ARC/INFO: A GEO-RELATIONAL MODEL FOR SPATIAL INFORMATION

Scott Morehouse
Environmental Systems Research Institute
380 New York Street
Redlands CA 92373

ABSTRACT

A data model for geographic information is described. Originally designed for thematic mapping and map analysis, the model lends itself to tabular data processing applications as well as automated cartography. The model is a combination of the topological model (to represent feature locations and topology) and the relational model (to represent feature attributes).

GOALS FOR THE DATA MODEL

A geographic information system is a spatial data base together with a set of spatial operators. Any spatial data base is derived from a model of geographic information. The usefulness of a geographic information system depends on having a data model appropriate for geoprocessing. This is particularly true when systems and data bases must serve a variety of purposes. The ARC/INFO data model was designed as the basis for a generalized geographic information system. The overall goal is for a practical data model with as much generality as possible. Specific goals are described in the following paragraphs.

Generality. The data model should support data bases developed at a variety of scales and for a variety of purposes. It should be suitable for applications ranging from thematic mapping to land inventory to topographic mapping to urban base mapping.

Simplicity. The data model should be as simple as possible and still meet its other goals. A simple data model is the key to implementing efficient and reliable geographic data bases and algorithms.

Efficiency. The data model should provide the basic data structure for all geoprocessing functions. It should support efficient geoprocessing functions directly without requiring conversion of data to special "analysis" or "edit" formats. For example, a function such as polygon overlay should be done directly with the data model rather than requiring a grid cell copy of the data base. Functions which should be easily implemented using the data model are: bulk digitizing, high quality map graphics, polygon processes such as overlay and dissolve, non-graphic query and analysis, and network simulation.

Adaptability. It should be possible for both the system user and system programmer to extend or adapt the data model for particular applications. This is especially true for

feature attributes. It should be easy to add new attribute information to an existing data base. More importantly, it should be possible to relate information in the geographic data base to existing non-spatial data. Geoprocessing applications often require more than just "map data" -- other non-spatial data are usually involved (e.g., soil interpretation matrices for land planning or property ownership records for urban planning).

Freedom from Restrictions. The data model should contain no inherent limitations on its size or content. It should handle both very large and very small databases well. Any limits that are placed on data volume or content (e.g., maximum 2000 points per polygon) will soon be challenged and require messy adaptation to application programming and system use. The absence of restrictions is particulary important since the data model is intended for very large production applications.

DERIVATION OF THE DATA MODEL

The ARC/INFO data model is based on the idea that geographic data can be represented as a set of features. Each feature has associated locational and thematic data. For example, if our features are cities, then their locational data might be latitude/longitude coordinates; the thematic information might be population, area, etc.

Early in the design of ARC/INFO, it became evident that data structures optimal for the analysis of locational data were not optimal for thematic data. In addition, it was clear that these two views of geography, features in space and features with thematic attributes, are both equally important. It is a mistake to think of thematic data merely as attribute codes tagged onto the end of the coordinate definition of a feature. It is equally incorrect to regard locational data as yet another item in a thematic data base management system.

For this reason, ARC/INFO was designed using a hybrid data model. Locational data are represented using a topological data model (similar to the USGS Digitial Line Graph, USGS 1984). Thematic data are represented using a tabular or relational model. In the name "ARC/INFO", "ARC" refers to the topological data structures and algorithms, "INFO" to the tabular data structures and algorithms, and "ARC/INFO" to the composite data model and associated processes. The data model is a geo-relational as it combines a specialized geographic view of the data with a conventional relational data model.

The topological model was chosen to represent locational data because it has a sound theoretical and practical basis and also met the goals outlined above. It has been studied theoretically (see, for example, Puecker and Chrisman 1975, White 1980, and Corbett 1979) and has served as the basis of a number of successful systems: DIME, GIRAS (Mitchell et al 1977), and ODYSSEY (Morehouse 1982). It has proven useful for a variety of applications, ranging from address matching and network flow simulation to detailed storage of map data.

A number of other well-understood spatial models were rejected because they could not meet the overall system goals. Grid cell encoding and its variants, such as quad trees, were rejected because of their inability to handle large amounts of data with precision. Conventional polygon encoding, as implemented in MOSS (WELUT 1982), was rejected because it is inefficient for many geoprocessing functions (e.g., polygon overlay). Graphic element encoding, used in computer-aided design systems (see, for example, SYNERCOM 1982), was rejected for similar reasons even though this structure is well adapted for interactive graphic editing.

For thematic data processing, the relational (or tabular) model was chosen for its adaptive and simple characteristics. This model is the subject of extensive theoretical investigation. In addition, it has been successfully implemented in a number of systems (SAS, INFO, ORACLE, and others). I prefer to use the term "tabular data model" here rather than the more fashionable "relational data model" because the term "relational" has a more restrictive usage. The conventional statistical matrix as implemented in SAS, for example, illustrates the power and utility of the tabular model. However, SAS is not strictly a relational data base. Another trend which enhances the utility of the tabular model is the emergence of "fourth generation programming languages". These programming languages contain program statements for screen input, report generation, file sorting, and merging and for computation of new record values. The INFO system, used by ARC/INFO for tabular data processing, is one of these.

An underlying strategy in the design of the ARC/INFO data model was the use of existing data models. ARC/INFO can thus benefit from continuing technological advances in these areas.

THE DATA MODEL

The Coverage

The coverage is the basic unit of data storage in ARC/INFO. A coverage is analogous to a single map sheet or separation in conventional cartography. It defines the locational and thematic attributes for map features in a given area.

A coverage is defined as a set of features, where each feature has a location (defined by coordinates and topological pointers to other features) and, possibly, attributes (defined as a set of named items or variables).

Feature Classes

There are several kinds of features that may be present in a coverage. Each of these feature classes may have associated locational and thematic information. Figure 1 shows some of the feature classes that may be present in a coverage.

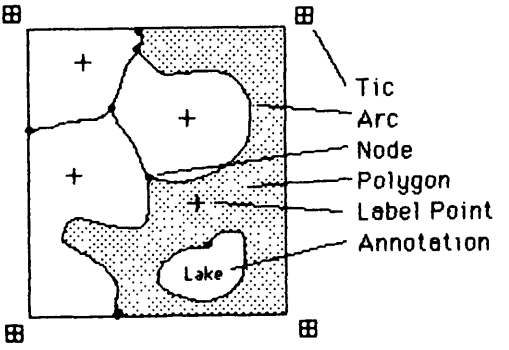

Figure 1: Feature Classes in A Coverage

Tics are registration or geographic control points for the coverage. They allow the coverage to be registered to a common coordinate system (e.g., UTM meters, latitude / longitude, etc.)

Arcs are lines used to represent the location of linear features (e.g., roads) and the borders of area features. Arcs may be topologically linked to the nodes at each end of the arc as well as to the polygons on each side of the arc. Also, an arc attribute table may be created, defining values for thematic attributes (e.g., type of stream, length, etc.).

Nodes are points at the end of arcs. Nodes may be topologically linked to the arcs which meet at the node. A node attribute table may be created, defining values for thematic attributes (e.g., section corner type).

Polygons are areas enclosed by arcs. All polygons are defined by topological pointers to the set of arcs which compose the polygon border and the set of label points inside the polygon. A polygon attribute table may be created, defining values for thematic attributes (e.g., section number, area, etc.).

Label Points are points used to represent information such as wells. Label Points are also used to associate thematic data with polygons and for positioning text within polygons. When used for polygon labelling, label points are topologically linked to the surrounding polygon. A point attribute table may be created, defining values for thematic attributes (e.g., well type).

Annotation is a set of text elements used to label the map on hard copy plots. Place names appearing on a map sheet are recorded in the coverage as annotation elements. Annotation is used solely for display purposes; it is not used in analysis processes such as overlays or section subdivision. Annotation may be topologically linked with features represented.

These feature classes are the basic vocabulary used to define

geographic information in a coverage. By varying the types of features contained in a coverage and the thematic attributes associated with each feature class, the coverage can be used to represent many types of map information. For example, the general land office grid can be represented as a set of arcs (section lines), nodes (section corners), and polygons (sections). Thematic attributes can be associated with each of these feature classes.

Feature Attribute Tables

In theory, each of the feature classes described in the preceding sections can have an associated feature attribute table. In practice, feature attribute tables have been implemented for arcs, label points, and polygons. The node attribute table is in the process of being implemented. All attribute tables have the same general structure (see Figure 2).

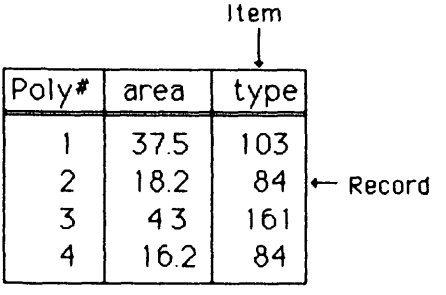

Figure 2: Structure of an Attribute Table

In ARC/INFO, the rows of the table are called records and the columns are called items. There may be one table for any feature class in the coverage. Within the feature attribute table, there is one record for each feature of that class. All records in the table have values for the same set of items (or thematic attributes). Items are defined by type and the number of bytes used to store the item.

The feature attribute tables are an integral part of the coverage and are processed by ARC/INFO commands which affect the coverage. For example, when two polygon coverages are overlaid to create a new composite coverage, the polygon attribute tables of the input coverages are merged and written as the polygon attribute table of the output coverage.

In addition to the feature attribute tables, the user can define any number of additional attribute tables. These tables can be related to the feature attribute tables and each other in a variety of ways. For example, the polygon attribute table for a forest stand map could simply contain a single item, the stand identifier. There could be a separate table containing detailed information for each stand in the entire forest. This would allow the stand attribute

information to be assembled and maintained independently of the stand maps, but still allow data from both to be related for any application.

The Workspace

A workspace is simply a directory which contains one or more coverages. In addition to the coverage locational data, the workspace has an INFO data base containing the coverage attribute tables and any other, related, attribute tables. Workspaces provide a convenient means for organizing coverages into related groups. They also provide a place for the storage of tabular data not directly tied to a coverage (soil interpretation matrices, for example). Each workspace is completely independent. However, ARC/INFO processing commands allow coverages from different workspaces to be used together. The decentralized organization of coverages in workspaces allows an unlimited number of workspaces (and coverages) to be managed on a single system.

The Map Library

A centralized data structure is useful for the management of very large geographic data bases. The map library is a device which allows coverages to be organized into a large, complex geographic data base. Coverages are organized simultaneously in two dimensions -- by subject or content into layers and by location into tiles (see Figure 3).

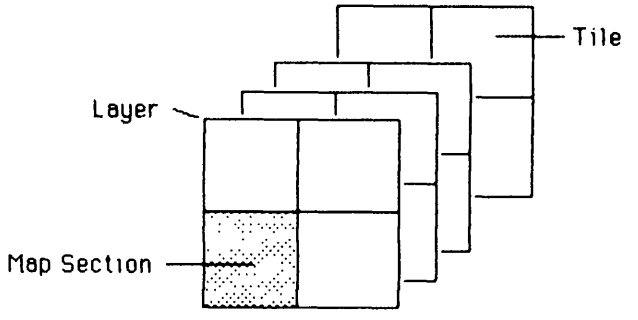

Figure 3: Structure of a Map Library

Tiles. The geographic area represented by a map library is divided into a set of non-overlapping tiles. Although tiles are generally rectangular (e.g., 30° squares), they may be any shape (e.g., counties or forest administration units). Tiles are a digitial analogue for the map sheets 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 the library. All data in the same layer have the same coverage features 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 crosses several tiles.

Map Sections. Once a layer has been subdivided into

tiles, it consists of a set of individual units called map sections. A map section is simply a coverage as defined previously.

In terms of the map library, a geographic data base is defined 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 in the same way as coverages, by defining the feature classes present and the thematic items associated with each feature class.

CONSTRUCTING A DATA BASE USING THE DATA MODEL

In order to develop a functional data base, geographic information must be adapted to the concepts and vocabulary of the data model. For example, roads are represented by arcs; lakes by polygons; and so on. Feature attributes are recorded in the coverage feature attribute tables. A critical part of data base design using this model is determining which geographic features are stored in the same coverage or layer. For example, should single-line drainage be stored in the same layer as double-line rivers and lakes? These questions can be resolved by considering the use for the data base and by striving to simplify the feature attribute tables associated with various layers. Issues raised in this aspect of data base design are similar to those encountered in the design of relational data bases (see the discussion of normalization in Kent 1983, for example).

As an example of an ARC/INFO data base design, a simplified design for the information contained on a USGS 1:24,000 quad sheet might be:

```
Layer:  General Land Office land net
      arcs:  section lines
             items: line type (e.g.,township border)
     nodes:  section corners
             items: type (surveyed, photo identified)
  polygons:  sections
             attributes: township, range, section, meridian

Layer:  Hydrographic Lines and Points
      arcs:  drainage lines
             attributes: type (intermittant, channel),
                         stream hierarchy, river mile index
    points:  wells, gage stations, etc.
             attributes: type

Layer:  Hydrographic Polygons
      arcs:  shorelines
  polygons:  lakes, rivers, mudflats,etc.
             attributes: type (lake, flat, river,etc.),area

Layer:  Transportation Network
      arcs: roads, railroads,transmission lines,etc.
            attributes: type,class,DOT segment
```

```
Layer:  Topography
      arcs: contour lines
            attributes: elevation
    points: spot elevations
            attributes: elevation
```

Several points can be made here. First, by separating hydrographic lines from polygons and by adding channels through the polygons, network analysis models can be used to easily determine upstream/downstream relations and to compute paths between points in the river network. In addition, because both the hydrographic lines and transportation network are simply arcs, the same application code can be used in both cases. This is also true for digitizing, editing, plotting, and all other GIS functions. Because each function simply operates on abstract coverage features (e.g., arcs, nodes, etc.), there is no need for specialized data-dependent logic.

A second point in this example is the inclusion of pointers to non-map data in the geographic data base. Every road arc has a DOT segment number which relates to a statewide transportation data base. This allows information carried in the external data base to be easily related to the cartographic features on the map.

APPLICATION OF THE DATA MODEL

The ARC/INFO data model has been designed to complement and support the geoprocessing functions of the ARC/INFO system. This is in contrast to most other geographic data models, which have been designed either as abstract "models of space" or as standardized repositories for map data. The applications which presently operate using the ARC/INFO data model are outlined in the following table (ESRI 1984).

Input Functions
- interactive digitizing and editing
- read DLG, GIRAS, DIME, and SIF
- build topology from unstructured data
- input screens for tabular data
- interactive edge matching
- coordinate geometry (COGO)
- GLO legal description
- grid cell conversion
- interactive districting

Analysis Functions
- map projections
- polygon overlay
- point in polygon
- line in polygon
- feature selection
- feature removal
- polygon aggregation
- ajust coordinates to control
- section subdivision
- network allocation
- minimum path calculation
- compute buffer zones

relational operators for tabular data

Output Functions
mapping based on feature attributes
- point, line, area symbols
- automatic text placement
- key and legend generation
report writer for tabular data
DLG files
convert to grid cell
interactive map query

Data Management
map sheet split
map sheet merge
map library management

The ARC/INFO geographic information system is presently in use at over 60 sites. In addition to data bases developed by other users, Environmental Systems Research Institute has developed a number of large data bases using the data model during the last three years.

UNEP/FAO World Data Base
area: world
scale: 1: 5,000,000
layers: coastline, FAO soils, FAO agro-ecological zones
size: 1.2 million points; 25,000 polygons

UNEP/FAO Africa Data Base
area: continent of Africa
layers: coastline, elevation, terrain units, rainfall, rainfall, wind velocity, hydrography, roads administrative units, railroads
size: 1.4 million points ; 53,000 polygons

Illinois Lands Unsuitable for Mining Program
area: State of Illinois
scale: 1: 500,000 (areas at larger scales)
layers: terrain units, hydrography, administrative units, oil and gas pipelines, transmission lines, coal resources, land net, roads, surficial deposits
size: 1 million points; 70,000 polygons

Alaska Land and Resource Mapping Program
area: Alaska - 80,000 sq. miles
scale: 1: 250,000
layers: terrain units, hydrography, administrative units, infrastructure, energy and mineral resources, elevation provinces, historic sites
size: 1.8 million points; 62,300 polygons

North Slope Borough Resource Inventory
area: Alaska - 97,000 sq. miles
scale: 1: 250,000
layers: terrain units, hydrography, administrative units, infrastructure, energy and mineral resources, elevation provinces, historic sites, North Slope Borough planning data, subsistence land use
size: 1.4 million points; 56,600 polygons

REFERENCES

ESRI, 1984, ARC/INFO Users Manual, Environmental Systems Research Institute, Redlands CA

Corbett, J. 1979, Topological Principles in Cartography, U.S. Bureau of the Census, Washington

Kent, W. 1983, "A Simple Guide to Five Normal Forms in Relational Database Theory", Communications of the ACM, Vol. 26, No. 2, pp. 120-125

Mitchell, W. et al 1977, GIRAS: A Geographic Information Retrieval and Analysis System for Handling Land Use and Land Cover Data, U.S. Geological Survey Professional Paper 1059

Morehouse, S. and M. Broekhuysen 1982, ODYSSEY Users Manual, Harvard Graduate School of Design, Boston

Puecker, T. and N. Chrisman 1975, "Cartographic Data Structures", American Cartographer, Vol. 2, No. 1

SYNERCOM, 1982, The Synercom Information and Mapping System INFOMAP, Synercom Technology Inc.

USGS, 1984, USGS Digitial Cartographic Data Standards: Digitial Line Graphs from 1:24000-Scale Maps, U.S. Geological Survey circular 895-C

WELUT, 1982, MOSS Users Manual, U.S. Fish and Wildlife Service, Western Energy Land Use Team, Fort Collins

White, M. 1980, "A Survey of the Mathematics of Maps", Proceedings of AUTO-CARTO IV, Vol. 1, pp. 82-96