consideration in designing and implementing a map library is to respond to the typical user inquiry as efficiently as possible. This is essentially an exercise in optimization and compromise. The degree to which an efficient response can be fulfilled is a function of the total amount of disk storage which can be devoted to a library and a sensitive balancing of tile size and layer organization. Both affect retrieval time and subsequent processing time. Additional considerations not to be overlooked throughout the design process are: 1) the cost in both manpower and computer time of implementing the library; and 2) the cost of maintaining the library once created.

Before proceeding, however, it should be pointed out that there are situations when a single map library may be inadequate to fulfill the requirements of a broad user community or prove inefficient for specific user groups. Among these are:

<u>The data base is composed of data differing in resolution and precision.</u>
Data captured at different scales may not be suitable for use together. Although digital maps are essentially without scale, they still have precision and resoultion. If a user has data sets at widely varying degrees of precision, it may be necessary to store them separately. The precision of a map produced by any given process is that of the least precise input into the process, therefore separation into libraries may be warranted.

<u>The data base is composed of incompatible data types.</u>
Separate libraries might be created from a data base including both natural history and detailed taxation data.

<u>The data may be extremely non-contiguous.</u>
When areas of interest are widely separated such as the producing oil fields of the world, it may be desirable to treat them as separate libraries rather than have numerous tiles with no data.

<u>Portions of the data are proprietary or sensitive.</u>
Even though a particular group of data might be well accommodated by an existing library's tile and layer organization, it may be the basis for a separate library if it is confidential in nature.

<u>A user group has highly specialized requirements.</u>
A single group within a larger user community may have highly focused data requirements. Ease of access and processing efficiency may indicate that a separate library is warranted.

With these considerations in mind, the steps in the design of a map library are outlined in Table 1 and discussed in the following sections. It is assumed that extensive data bases are in place.

<u>Step I Assess User Needs</u>

The fundamental step in designing a library is performing a user needs assessment. This is true whether a broad user group is envisioned as the library consumer or a single department. The user needs assesment provides the conceptual basis for defining tile size and layer content in addition to verify that the construction of a library is warranted or

Table 1
Library Design Sequence

Steps

I Assess User Needs:

Quantify volume and type of user requests (simple reference or will data be processed?)
Determine areal extent of typical user inquiry
Document attribute data being processed
- identify logical relationships between data elements
- identify frequency of specific data item utilization
- identify missing data elements

II Evaluate Cartographic and Thematic Data
- scale of data when automated
- mapping resolution
- processing precision
- content of existion files
- spatial organization of coverages
- original intent or use of the data
- classification systems used

III Select Tile Structure and Size

IV Select Layer Content

V Prepare Pilot Study
- select study area
- set up library
- evaluate library construction time
- evaluate retrieval time
- evaluate tile size (too small, too large, just right?)
- evaluate layer contents (determine number of reselects and appends
required in operation)
- quantify manpower and computer resources expended to implement the pilot study and perform at least a rough cost/benefit analysis.

VI Re-evaluate Tile System and Layer Content Based On Pilot Study

VII Implement the Real Thing

feasible. The volume and nature of user interaction with the data is established during this step in the design process.

A. Establish volume of user interaction with the data base. How many different coverages are currently being manipulated and processed? A recording period of at least several months should be used to avoid anomalies in the sampling process.

One should also make an estimate of whether this interaction with the data base would increase if a map library were installed and by how much. This evaluation of use should be scrutinized carefully to determine if the construction of a library is justified.

B. The aerial extent of the user's typical inquiry should be identified. Care should be taken to assure that the current coverage organization does not influence the definition of the area. For example, the fact that data are currently stored in coverages based on state boundaries should not prevent the user from indicating county level definition, if processing at the county level comprises the bulk of user manipulations.

C. Attribute data being manipulated by the user should be listed both by variable and feature type (point, line or polygon). Logical relationships between the variables being used should be identified and recorded. For example, if the user is selecting marshlands from one file and bogs from another these should be grouped together.

The frequency of use of different variables should also be noted, particularly when the same groups of data are being used repeatedly. It is also important to identify variables for which there are only partial thematic coverage. The way the data are being manipulated has direct impact on the way map library layers should be subsequently defined and organized.

Step II Evaluate Cartographic and Thematic Data

At the same time the use of the data is being evaluated the qualitative nature of the data should also be documented.

A. The scale of the data when automated is extremely important to record. It has ramifications on both the resolution of the data as well as its precision.

B. Equally important to record is the current overall precision of the processed digital geographic file. Tolerances used in processing a file may result in a file whose precision is considerably less than when it was first automated. Ideally, tolerances used in processing the original data file should be recorded.

C. Extreme care should also be taken to document compatibility between systems used to classify thematic data. This is particularly true for spatially and temporally separated digital maps. It also applies to documenting spatial fidelity between features recorded at different times for the same location (for a comolete discussion see Chrisman, 1982). Otherwise, the information content of digital maps having the same resolution, precision, and scale can be severely corrupted when assembled as composites or when identical processing operations are performed in spatially separate areas.

D. The boundaries of all coverages currently being used should be explicitly plotted as a composite. This permits the identification of irregularly shaped coverages and coverages which overlap. It also permits islands of non-contiguous coverages to be located. This graphic documentation of existing spatial data base can be thought of as a crude INDEX cover for the later definition of tile shape and size.

Step III Selection of Tile Structure and Size

A good tile definition is one which corresponds to the typical usage of spatial data. If 90% of user requests are 7-1/2′ quad sheet based, then that should prove to be a good tile size. However, if operations frequently result in multiple tiles having to be extracted for processing, a larger tile size should be considered. The goal is to select a tile which is large enough to accommodate the user requirements, but not so large as to cause unnecessary overhead in file transfer and processing.

The selection of a tile structure should be based upon the user needs assessment. It should, however, also be tempered by a realistic evaluation of the potential costs involved in restructuring the original data base, since coverages will need to be either aggregated together or subdivided.

Step IV Selection of Layer Content

Like the tile structure the creation of data layers should reflect typical usage. A layer is a map type or an association of related map types. Examples of a single map type include: streams, deciduous forests, lease areas, and zoning. Associations of types might include: an aggregate of deciduous forests, mixed forest, coniferous forest, land use and waterbodies. A good layer has a logical single nature. If one is conducting a highway transportation analysis, it would not make sense to have streams in the same layer as roads, as one would constantly have to reselect the desired data. Likewise one would want to make sure that roads, highways and interstate highways are all in the same layer rather than extracting them from separate layers and appending them. It should be remembered, morever, that the data in a layer must be sympathetic in feature type; lakes and streams are both water, but one is a polygon layer and the other is a line layer.

Another consideration in constructing layers, particularly when several items may be merged into a single layer, is the volatility of any of the data items. Items which change frequently, such as land ownership or zoning in transition urban areas, should probably be segregated into a single layer to reduce the need to update several layers because of the presence of a rapidly changing elements.

Step V Prepare a Pilot Study

Prior to embarking on a full scale data base conversion effort, a pilot study should be conducted on a limited study area. This allows the library designer to quantify cost, and response time, and to empirically test tile size and structure and the layer organization.

A. The cost of the data base conversion effort should be carefully tracked throughout the pilot study. While some of the energy expended in the pilot data base conversion will be expended in learning conversion procedures, the balance of work will be useful for projecting final data base conversion cost. This may cause a redesign of the library.

B. The appropriateness of the tiling structure can also be evaluated by carefully tracking the number of occurrences where multiple tiles must be extracted to fulfill user requirements. Likewise, if there are never situations in which more than a single tile must be extracted then the tiles may be too large.

C. The layer organization should also be evaluated by determining the frequency with which unwanted variables must be dropped out or appended to include missing items. The user should also perform revisions of the pilot layers to document the ease or difficulty with which changes can be made.

D. A thorough pilot study might include several different tile structures and layer organizations, which should be compared and contrasted with regard to labor in conversion, processing requirements, retreival time, flexibility of usage, and ease of data revision.

Step VI Redesign

It is almost certain that the final map library which is implemented will differ from the pilot study's original specifications. The primary question which should be asked is whether the pilot study library meets the user's typical needs.

Based on an evaluation of the pilot study results, design decision should be revised and incorporated into the final design. The cost of total data conversion may force the reduction of the scope of the library by either increasing tile size and reducing the number of layers or by reducing the aerial extent of the map library's coverage.

Because the end design is both both difficult and expensive to revise once implemented, conducting one or more revisions of the pilot study may be prudent.

SUMMARY

A digital geographic library is a useful organization tool for managing and accessing large geographic data bases. The speed of access to specific thematic data or cartographic data can be enhanced by careful design. The design of a digital geographic library is much simpler if the data are homogeneous in type, scale, precision, resolution, and classification. The design becomes more complex give permutations among these. This may require large dissimilar collections of digital maps to be organized in separate libraries.

REFERENCES

Aronson, P., Morehouse, S., 1983, "The ARC/INFO MAP LIBRARY: A DESIGN FOR A DIGITAL GEOGRAPHIC DATA BASE"., Presented at AUTO-CARTO VI.

Chrisman, R., 1981, Methods of Spatial Analysis Based on Maps of Categorical Coverages, Unpublished Ph.D. Dissertation, University of Bristol.

Morehouse, S., 1985, "ARC/INFO: A Geo-Relational Model For Spatial Information"., To be presented at AUTO-CARTO VII.

ACKNOWLEDGMENTS

The authors wish to acknowledge the kind assistance of Susanne Rohardt and Scott Morehouse for their comments and suggestions.