DIGITAL CARTOGRAPHIC DATA BASE
PRELIMINARY DESCRIPTION

Dean Edson
U.S. Geological Survey

CARTOGRAPHIC DATA BASE DESCRIPTION

The expanded scope of the National Mapping Program includes the establishment of Cartographic Data Bases which reside in the public domain and are available for retrieval and reproduction on demand. These data bases fall into two general media categories: graphic and digital. The types of data involved are those generally included in the U.S. Geological Survey topographic maps and are referred to in the National Mapping Program as base category data.

The National Topographic Program will be modified, extended and renamed to serve better the basic cartographic data needs of the country. The new National Mapping Program (NMP) will continue to provide a family of general purpose maps of greater scope than heretofore. In addition it will provide basic map data in a variety of forms useful to the division, other agencies, and the public both for preparation of other maps at various scales and for numerical and statistical analysis of map data. These data forms will include, but not be limited to, color separates, feature separates, or digital data (based on geometric distribution of digitized information).

Base map data categories are:

1. Reference systems: geographic and other coordinate systems except the public land survey network.
2. Hypsography: contours, slopes, and elevations.
3. Hydrography: streams and rivers, lakes and ponds, wetlands, reservoirs, and shorelines.
4. Surface cover: woodland, orchards, vineyards, etc. (general categories only).
5. Non-vegetative features: lava rock, playas, sand dunes, slide rock, barren waste areas.
6. Boundaries: portrayal of political jurisdictions, national parks and forests, military reservations, etc. This category does not fully set forth land ownership or land use.
7. Transportation systems: roads, railroads, trails, canals, pipelines, transmission lines, bridges, tunnels, etc.
8. Other significant manmade structures such as buildings, airports, and dams.

9. Identification and portrayal of geodetic control, survey monuments, other survey markers, and landmark structures and objects.

10. Geographic names.

11. Orthophotographic imagery.

It may be noted that this list does not include photography other than that implied in producing the data categories. However, aerial photographic coverage, without being converted to orthophotos or other base map categories, will be a significant component of the NMP.

The Digital Cartographic Data Base (DCDB) to be developed and maintained by the U.S. Geological Survey (USGS) will be a standardized source for base categories of digital cartographic data principally for the United States. Its implementation will apply the techniques of generalized data base management to produce an integrated approach to the storage, retrieval, and maintenance of digital cartographic data. Data for the DCDB will initially be drawn from existing qualifying sources such as the several series of current USGS topographic maps. When complete, the DCDB, along with the current graphic data base (maps), will be an additional medium for USGS distribution of cartographic data to the community of cartographic data users through the National Cartographic Information Center.

Figure 1 illustrates the DCDB concept, its software interfaces, and its relationships to users of digital cartographic data.

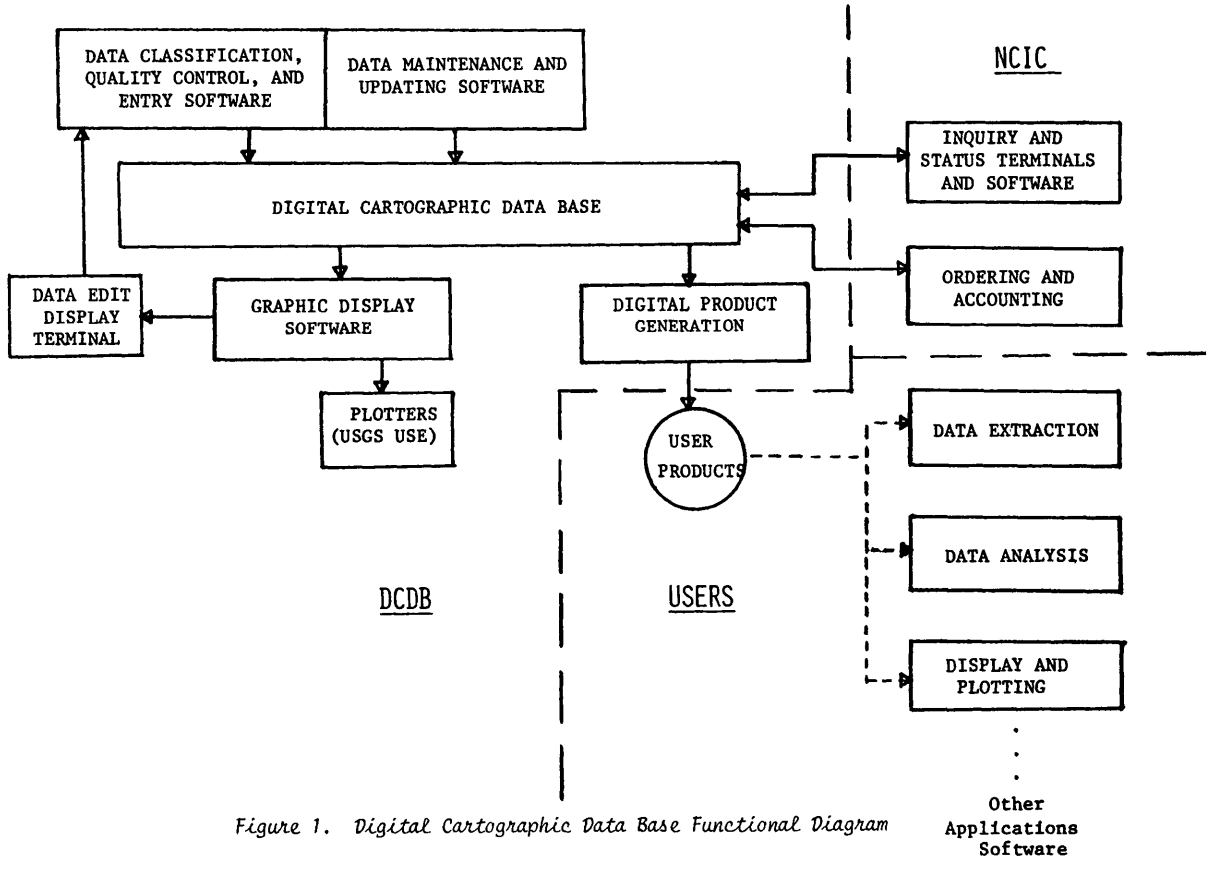


Figure 1. Digital Cartographic Data Base Functional Diagram

Cartographic data for the United States are currently being made available by the USGS in several series of topographic maps at standard scales combining data of many base categories. These data normally comply with National Map Accuracy Standards and are presented in a standard map projection with standardized symbology. Most of these data are also available, on special order, as color-separation film reproductions. Color separates contain the data that would be placed on the map by a single printing plate using a single color, and when available, are in the same scales, projections, and standards as the published maps. These color separates, which constitute the graphic data base, form the major input basis for the definition of data to be included in the DCDB. Five base categories of data have been selected for inclusion in the Digital Cartographic Data Base pilot project: County and State Boundaries, Rectangular Survey System (section corners, etc.), Surface Hydrography, Terrain Surface Elevation, and Transportation. These categories have been selected based on availability of data and initial indications of need expressed by potential users of the DCDB.

DCDB OBJECTIVES

The DCDB will provide selected cartographic data in digital format responsive to current known requirements and structured to expand and evolve as the user community gains experience with digital data and refines its requirements.

By providing management to the DCDB, USGS will establish standard data formats and software interfaces throughout the community of digital cartographic data users, which should help to coordinate orderly development of digital cartographic analysis technology and reduce separate developments of equivalent applications in different formats by different users. USGS will also have the background responsibility to serve as a clearinghouse for applications software developed in forms compatible with the established formats, and for inter-user status communications, with the DCDB as the common reference point, to provide answers to such questions as, "Is what I intend to do with this data already done ...? -- by who, when, etc.?"

Through use of the DCDB and automatic plotting equipment, USGS and DCDB users may acquire the capability to produce certain graphic products not now available. These products would include maps to non-standard scales and projections, combined maps from different base categories, non-standard symbols, colors, etc., and other products that users might identify.

Digital and graphic data taken from the DCDB will be made available to users through the National Cartographic Information Center (NCIC). Although the details of the inquiry and order handling processes have not been resolved, it is known that requests will be handled by a combination of index graphics, which will contain approximate assessments of the data coverage over geographic cells, as well as by detailed computer data base search and retrieval routines. The codes used for the storage and retrieval of cartographic data will be consistent with those used in NCIC's Indexing and Referencing System. In this way DCDB data will be treated as subsets of the total cartographic data set available through NCIC.

In addition to the new products available to users, the DCDB will enable USGS to provide more effective service to users of current standard products through improved operations internal to USGS. These improvements will result from a more rapid update cycle for the materials used to produce topographic maps, improved availability of status information on map revisions and other projects in progress, and possible release of new data in dynamic digital files in less time than required to produce finished maps.

SCOPE OF PROJECT

The development now being planned will be a pilot project to define, design, implement, demonstrate, and evaluate the DCDB. The procurement will include definition and implementation of a data base structure, installation on a USGS-designated computer, and demonstration of data entry, data access, retrieval, and graphic display, data maintenance, and status summary software. Data standards will be developed along with the ability to grade data according to accuracy, reliability, and standard classifications within base categories. Table 1 describes a tentative two-digit spatial accuracy code for the DCDB. A primary objective of the DCDB will be standardization of feature categories (such as kind of highway, etc.) and data reliability codes.

A set of sample data complying with the specified standards will be included in the pilot DCDB for demonstration and evaluation. The pilot project will demonstrate and evaluate all primary data base program modules.

BASE CATEGORY I - COUNTY AND STATE BOUNDARIES

This data category will be designed to include the boundary lines of the 50 States and the county boundaries within each State. State and counties will be described as areas enclosed by boundary lines and the boundary lines will be defined as to type (State, county) and feature identification where the line corresponds to a physical feature such as a river or highway. The exact format of the data storage will be determined as part of the DCDB pilot project definition and methods such as points connected by line segments, polynomial fits to points line segments defined by end points, etc., will be evaluated for storage efficiency and ease of entry and retrieval. The areas will be described by FIPS Codes and access to the data will be through the FIPS Codes or through geographic coordinates. A typical access request by geographic coordinates would specify all boundaries of a given type enclosed within a polygon defined by corner coordinates. The data would be returned in spherical coordinates -- geocentric latitude and longitude -- in radians.

Potential growth to Base Category I Data might include civil townships, towns, cities, State and national recreation and preserve lands. The common attributes of these data are that they are manmade political data, and are most frequently not apparent from physical features or aerial photography.

BASE CATEGORY II - RECTANGULAR SURVEY SYSTEM

This data category is somewhat similar to Base Category I in that it contains the areas and boundary lines defined by the rectangular survey system administered by the U.S. Bureau of Land Management. It therefore includes corner and closing point data and monument data in addition to boundary lines and enclosed areas. Certain land grant areas -- French in Louisiana and Spanish in California -- will also be included where required.

Figure 2

Proposed Accuracy Codes

Positioning Accuracy Code	Scale Equivalent	Positioning Method	Expected 90% Tolerance	Decimal Storage	D M S Output
9	1:2,400 and larger	Precise geodetic position	NTE 1 metres	+31.7051875 +111.6070167	31° 42' 18.675" 111° 36' 25.260"
8	1:2,400 to 1:6,000	3rd-4th Order Geodetic	NTE 5 metres	+31.705188 +111.607017	31° 42' 18.68" 111° 36' 25.26"
7	1:6,000 to 1:14,000				
6	1:14,000 to 1:20,000				
5	1:20,000 to 1:35,000	24,000 map scaling	NTE 13 metres	+31.70519 +111.60702	31° 42' 18.7" 111° 36' 25.3"
4	1:35,000 to 1:70,000				
3	1:70,000 to 1:130,000	Intermediate-map scaling and protracted plate	NTE 30 metres	+31.7052 +111.6070	31° 42' 19" 111° 36' 25"
2	1:130,000 to 1:300,000				
1	1:300,000 to 1:1,000,000				
0	1:1,000,000 and smaller				

The suggested identifier codes for rectangular surveys are given below.

Marker Code		Marker Type Code
Marker Code (Map)	Type of Mark Code	Symbolization (Plotter Output) Code
9	Brass cap monument	00
8	Other found corner	01
7	Accepted	02
6	Plotted	03
5	Witness corner	04
4	Reference monument	05
3	Location monument	06
2		07
1		08
0		09
		10
		11
		12

The spatial accuracy and data origin codes of Table 1 will also apply to this data category and access eill be through the same parameters as Base Category I.

This category will have the potential to expand to include new survey data.

BASE CATEGORY III - SURFACE HYDROGRAPHY

Surface hydrography data will include all perennial drains, intermittent drains greater in length than 600 metres, and perennial open water where the smallest dimension is 15 metres. Drains will be defined as line segments and nodes, described by type, implied flow direction by ascending node numbers, and text. Open water will be defined as areas described by boundary lines consisting of the land/water interface and text. Both types of hydrography features will contain horizontal data only and be accessible through the text names or geographic area methods.

The Water Resources Division of the U.S. Geological Survey maintains an extensive data base of surface hydrography data. A logical extension to the DCDB would be the cross-referencing to these data. In addition, the DCDB has the potential to include hydrology features such as marshes, flats, dry lakes, intermittent open water -- controlled or uncontrolled, rapids, falls, aqueducts, wells, springs, glaciers, etc., in addition to the base data described above. Selected expansion to include these data types will be determined by the needs of DCDB users.

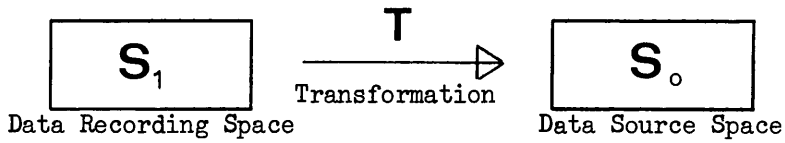
BASE CATEGORY IV - TERRAIN SURFACE ELEVATION

This category will contain all basic elevation data for the DCDB. These data represent the third dimension of the real world and are graphically presented as contour intervals and spot elevations on topographic maps.

In the digital domain, three-dimensional data are as easy to represent as two-dimensional data, and ideally it would be desirable to implement an elevation function above a horizontal grid. This function could be described as points defining segments of planes, fitted analytical coefficients, or other forms, and would directly provide the data most commonly required for digital applications. In practice, an enormous volume of data has been collected; these data are represented as contour lines on a flat surface, and this form of elevation data presentation is widely used and will continue to be demanded by users. General purpose transformation software between contour intervals and three-dimensional grid formats for terrain elevation data will, therefore, be appropriate as part of the DCDB applications software if grid format data are widely utilized. However, the costs of converting existing data would be high, and no clear mandate currently exists which makes conversion necessary. It is, therefore, proposed that the DCDB provide for entry, storage, maintenance, display and retrieval of terrain elevation data in three forms, and for data to be entered in whichever form they are available for given areas.

The three different ways for documenting hypsography of an area are: elevations at regular intervals along terrain profiles, planimetric traces of contour lines, and elevations at planimetric points of critical elevation or slope change. In every case, a terrain point is spatially defined by its coordinate components with respect to some three-dimensional space S. It is essential that the space S be itself a Euclidean space or can be transformed to one, in order for the terrain data to be susceptible to geometrical manipulation. The space S may or may not be Earth related. However, the usefulness of terrain surface data in a data base setting will almost always require a space S which is Earth related.

For any set of terrain surface data, two coordinate spaces are assumed to exist: Data Source Space S_0, and Data Recording Space S_1. The Data Source Space S_0 is the coordinate space in which source material for terrain surface data are documented. It is usually a non-Euclidean space with planimetry represented on some map projection and elevations reckoned from some reference surface. The space S_0 could be also a three-dimensional Euclidean space; such as geocentric, local secant, etc. In every case, spatial coordinates of points in space S_0 are usually transformed during the digitization process to some arbitrary coordinate space which we will refer to as the Data Recording Space S_1. Efficiency of storage, digitizing equipment, and terrain data source material are factors which affect the choice of this transformation T. It is essential that the terrain surface data file contain a definition of the transformation T and its parameters. The relationship between the coordinate space S_0, space S_1, and transformation T are schematically represented by the following figure:



The logical organization of terrain surface data files will be presented in terms of two data records: header record and terrain surface record. A typical file will contain one header record and many terrain surface records. In terms of storage space, requirements for terrain surface records are by far the most imposing, and optimization efforts in their structuring are most rewarding.

This category of data will be limited to terrain elevation data.

BASE CATEGORY V - TRANSPORTATION

Base Category V will contain primary elements of the U.S. transportation network including roads, railroads, and power and pipelines. Roads will be categorized by type such as limited access, heavy duty, medium duty, light duty and unimproved. Several classifications exist for roads and it will be a goal of the DCDB to identify, standardize, and integrate these systems. The pilot project DCDB will include single and multiple track standard gage railroads and selected (trunk) pipelines and powerlines. Transportation features will be defined as lines with nodes, and the lines will be associated with feature classification. The format of their description will be determined in the definition of the DCDB and result from trade-offs considering accuracy, storage required, and ease of entry, maintenance, display and retrieval. The data will be horizontal only, with elevation available from Base Category IV. Data origin, accuracy, and reliability codes will be described by codes similar to those defined in Table 1.

Roads under construction and proposed roads constitute important data to planners and are a potential area of Base Category V expansion. In addition, other transportation features such as non-standard gage railroads and carlines, canals, inland shipping routes, interstate route overhead clearances, and communications links could be later included as need indicates.

DATA STRUCTURE REQUIREMENTS

The potential benefits of an integrated Digital Cartographic Data Base are only realizable if the data are structured (organized) in a way that meets the system objectives. A considerable technology has evolved concerning the management of large computerized data bases, and this technology focuses on the data themselves as the end product to be created, maintained and distributed to users. This development is in contrast to earlier views of digital data processing which focused attention on the processing programs and data were simply "input" and "output." The emerging Data Base Management (DBM) technology has resulted in several DBM languages of which one system (System 2000 developed by MRI Systems Corporation) has been procured by the USGS for application in our Water Resources Division. Computer main-frame manufacturers are now generally offering their own DBM systems also. These systems are in general targeted at commercial applications such as personnel systems, sales, and inventory control, etc., and may be applicable to the DCDB. Special emphasis will be directed at evaluation of these available systems in the DCDB pilot project definition to determine their application potential to the DCDB. Good data base design requires trade-off of conflicting requirements to achieve optimum response to the

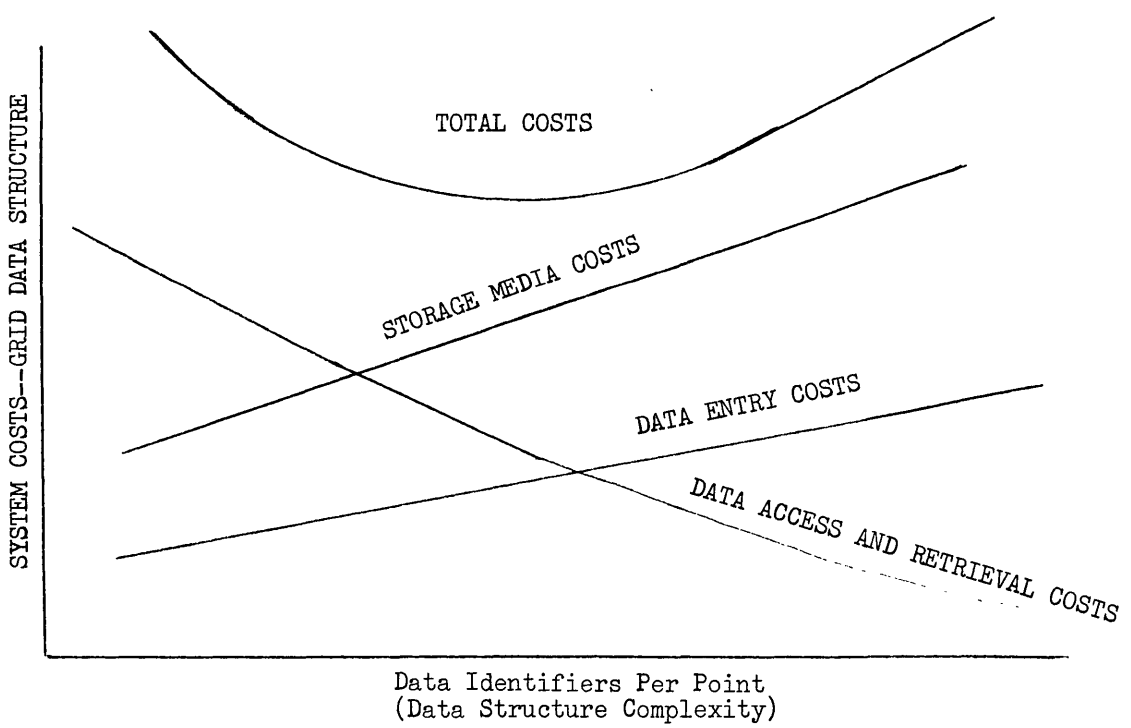
objectives of a specific data base. Requirements definition and trade-offs will be a significant task in the DCDB development.

Figure 3 illustrates a typical trade-off the DCDB optimization must consider. This figure shows that, for a typical Base Category of data and a typical data structure, how minimizing storage media cost conflicts with minimizing data entry, access, and retrieval costs. Optimum resolution of this conflict requires evaluation of the combined effect on total system cost. Figure 3 illustrates summary data. Data entry costs, for example, would be supported by frequency of data entry and unit entry processing cost, data entry software development cost, etc. Although the numerical values attached to the curves shown in Figure 3 are not often well known, the trends are usually well understood and good system design requires early evaluation of these factors for a particular system.

Since the DCDB will be an active, working data base, it is necessary to organize the data in such a way as to minimize massive searches for data access, updating, and retrieval.

Data entry, updating, access, and retrieval software will itself be a significant development item if an existing DBM language is not selected, as will maintenance of this software. The DCDB structure should minimize the complexity required of this software. In addition, users of digital DCDB projects will be acquiring and developing applications software to accomplish their analyses of the data. It is necessary to avoid placing a requirement for a high degree of software sophistication on the data users because of the direct relation to their costs for applications software development.

Figure 3. Illustrative CDB Trade-Off



As one views the spectrum of user requirements for cartographic data in digital form, it becomes immediately apparent that the initial DCDB will not satisfy everyone. However, there are some concepts related to structuring data which will enhance the usefulness and flexibility of the data base at a reasonable dollar investment in terms of structure design, file building and storage. The most fundamental of these concepts deals with the total relationship of all features identified in the DCDB. This relatedness concept is best described as the topology of a region.

In considering a topological structuring approach, we find that the relatedness of features can be expressed in two-dimensional space by the intersections or junctions of like or unlike features. In the context of Base Category features, if each of these intersections or junctions, which are referred to as nodel points or nodes, is assigned some sort of identification and has a spatial reference such as a latitude or longitude or map projection coordinates, this group of points represent a topological framework to which other data can be related. When the topological framework is filled in with Base Category feature data and plotted, a base map is formed.

Some specific examples of nodes are given in Figure 5. Within this illustration we find the following:

Road intersections	nodes A, C, E
Road ends (terminations)	nodes G,
Road and drain intersection	node B
Section line and drain intersection	node K
County boundary and road intersection	node D
Drain and open-water intersection	nodes H, J

The formation of the features which connect the nodes is of three types:

1. Straight lines which need only end points to be defined;
2. Simple curved lines which require either end points and a radius or three points; and
3. Random-shaped lines which require a string of closely spaced points to approximate the line location. Such lines may represent centerline location, boundary location or physical interface such as shorelines.

In any case, the actual location of features is spatially definable as coordinate points. Figure 4 illustrates the use of coordinate points to define some typical features. Note the numbering is random.

These coordinate points constitute the root level of data, so this is where the data base structure begins. In order to store and retrieve the coordinate (root) level of data, a point directory is assembled. As illustrated in Figure 4, each set of point coordinates is stored with a corresponding point number.

The second data structure level to be established is accomplished by setting up a directory of nodes based on corresponding points. Since we store point coordinates only once and at the root level, a node framework could be plotted by noting point numbers in the node directory, retrieving the corresponding points and plotting each as a separate coordinate point. Figure 5 indicates this relationship and notes the start of the topological data structure levels.

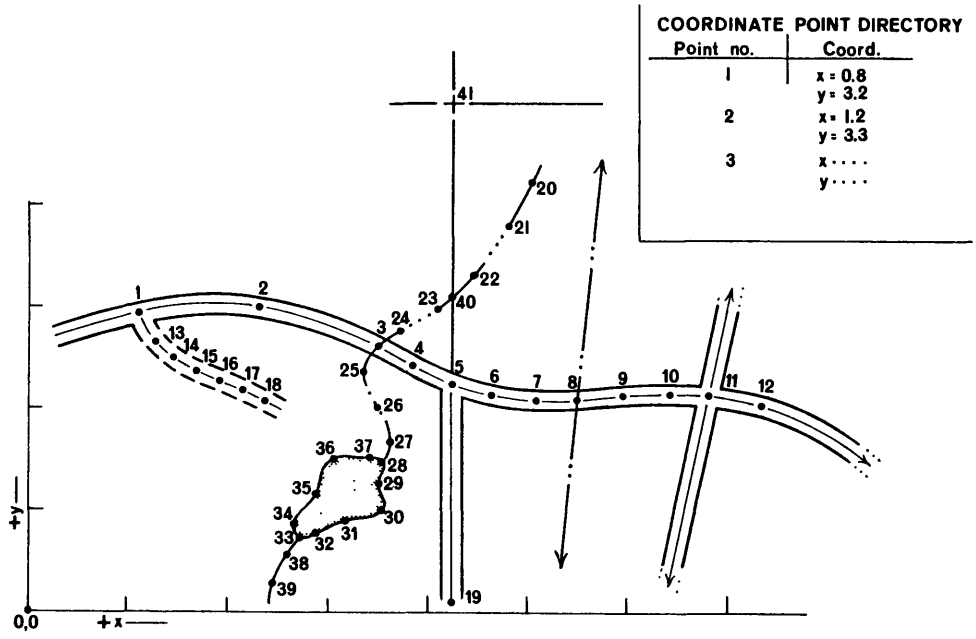


Figure 4. Points - The minimum number of spatial coordinates needed to define the centerline of linear features, the center of point features, and the edge of area features.

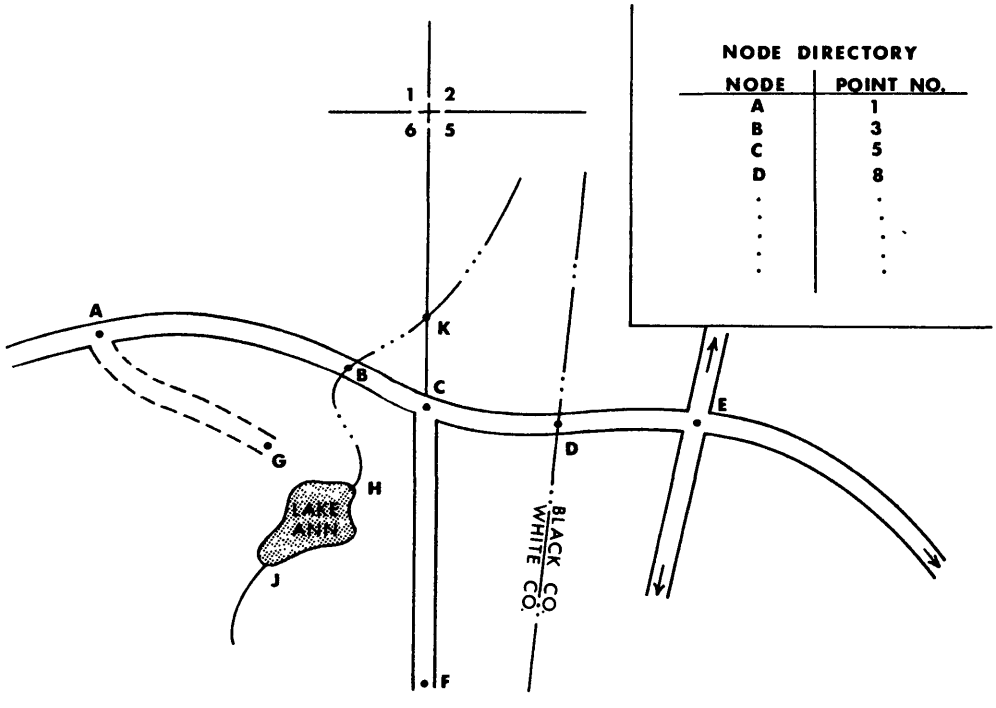


Figure 5. Nodes - Those points which represent the intersection of two or more like or unlike features or the point which represents the end of a linear feature.

The next data structure level involves identifying the string of related points which define a feature from node to node. These are called chains. Figure 6 illustrates the chain identifier for strings of points between nodes. A chain therefore represents a line segment between two nodes.

A feature such as a road, a drain, a lake, or a boundary is usually made up of several line segments or chains, so it is important to establish the third data structure level called chain groups. As an example of chain groups, let's consider a road several miles long. This road would be identified as a series of chains and would be assigned a single chain group number. In order to plot the road, the chain group directory is consulted to find the proper chains, then the node directory is consulted to find the proper coordinate points from which the road is plotted.

Networks of roads and drains, which can be identified as networks, will appear in a network directory under two specific types -- branch network for hydrology and block networks for transportation. This directory permits important name identifiers such as drainage basins and major river networks to become data retrieval descriptors. The network directories constitute the fourth level of data structure. These relationships are illustrated in Figure 7.

The user query level utilizes geographic names and feature codes for basic data retrieval. Data generally will be organized into logical area modules such as 1:100 000-scale quadrangle which represents 30 minutes of latitude by 60 minutes of longitude. Each module would be referenced by 32 sub-modules which would be called a data reference page. These 32 pages correspond to the $7\frac{1}{2}$-minute quads in each module.

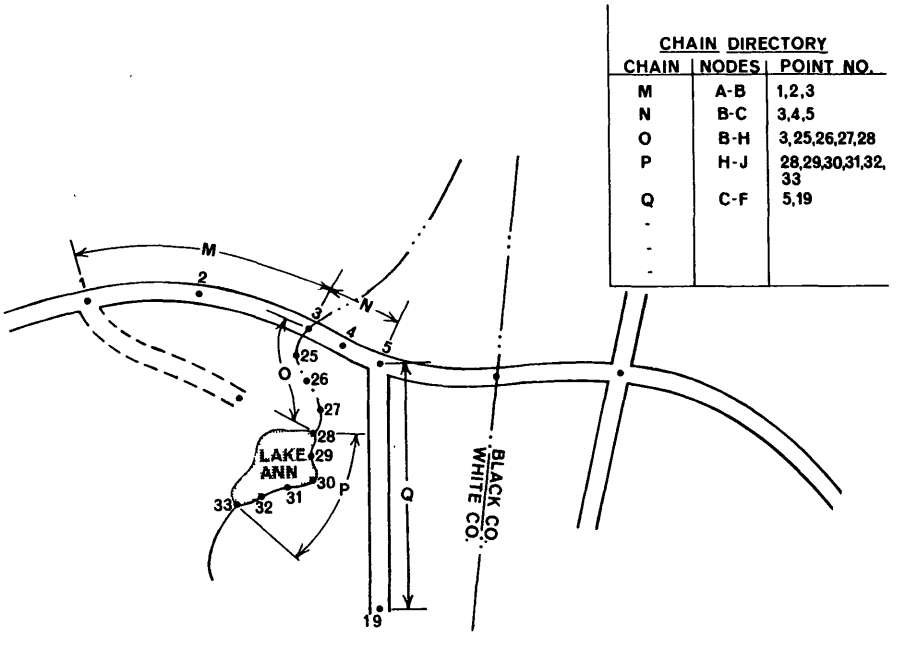
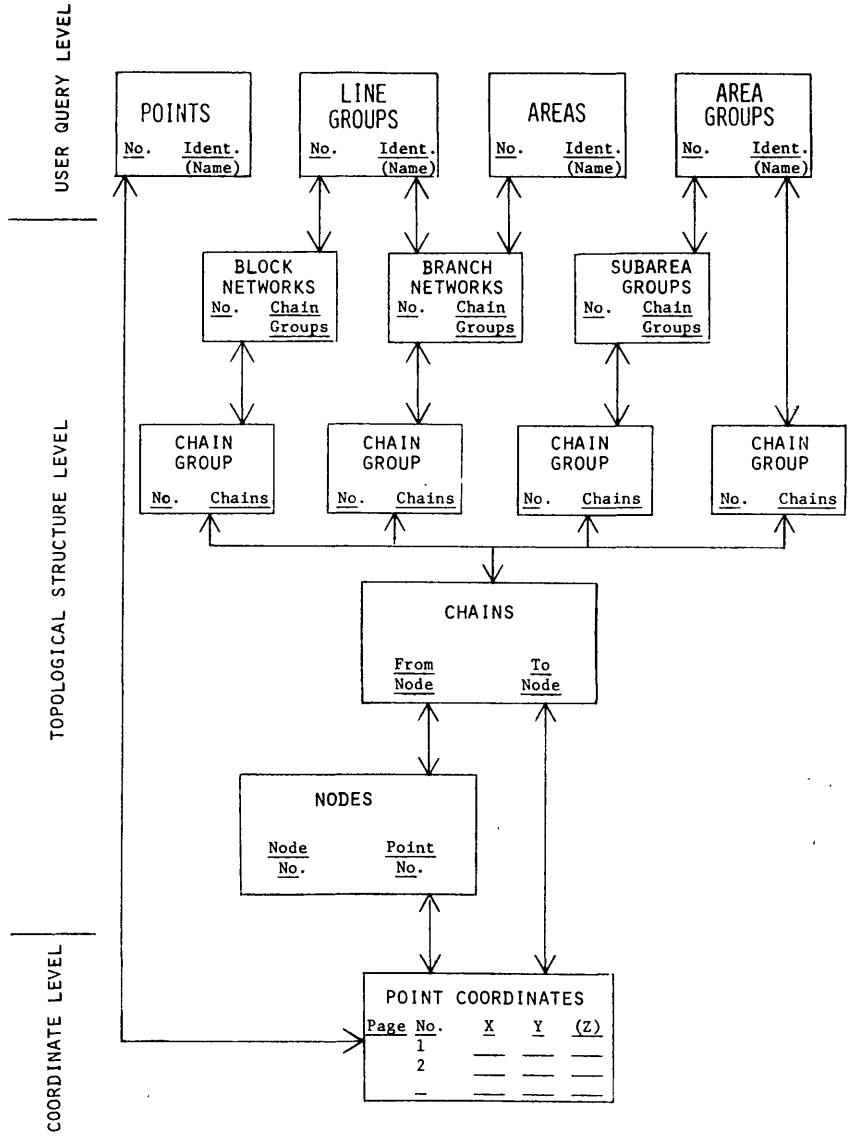


Figure 6. Chains - The series of coordinate points needed to define a linear feature between two nodes. When the linear feature is straight, only the end points (nodes) will appear in the Directory (see example Q).

Figure 7. Data Base Structure Concept

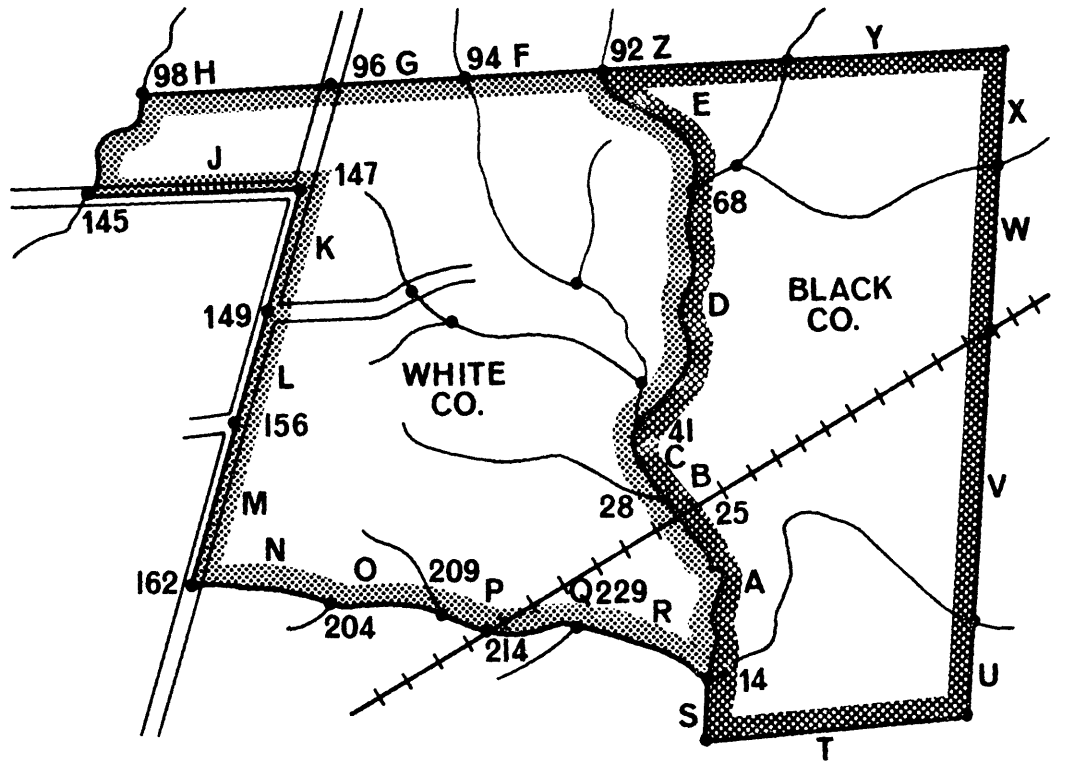


It is important to recognize that points, chains, and chain groups can be shared by almost any number of features. This structuring technique provides a basis for potential users to obtain copies of the point, chain and node data and establish their own higher order directories for specialized use without having to redigitize the root data.

The structuring of area groups such as counties, States, sections, etc., is established in a manner similar to networks. Using this approach, a complete closed area is identified at the chain group level such as sections, counties, land grants. These small area units are aggregated upward to become regions, States, national park management areas, etc.

One of the important aspects of this structure concept is the use of geographic names at both the highest topological structure level and the user level. Figure 8 illustrates how a name is used as a retrieval descriptor for an area such as a county.

Figure 8. Chain Group for White County



CHAIN GROUP NO.	CHAINS
1	A through R
2	A through E S through Z

AREA GROUPS CHAIN GROUP	NAME
1	= WHITE CO.
2	= BLACK CO.
1+2	= WET BOTTOM WATER DIST.

Some of the attributes of a topological data base structure can be identified at the outset as follows:

1. We estimate that the amount of data residing in the DCDB will reach or exceed 2×10^{10} points. This is based on 50,000, 7.5-minute quads as the primary data source, each containing an estimated 400,000 points. In terms of reels of magnetic tape, if standard 2,400-foot reels are used as a recording medium and the recording density is 1,600 bits per inch and it takes an average of 20 digits to define a point, then it would require 11,000 reels of tape to record just the point coordinate data. Obviously searching a file this size is not a minor consideration and with the proposed data structure which includes appropriate pointers and identifiers retrieval time can be minimized.

2. The topological structure provides a point framework upon which the data file can be easily updated just as junction points on a map are now used as a basis for revision. This is probably the most significant attribute from a data management standpoint.

3. Retrieval of specific data for a given area will be a high demand retrieval mode. This structure permits logical and direct retrieval as opposed to a spiraling type search.

We envision a system of software modules organized as follows:

Creation Modules: This set of modules is responsible for capturing raw data and extracting from it the information required by the data structure to process the data to meet users' requirements.

- Data Capture: Converting data from graphic (e.g., mylar separations) form to digital form.

- Data Formatting: Converting the digital data created in the previous step to a form which includes all necessary identifiers and topological relations implicit in each data set.

- Data Structuring: Entry of data into the archive, using the internal data structure that will be used for all subsequent references to the data.

Maintenance Modules: This set of modules will permit verification by staff members of the correctness of items in the DCDB, and editing of them where necessary. The data structure can be expanded through these modules to enable possible new categories of retrieval, and tests performed to insure that performance standards continue to be met.

- Updating: File maintenance, primarily by interactive graphic inspection, to allow manual insertion, replacement and deletion of data items.

- Data Structure Enhancement: Defining new entities within the data structure (based on information already in the system), thus expanding the categories of data available for retrieval. This does not mean altering the data structure, necessarily; rather, the data structure should be inherently capable of extending itself.

- Testing: Routines will be involved, routinely and as required, to examine the performance of the data base in crucial tests.

Retrieval Modules: This is the user interface and includes preprocessing of the DCDB to extract user-requested information, followed by output of data from the system in proper format.

- Preprocessing: A user may request intricate subsets of information from the DCDB. This may involve, for example, assembling a number of quad sheets together, changing scale and/or projection, requesting certain features that are spatially coincident or arbitrarily close together, or multiple overlays with certain features selected out. Such tasks are inherently within the power of the data structure, but would not normally take place in the absence of a user request.
- Graphic Output: Interactive display and plotted output must be available, the sophistication of which is essentially limited by the graphic output hardware.
- Digital Output: Machine-readable files can also be output, mainly on magnetic tape. A variety of formats should be possible without special intervention. DIME files or World Data Bank format files, for instance, ought to be among the standard output options, as well as matrix conversion for certain types of data (especially terrain).
- Software Dissemination: Users should be able to acquire software for using data from the DCDB.

The DCDB will include status and summary record data such as entry/update date, source, accuracy and reliability, and data access history with applications data. This will enable users to determine if the analysis they intend to perform on the data has already been accomplished or if applications software is already developed.

The DCDB must ultimately include interactive graphics for data editing, file building, update and retrieval. The requirements to provide for this capability will be evaluated during the design phase of DCDB development.

REFERENCES: RECENT ARTICLES AND REPORTS ON DATA BASE SYSTEMS

1. Musgrave, B., "The ABCDs of DBMS," Computer Decisions, January 1975.
2. Hunter, J.J., "Pointers in Data Base Management," Computer Decisions, January 1975.
3. Cuozzo, D.E., and J.F. Kurtz, "Building a Base for Data Base: A Management Perspective," Datamation, October 9173.
4. Eriksen, S., "The Data Base Concept," Honeywell Computer Journal, volume 8, number 1, 1974.
5. System 2000 General Information Manual, MRI Systems Corporation, 1972.
6. Feature Analysis of Generalized Data Base Management Systems, CODASYL Systems Committee Technical Report, May 1971.
7. Good, E.F. and A.L. Dean, 1971 ACM SIGFIDET Workshop -- Data Description, Access, and Control, Association for Computing Machinery, November 1971.