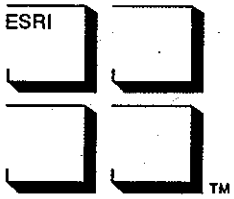
Environmental Systems Research Institute, Inc.



Overview of GIS Database Design

■ 380 New York Street, Redlands, California 92373 ■ Telephone: (714) 793-2853 ■ Fax: (714) 793-5953 ■

Environmental Systems Research Institute, Inc.

ESRI

ARC News

Reprint
Spring 1989 Issue
Vol. 11, No. 2

Overview of GIS Database Design

by Don Chambers, ESRI

A geographic information system (GIS) is comprised of several elements, including hardware, software, users, procedures, and data. GIS organizations select the hardware and software that meet their needs, the staff members are trained, procedures are developed, and the organization evolves so that the new technology is incorporated in day-to-day activities. However, after procurement and transition into the technology occurs, the ultimate success and ability of the system to provide decision makers with quality information depends in part on the quality and usability of the data that reside in the system.

Many users are beginning automation of commonly used data sets, either in house or through a vendor, immediately after installation, or sometimes even before a specific system is chosen. This often occurs because users are pressured by management to demonstrate results that have been obtained with the newly acquired system. The consequence of this pressure is often hasty action that can potentially lead to a series of ill-conceived data files rather than a well-integrated comprehensive GIS database designed to meet user requirements.

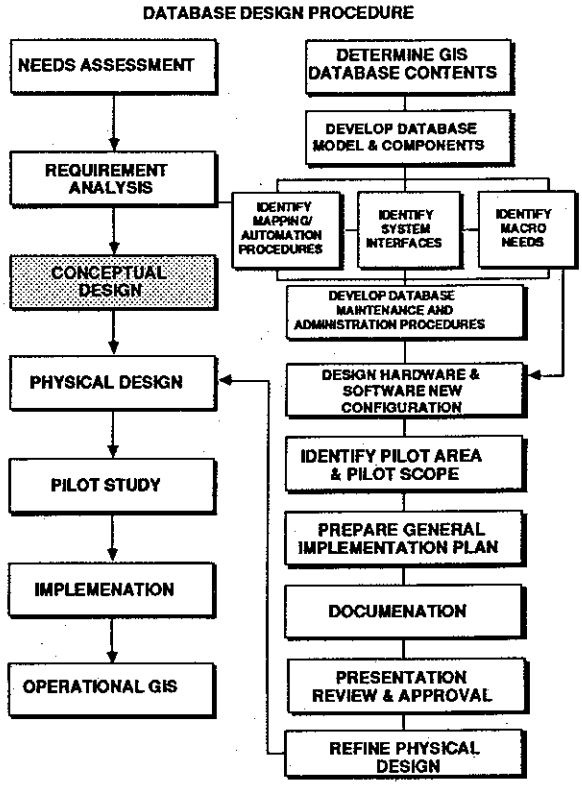
A more systematic approach to GIS implementation is recommended. This approach involves assessing user needs and requirements, developing a database design based on these needs, and testing the design in a pilot study before production automation begins.

This article is an introduction to these various components.

What Influences a GIS Database Design?

Some of the major factors that influence a GIS database design include the data needs of the applications that will be developed, availability and format of existing data required to support the applications, update and maintenance procedures, size of the database, hardware platform/configuration, number and sophistication of users, organizational structure of the users and facility, schedule, budget, and management support.

It is important to understand these factors before automation begins. For small organizations, or for specific projects, an informal series of brainstorming sessions and review of what has worked elsewhere can often clarify what the GIS will be used for and what products it is expected to produce. It will also be necessary to review data sources and identify how the data will be maintained, as well as develop schedules and budgets. This may take from two days to two weeks, depending on the size of the organization and sophistication of intended GIS applications. Simple documentation of agreements, responsibilities, and action items may be all that is required.



For large organizations, or those in which the GIS must support a large diversity of users, a more formal approach is desirable. It is important in this type of setting for a single individual or small group to be designated as responsible for the design. It is their responsibility to ensure that users are identified and involved, appropriate information is collected, and sufficient documentation is developed that will allow the database design to be conducted. This task usually requires a more formal and structured process than the one required for the small organization or project. Interviews with management, users, and existing system support staff are necessary. This process may take from two weeks to several months to complete. Documentation needs to be comprehensive so that the large number of potential users can sign-off on the proposed usage and direction of the system, as well as understand their roles and responsibilities during implementation.

Whether you are from a small or large organization, this initial identification of requirements and constraints must be done. It can be simple or complex, informal or formal, but it must be done. The database design that results from this step will be based on real needs and the final database will more efficiently and accurately support the organization in the future.

What Needs To Be Designed?

Five primary components need to be designed for an ARC/INFO database: the cartographic layers, feature attribute tables, lookup tables, annotation (not discussed in this ar-

ticle), and the map library. While designing these components to meet user requirements, the underlying goals are to maintain data consistency/integrity, reduce data redundancy, and increase system performance while maintaining maximum user flexibility.

Layer Design

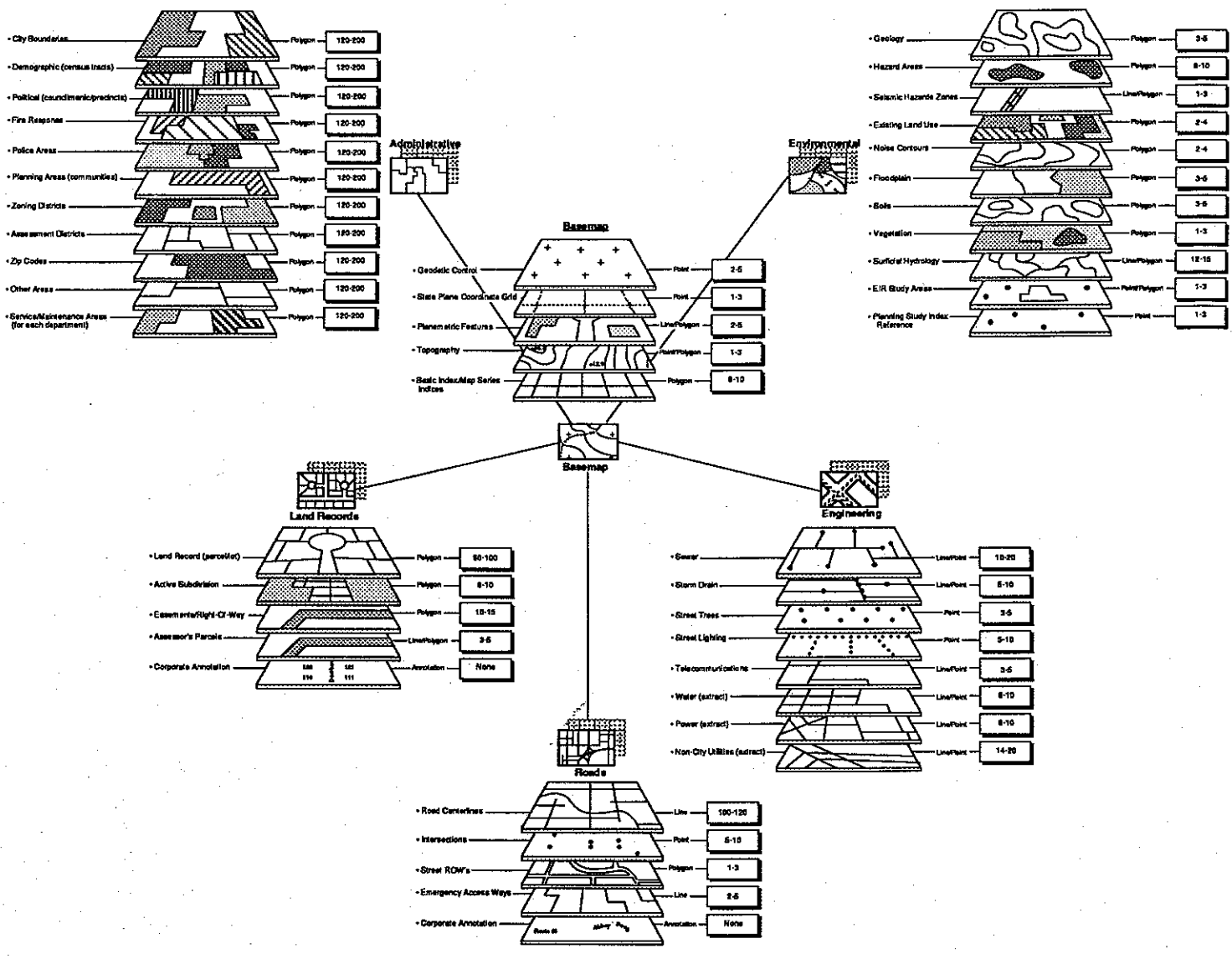
There are three basic layer (coverage) types in ARC/INFO and two variations. Basic layer types are polygons (soils, parcels), lines (streams, street centerlines), and points (eagle nests, manhole covers). Variations on these layers include network coverages that contain polygons and lines (such as roads and blocks) and link coverages containing lines and points (such as roads and street intersections). In future releases, these layer (coverage) types will support more complex combinations of spatial features that can be constructed from basic primitives. Selection of the correct layer type for a database depends on anticipated uses of the data and often depends on the scale and resolution of source data. A stream may be a line at 1:250,000 scale, but a polygon at 1:24,000 scale. An archaeological site may be a point at 1:100,000, but a polygon at 1:24,000.

Many factors influence which data sets should be combined into ARC/INFO layers. Two of the most important are data to data relationships and data to function relationships.

Data to data relationships: Many manually drawn maps have common or coincident lines between them. Examples include land-water interface, survey section lines and roads, streams and political boundaries, parcels, and zoning. In addition, many data sets share common boundaries through hierarchical relationships. An example is a boundary shared by a county, state, and country. It is important to recognize these relationships during the design process and to ensure that these relationships are organized for consistency, minimum data capture, and ongoing data management.

There are four principal methods for capturing these relationships: pre–automation data preparation, where data is integrated into one layer before digitizing (e.g., census geography); creation and use of templates, where data from one automated layer (e.g., land-water interface) are used as a spatial template for use in other layers; automatic snapping one feature to another in ARCEDIT; and copying and moving features from one coverage to another. Each method has its own use, depending on the size of the database, number of coincident features, and where in the automation process the technique is used. The establishment of procedures that manage and update these relationships automatically is both important and possible and should be considered at the design stage.

Data to function relationships: Some types of geoprocessing appli-



cations in an organization may dominate others. The database design should reflect these priorities. A database that must support a wide variety of user views and applications should have a more basic and primitive organization. Data that are not coincident, but tend to be used together, should be considered for inclusion in the same feature layer. Data that are maintained by different departments, have security restrictions, or are frequently updated should be isolated into separate layers.

Feature Attribute Table and Lookup Table Design

Just as important as the cartographic database design is the tabular database design. The way tabular data are organized in a relational database management system (RDBMS) has a tremendous impact on system performance. In designing the feature attribute tables and noncoverage related files, or lookup tables, it is necessary to anticipate uses as well as update procedures.

Similar to layer design, the goal of tabular database design is to create files that are easy to maintain, update, modify, and protect. For an RDBMS, this is accomplished through a process called data normalization, which provides guidelines and rules for organizing data into a series of tables that are related to each other by common keys.

Normalization requires thorough knowledge of the data and its relationships. Using RDBMS operators, normalized tables can then be cut and pasted to form new relationships. This is accomplished in part by assigning a simple key or geocode to all cartographic features in the coverage, and having all descriptive attributes about the feature in separate lookup or "related" tables.

For example, in a parcel coverage, each parcel can be identified with its unique parcel number, the common key. Descriptive attributes about each parcel are then maintained in a RDBMS such as ORACLE, INGRES, or SQL/DS. Another table may provide detailed information about the owner of each parcel, such as address and telephone number.

The new RDBMS packages that can be linked through RDBI to ARC/INFO are powerful and flexible. As an organization's database and applications grow, new records can be easily added. Using RDBMS operators, these tables can then be cut and pasted to form and define new relationships. The separation of the geographic database (ARC) from the RDBMS increases performance and allows users to separate the simple daily transaction updating of the tabular data from the more complex and less frequent updating of cartographic data.

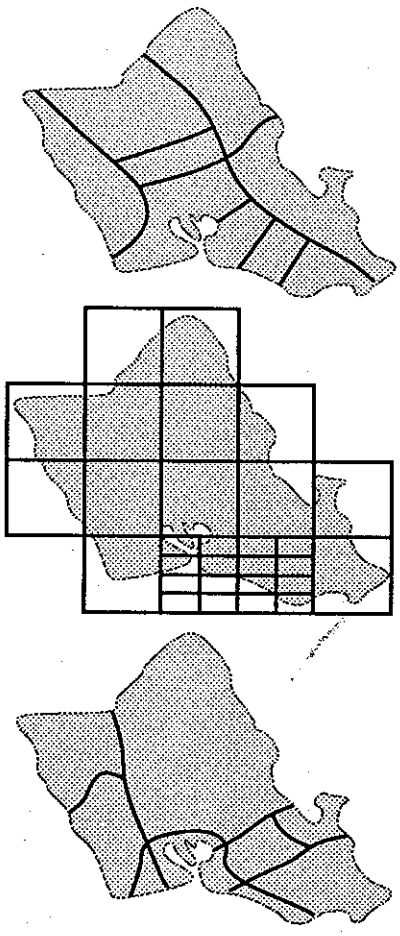
Map Library Design

ARC/INFO map libraries are a useful mechanism for organizing coverages under the following conditions: large amounts of standardized data must be handled; the database will be used for a long period of time; and general systemwide access by many users is desired.

Map library design must be considered because the selected structure will affect database maintenance, data query, and system performance. Different designs used for storing the same data can have highly variable requirements for disk storage. The two primary elements that influence these factors are the tile structure and map projection.

Although tiles may be any shape, there are several general guidelines that should be considered when choosing a structure.

Alternative Tile Structuring Schemes



First, tile boundaries should be stable because the tile structure is the map library component that is most difficult to change. Physical objects, such as watersheds, are a better choice for the boundaries than political units, which may change over the life span of the library. Abstract grids, such as U.S. Geological Survey quadrangle boundaries, are stable and used in many databases as tile boundaries. Regular grids have the added advantage of increasing performance during certain operations, such as searches. If there are geographic units on which analysis or updates are commonly conducted, that unit would make a logical tile structure.

Second, the tile structure should enhance data access performance. Tiles should contain approximately the same amount of data, and should not have more data than can be readily manipulated in an ARCEDIT session. Not all data need to be stored in the map library structure. If data are found to be sparse during automation, or if project specific data are developed for a small area, that data may be better left in a single coverage.

Most large users of ARC/INFO make use of the librarian function for managing their database collections. Reasons for this are many, but are largely associated with performance and shareability of data. The principles of the LIBRARIAN function involve physically managing the cartographic data in a

series of layers and tiles with relational indexes integrating the maps to a seamless tabular database (e.g., SQL DBMS).

This approach provides high performance, particularly for operations dealing with updating and spatial clipping. The LIBRARIAN technology (particularly with version 5.0 of ARC/INFO) allows users to view the database as if it were physically continuous. To obtain the best performance, however, we must still consider the optimum tiling size and configuration.

ARC/INFO is capable of storing all data and their relationships in a continuous cartographic coverage. Recent project work at ESRI involved organization and use of this type of coverage of the United States with nearly one million spatial features and fast query access. This application is designed for network analysis and query only. Therefore, ARC/INFO's spatial indexing tools have been implemented for providing the users with very fast spatial query and analysis functionality.

Given a more diverse application setting, particularly where updating and spatial clip/map topology reorganization is necessary or desirable, the spatial tiling approach provides far better performance and flexibility.

In any case, it is important to design any tiling structure in the early stages of the database design, and consider very carefully the applications that will be served to ensure performance and functionality.

Database Design Documentation

It is difficult to overstress the importance of adequately documenting the database design and subsequent implementation efforts. Documentation is necessary if users are to have confidence in the data, and if the database is expected to remain functional beyond the tenure of those who originally conceived and built it. At a minimum, documentation should include a comprehensive data dictionary with descriptions of all items and codes for each layer. Ideally, the data dictionary is implemented on line and linked directly to the database. An on-line data dictionary supports development of legends for plot series and enhances quality control operations by providing input into routines used during automated attribute checking (using CODEFIND and CONSIST).

Beyond the data dictionary, documentation may also include diagrams and discussions explaining the concept and content of each layer and map library; data sources for all layers and attributes; and implementation procedures, including processing tolerances. If this supporting procedural and data source information is also on line, users can more readily assess the appropriateness of particular layers for a given application.

Pilot Study

Database designs and implementation plans more often than not require modification when tested under production conditions. As a result, a pilot study that incorporates all layers is strongly encouraged before users embark on large-scale data development work. Pilot studies are the implementation of database designs over limited geographic areas. They yield several benefits, including the following: (1) testing of physical database design performance, (2) development of procedures for performing tasks under production conditions, (3) identification of obstacles to system implementation, (4) development of specifications for contracting data loading efforts, and (5) yielding timely results or products for management presentations and gaining continued management support.

Several guidelines should be followed in a pilot study. To begin, the sample site must be representative of the entire study area and exhibit a full range of complexity. This will help ensure that lessons learned during the pilot can be extended to the remainder of the database. If a single, representative area cannot be identified, it may be necessary to select more than one. Another consideration is the scope of the study. To be effective, the applications and processes being tested must be well-defined and should be completed over a three- to six-month period. This will help to ensure that results of the pilot provide feedback and are readily interpretable and meaningful. A "peer review" of results should be conducted with major users of each layer and application type. Finally, it is important to document peer review comments and ensure that they are incorporated into the final database design. Unless they are documented, pilot results will become increasingly vague and meaningless over time.

Who Should Do the Database Design?

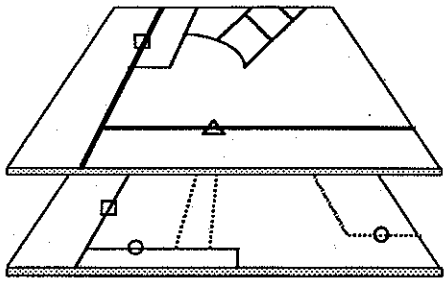
Database design is a process; it is done by a team with strong leadership. The process can be done in house, if there are people with experience, or with consultants. Most people prefer to do it themselves and learn as they go, but there are risks in doing this. There are few people with the academic background and practical experience in the wide variety of disciplines necessary for comprehensive design. Generalists, such as planners, geographers, and landscape architects with a background in GIS, make good design team leaders. They have the necessary tools to synthesize the variety of information required for a design. Depending on design requirements, it is important that they also be supported by topical specialists in fields such as cartography, surveying, planning, forestry, geology, biology, botany, economics, emergency preparedness, and facilities management.

The users must be fundamentally involved. They know the applications and data and have a lot of common sense about what they want to do. If your design seems too complicated, it probably is. Make it simple and clear. Remember, the next step is to implement the design. It must be kept simple to manage it effectively. Vendors or in-house staff must also be able to understand the design to implement it.

It is necessary to review data source documents and consider the intended applications and structure of the organization in which the GIS will reside. Successful design requires acquisition of these components and immersion of oneself in the information. Close examination of source maps and overlaying of the various thematic maps, basemaps, and aerial photographs on a light table help the designer know the data. The database designer must become an expert in the data and its intended uses. Only after this can good design begin.

The following pages show database layer diagrams created as part of a database design project performed for the city and county of Honolulu.

Network Facility Layers



Water Distribution Facilities
(Link)

Wastewater Collection Facilities
(Link)

Feature Attribute Tables

Water Distribution AAT
- Feature ID
- Type
- Size
-Material

Water Distribution XAT
- Feature ID
- Fixture Type

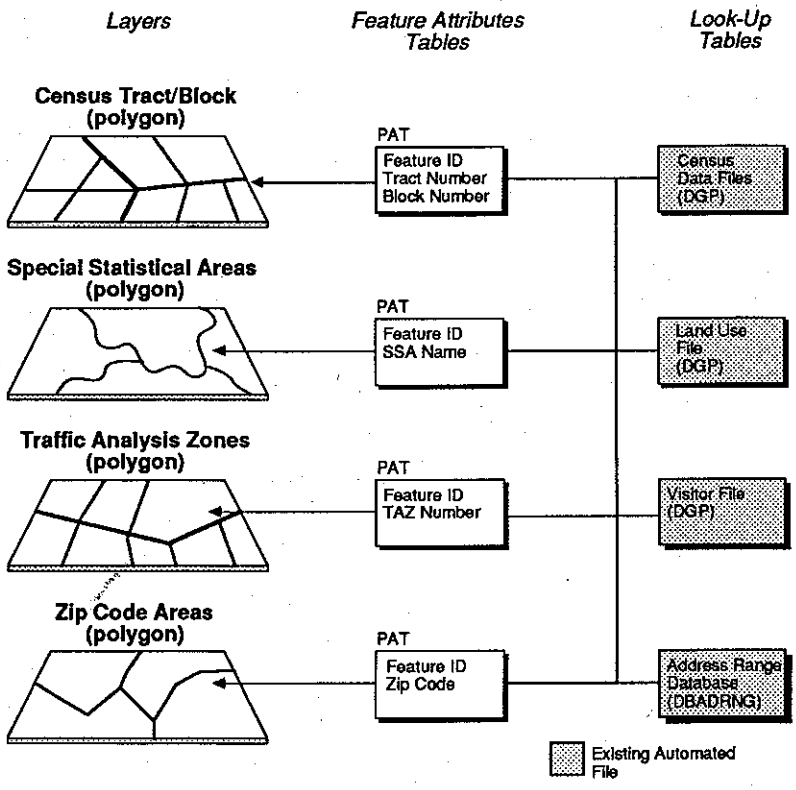
Wastewater Collection AAT
- Feature ID
- Wastewater Collection Feature ID
- Type
- Operator
- Size
- Material
- Slope

Sewer XAT
- Feature ID
- Fixture Type

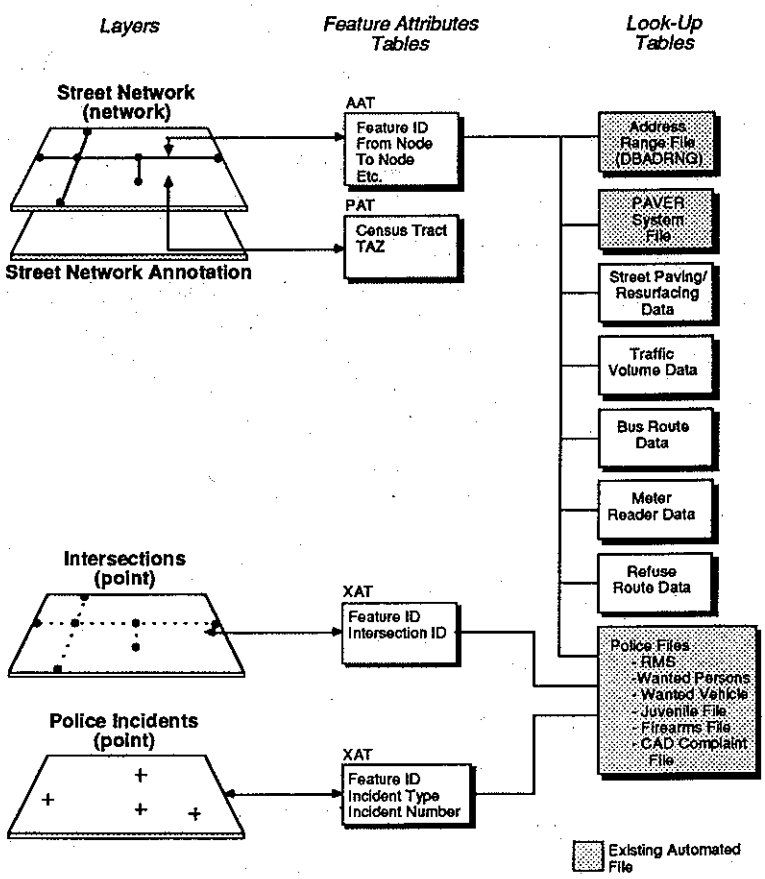
Lookup Tables

Many lookup tables can be developed for these data

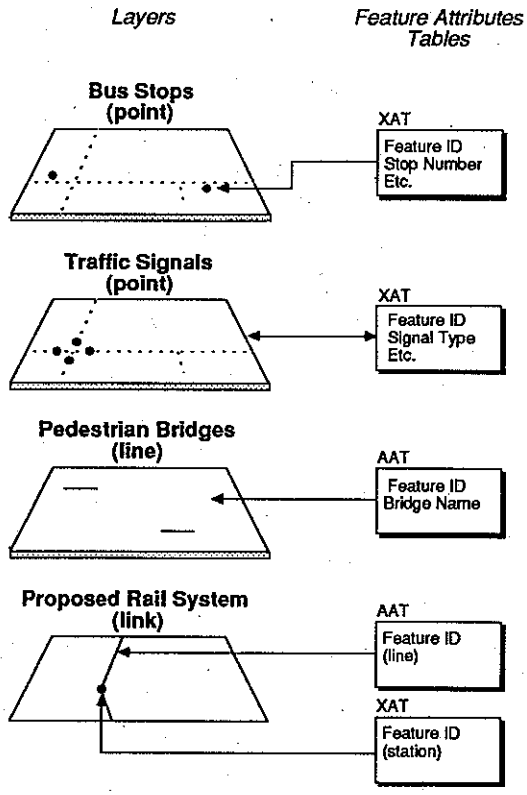
Statistical Area Layers



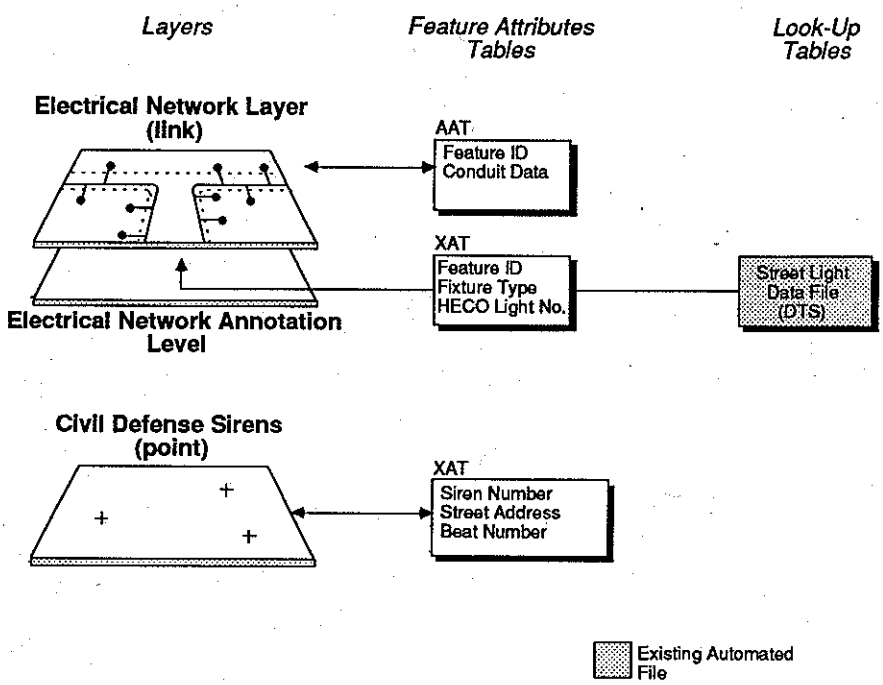
Street Network Layers



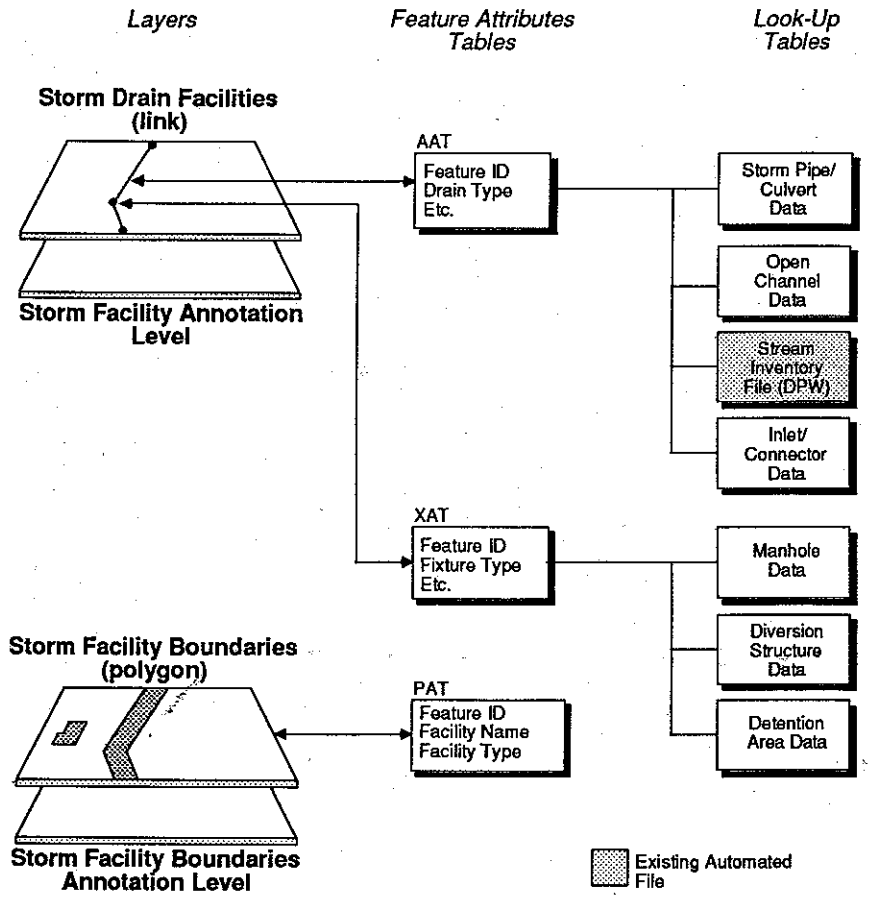
Other Transportation Layers



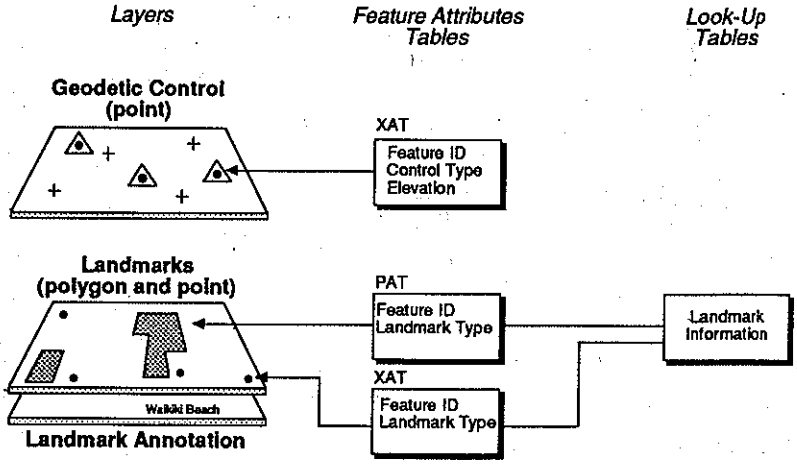
Electrical Facilities Layers



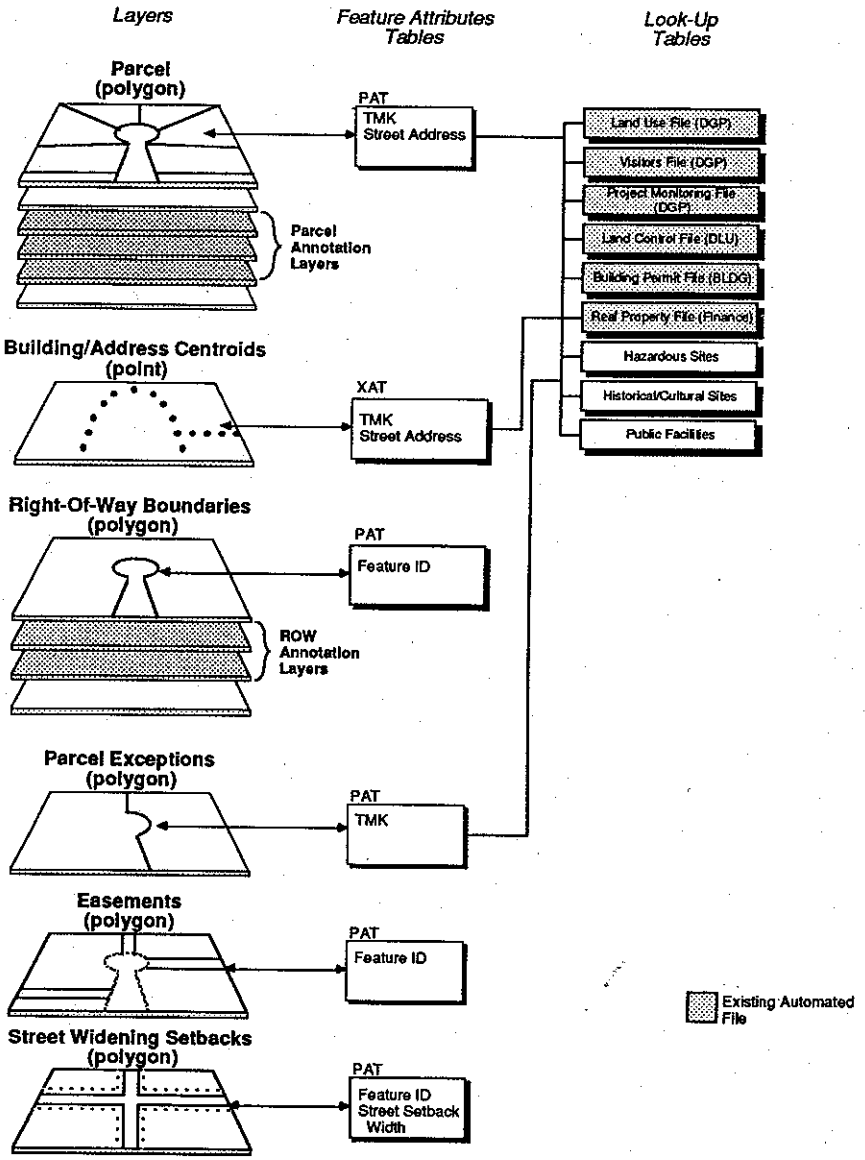
Storm Drain System Layers



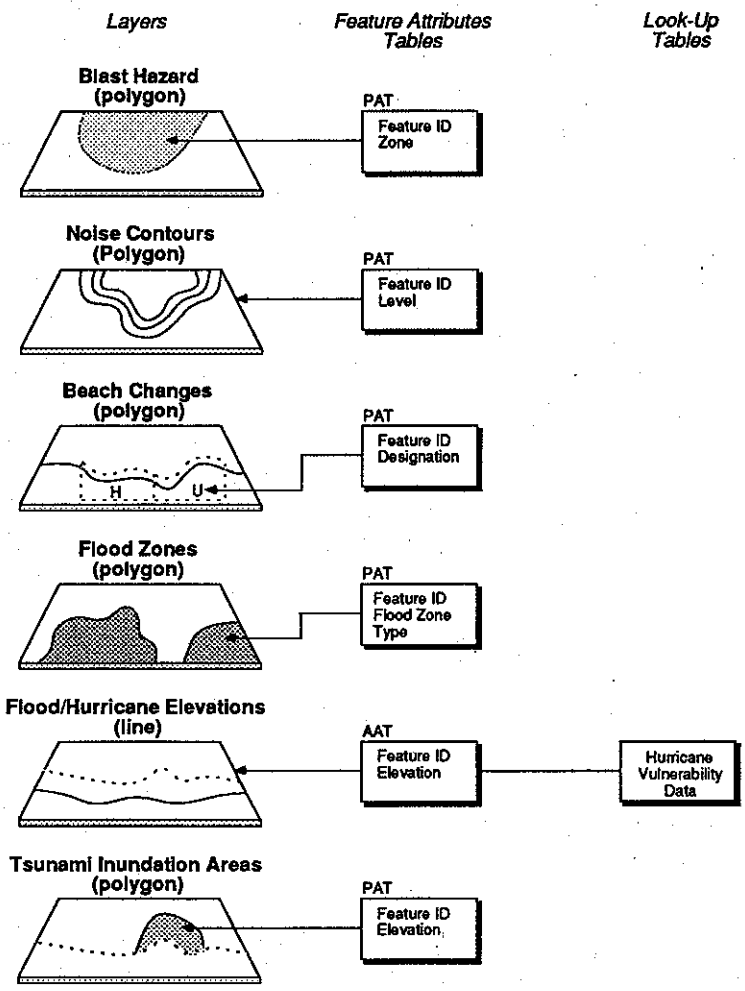
Basemap Data Layers



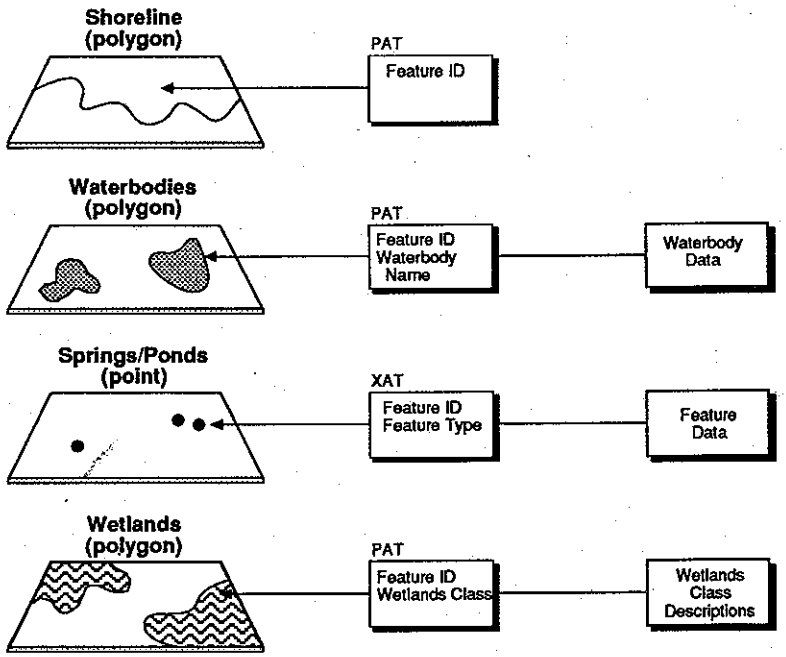
Land Records Layers



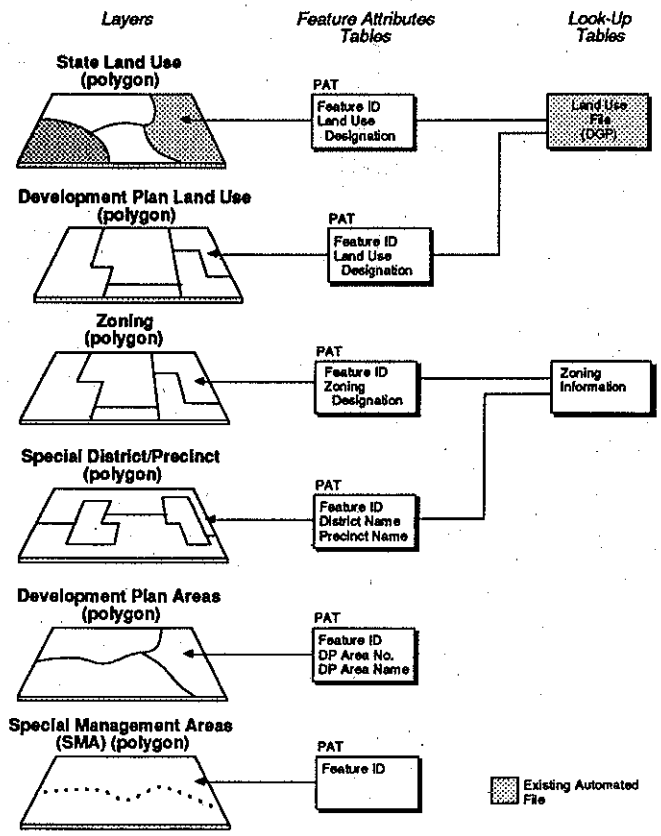
Environmental Constraints Layers



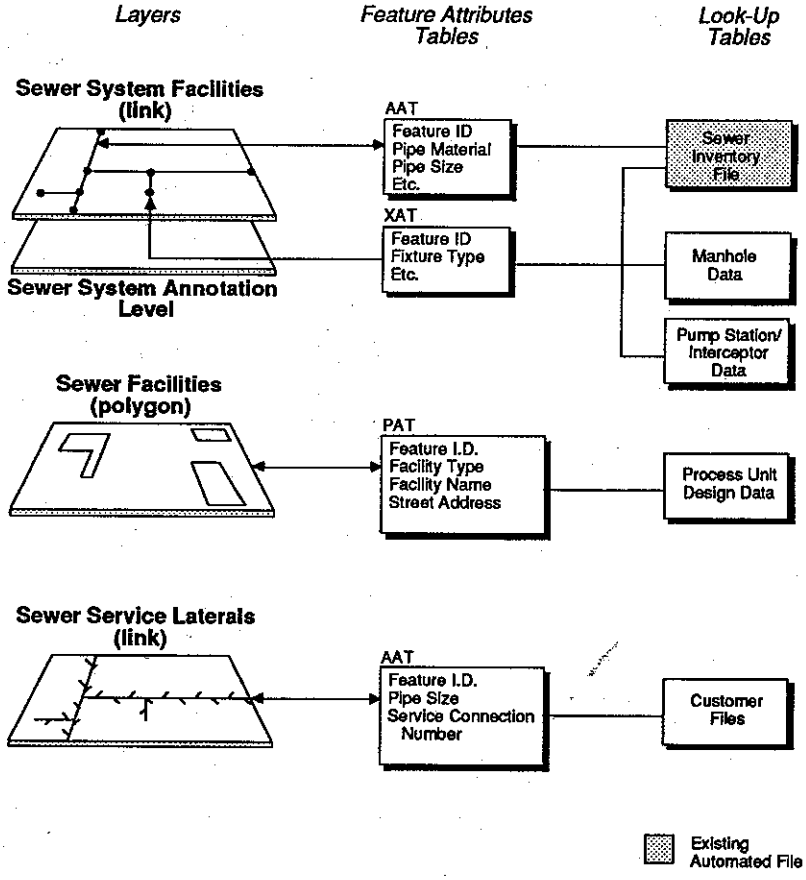
Hydrology Layers



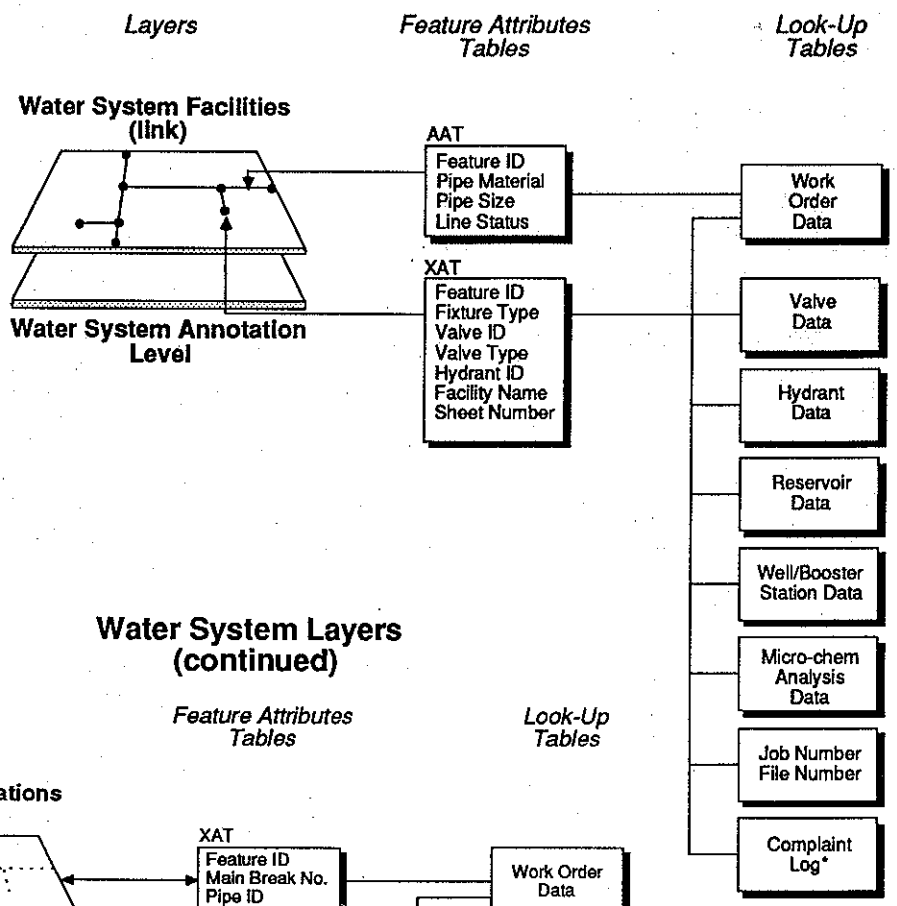
Planning Related Layers



Sewer System Layers



Water System Layers



Water System Layers (continued)