DATA BASE MANAGEMENT SYSTEM APPROACHES TO HANDLING CARTOGRAPHIC DATA

DR. MARBLE: Our third session today is organized around the problem involved in storage and management of spatial data. Many cartographic applications have utilized moderately sized to small data files. Very few of them, until relatively recently, have had an opportunity to use large volumes of spatial data. One of the truisms in computer processing of any type of data is that if you have only a small amount of data on hand, it is difficult to produce large increases in efficiency and savings, simply because you are not carrying out very many operations. However, as data volumes increase we find ourselves rapidly confronted with the problem of not being able to afford access to the data. As we look at the development of potentially large digital data bases, we must face the fact that we are going to have to worry and worry hard about managing these in an effective fashion in order to attain economic viability in their use.

One of the techniques that has been adapted for non-spatial data is the notion of the data base management system. This is a very complex software system which essentially stands between the applications user and the physical data itself. It permits the applications programmer to maintain a logical view of the data which may be, and quite often is, greatly different from the actual physical organization of the data in the computer. The important notions here are those of logical and physical independence of the data. However, when dealing with spatial data, particularly that containg large volumes of coordinate information, we have found that we run into some peculiar problems. The speakers today are going to address, first, the conceptual problems involved in handling spatial data, and then discuss a specific example which tries to utilize a data base management system approach to the manipulation of cartographic data.

The first speaker is Dr. Roger Tomlinson. Roger is chairman of the IGU Commission on Geographical Data Sensing and Processing. Roger?

DIFFICULTIES INHERENT IN ORGANIZING EARTH DATA IN A STORAGE FORM SUITABLE FOR QUERY*

DR. R. F. TOMLINSON: The purpose of this paper is to identify some of the methodological problems inherent in organizing a store of earth data in a form suitable for query. Earth data are here defined as those that describe the earth's shell, ocean and atmosphere and they are attached to a specific location; they are usually stored and displayed in the form of maps. Topographic maps, land use maps, soil maps, geological maps, vegetation maps, weather maps, oceanographic charts, population maps, and geophysical maps are well-known examples of stores of such data. To take advantage of the calculative capacity of existing computers to analyze these data, increasing amounts of them are being converted to digital form. Furthermore, instruments that gather earth data, such as sensors mounted on satellites, automatic gauges in streams, sounding devices on ships, and ground topographic surveying instruments, are now providing their data directly in digital form. The volume of earth data in digital form is thus growing rapidly. However, because of the discipline imposed by use of current computers, many of the relationships between data elements hitherto visually derived from maps must be more explicitly specified if the digital data are to be organized effectively for query. This raises some questions about the nature of such spatial data and spatial relationships that have not been widely discussed or resolved within the discipline of geography, or in other disciplines. These questions are outlined below, and some initial steps to resolve the problems are proposed.

A map can be thought of as a structured file in which entities, conditions and events in space are recorded. For maps of earth data the structure presupposes some conceptual model of the space occupied by the globe, suitable units for its measurement, an adequate transformation from the curved surface of the earth to

* This paper is the outcome of discussions in the past year between Stephen Gale, Michael Goodchild, Ken Hare, David Hays, Fred Lochovsky, Duane Marble, Dick Phillips, Azriel Rosenfeld, Mike Shamos, Dennis Tsichritzis and Roger Tomlinson, held under the auspices of the International Geographical Union Commission on Geographical Data Sensing and Processing.

the plane surface of the map, and the use of graphic conventions for representation of real world entities. The appropriateness and validity of current cartographic practice are not called into question in this discussion. Extremely large volumes of useful information can be, and are, stored on conventional maps. In general, the locational values of the contents of the map rely on establishing a series of identifiable points on the ground by multiple measurements between them, and between the points and extraterrestial bodies. The established points are "filed" on the maps according to their measured relative positions. Elements of thematic data are then located by observing or measuring their relationships with easily identifiable features already stored in the file, and they are recorded by inserting them in the appropriate place in the file, that is, by plotting the observations on a map.

The spatial relationships between entities,[1] relationships such as contiguity, adjacency, nearness, connectivity, above, below, between, and inside are occasionally explicitly defined by conjoint symbols (villages connected by roads, stations marked on railway lines) or by written values (distances between points). However, more frequently they are implicit in the file structure and must be determined by visual estimation or measurement and calculation.

When information is extracted from a map, it typically includes a mixture of the values of entities and the relationships between entities, appropriate to the question being asked. The utility of a map as a source of information is good at first consideration, in that the storage medium is also the display medium. When the data of concern are explicitly recorded on the map, retrieval is swift. Similarly, brief and simple estimations or a few straightforward measurements seem to yield a reasonable return for effort. However, when information extraction requires many measurements or calculations to determine relationships implicit in the file structure of a map, the task rapidly becomes tedious and error-prone. Although map sheets can be a compact and symbolic form of data storage, the number of sheets required to contain even a small amount of the spatial data already gathered from a particular area of the earth may be large. The time and effort of information retrieval increase in proportion to the volume of graphic data to be handled. In fact, increasing the data

1. Conditions and events are subsumed.

volume rapidly limits the type of retrieval operations that are economical, to the point where it can be extremely time-consuming, laborious, and costly to extract even the data actually written on one or more maps. In short, the limitations of human retrieval capabilities place a severe constraint on the utility[1] of maps as sources of large volumes of spatial data.

The continually improving storage and calculative capabilities of computers have been seen as a way to overcome the limitations of human efforts in retrieving and handling mapped data. Numerous systems for the storage and handling of map data in digital form have, in fact, been developed since the early 1960s. The use of computers requires that the data be in machine-readable form. At present, in 1978, the process of conversion from map (graphic) form to digital form, usually referred to as "digitizing," is still technically cumbersome and demands effort and expense. More significantly, the volume of digital data required to reproduce adequately the information content of a conventional map is substantial, perhaps rather more so than anyone had realized. One example can be drawn from the Land Use Mapping and Data Project in the U.S.A. The entire project involved 359 map sheets at a scale of 1:250,000. This relatively small number of maps is estimated to have more than 1.5 million inches of line data which will be digitally described by approximately 68 million x,y coordinate pairs. Topographic maps, as a category, appear to have somewhat larger amounts of information per square inch. A preliminary estimate of 235 million line inches of contour data alone has been made for the sheets available in the 7.5-minute, 1:24,000 U.S. Topographic Series. At a resolution of 12 points per inch, the contour data would require a digital record of 2.8×10^9 x,y coordinate pairs. At 175 points per inch, they would require 4.1×10^{10} coordinate pairs. Decisions have not been made on whether to digitize contours as lines, and on the resolution with which such lines should be recorded, but these figures clearly indicate that data volumes are so large that they must have a significant impact on design specifications for stores

1. The "utility" of a source of information for decision making purposes is dependent on 1) the relevance of the information to the decision, and 2) the ease with which pertinent information can be sensibly extracted from the store of data. Human retrieval capabilities affect the latter aspect of utility.

of digital spatial data. One logical way of handling the problem of data volume is perhaps to reduce the amount of information demanded for activities that use digital spatial data, to a volume more directly related to their needs. There is, however, a long and valuable tradition of accuracy in cartographic displays that will not be overturned overnight, and present practice is to attempt to portray the information as accurately as possible within the limitations of the instruments used.

As increasing numbers of maps are digitized and as data gathering institutions, particularly those concerned with environmental earth data, develop and implement techniques that generate data directly in digital form, some of their stores of data become very large. The U. S. Geological Survey, for example, has over 50 systems handling a wide variety of earth data in digital form. The aggregate volume of such data already in machine-readable form in 1977 is approximately 500,000 million bits.[1] Conservative estimates indicate that this will grow by more than 250%, to 1.7 million million bits, by 1981. Other institutions have similar objectives and expected growth patterns. It can be assumed that computers will become better and cheaper, and that the developing processes of institutional management will tend to match data production to handling capability, or more particularly to computing capacity. There is, however, cost associated with the use of computers, and the volume of data to be processed has a marked impact on that cost. We are rapidly passing the point where it can be assumed that "the computer will handle it," and we should ask ourselves how large volumes of spatial data can be organized efficiently.

Large volumes of data are not new in the world of computer science. Certainly, on a commercial basis, data base management systems have been developed that permit efficient handling of very large data bases for specific requirements of retrieval and manipulation. The principle of any current data base management system is to organize the data in such a way that paths are established to retrieve the entities required for specific enquiries, and at the same time to specify adequately the relationships between the entities that are pertinent to frequent enquiries.

1. A bit is the smallest unit of normal machine-readable information. One regular reel of magnetic tape can hold up to 300 million bits and one regular disk pack can hold up to 800 million bits.

To achieve this, the entities of concern, the relationships of concern, and the operations to be performed on the data must be unequivocally specified before the required data base can be generated for the data base management system. Defining spatial entities of the kind usually found on maps presents no major methodological problem. An adequate schema based on the representation of entities as either points, lines, or areas (with areas being a peculiar kind of line), can be devised. The entities have a variety of spatial or aspatial attributes attached to them. The spatial attributes, the coordinate or locational information attached to the entity, define the selected information content of the graphic image of the entity and its spatial position. The aspatial data record the desired information content of the value or values of the entity. Point entities are usually adequately spatially defined by a coordinate pair. Line entities, however, are typically characterized by a great deal of locational information. This difference in the volume and nature of spatial identifiers is what has made some types of earth data relatively easy to handle (those adequately represented by points), whereas others impose a substantial burden of data processing.

It is in defining the spatial relationships of concern and the suite of operations to be performed on the data that methodological problems arise. The notion of relationships in two or more dimensions has had some discussion within the field of geography, and in other fields, but we still do not have too clear an idea of what we mean by "relationships between entities." It is not certain at the moment that we could adequately define a comprehensive, internally consistent set of relationships that would allow us to devise a logical storage schema for a general-purpose store of earth data. Nor is it clear how the relationships might best be stated. The relative utility of languages of dimensionality has not been widely examined.

The spatial relationships that need to be defined within a data base, the most suitable language or languages for defining such relationships, and the selection of the suite of operations to be performed on the data must be determined in the context of the purpose of the data base. Perhaps the concept of a general-purpose store of earth data is not a useful (or desirable) objective, and earth data should possibly be assembled in a wide variety of logical schemas, each related to a certain category of questions. The problem still remains of defining the purposes of concern, identifying the

methodology, and establishing the patterns of inquiry associated with each purpose.

Because earth data are usually gathered by some institution and are arranged and stored for the perceived constituency of that institution (or parts of that institution), there is a commonly expressed feeling that pragmatic choices of data organization have already been made in the light of user needs, and that despite a possible risk of institutional bias in the provision of data, no serious problems exist. That view is being called into question, frequently by those most closely involved with digital data handling in the traditional data gathering institutions themselves. There are several reasons for this. Many of the systems for handling spatial data developed to date have fallen into disuse because they served no users adequately or economically, or served only a very limited range of users. Many data sets are multipurpose in nature (topographic maps, for example) and can reasonably be included in many logical schemas for different systems of inquiry. The resolution of complex questions concerning the environment and social interaction with physical resources will require data from various sources to be used in concert and in a way that will allow the relationships between disparate entities to be adequately determined. In fact, little work has been done on the nature of the questions that we ask of spatial data, and we do not adequately understand the relationships between the logical schema of a data set and the types of questions that can be answered from it.

A map produced and used by a human is, in a very real sense, a data base management system. The graphic product represents the organization of the data in a logical schema, and displays the data so that a human can determine the nature of the entities and the relationships of interest. Some of the drawbacks of this process, which were mentioned at the beginning of this paper, seem to be repeated in existing digital data base management systems. The following brief comparison of the human and computer-assisted approaches is instructive, as it focuses attention on the underlying nature of the problems.

1) In both human and computer-assisted approaches, if the required information is to be easily found in the file and extracted from it, the entities and relationships concerned must be explicitly defined (written) in that file. The greater the volume of data to

be handled, the more this holds true, and, by definition, the digital data base management systems are designed to handle large volumes of data.

2) If in either approach the relationships between entities are not explicitly defined but are implicit in the file structure and have to be derived by measurement and calculation, then retrieval of required information is laborious. It can be argued that computers have a vastly greater capacity for explicit measurement and calculation than humans have, and that therefore they are useful for such spatial data handling. However, the fundamental purpose of digital data base management systems and, presumably, the gains in retrieval efficiency inherent in them are based on the premise that they provide paths that allow explicit determination of the required entities and their relationships. It seems to be defeating that purpose (and hence the current utility of data base management systems) to rely heavily on computer capability to calculate relationships within a data base management system. Obviously there must be a trade-off between explicit definition of spatial relationships and the calculative capacities of computers. This trade-off is not fully understood and probably depends on how computers compute as much as how fast they compute. It also depends on the capabilities of a particular data base management system and how frequently a particular relationship is queried. This will be explored further below.

3) Computer data bases as well as maps can be displayed in graphic form. Modern interactive display devices also allow the human to manipulate, to some extent, the contents of computer storage so displayed. It can be argued that this capability makes it unnecessary to define all spatial relationships explicitly in the digital file; when they are needed they can be observed. Undoubtedly a human has an excellent mental facility for pattern matching and pattern recognition, and can use this capability to advantage on a small amount of displayed material, for recognizing both the nature of entities and the spatial relationships between them. Given simple images and straightforward tasks, such as allocating a contained centroid to a polygon, this approach can be very efficient. The weak link in the process is the sensory channel capacity[1] of the human, which limits the volume of graphic

1. Human "channel capacity" is the maximum rate at which bits of information can be transmitted to the brain through all human sensory channels.

information that can be made available to the human mind. This limitation thus constrains the mind's effectiveness in scanning large amounts of data, such as many maps sheets, or examining complex features, to determine the shortest path through a very intricate network, for example. Again, there must be a trade-off between the explicit definition of spatial relationships in digital spatial data management systems, and the use of display to permit human observation and interaction. This trade-off is not now understood clearly. Similarly, there must be advantages to be gained from increasing the pattern recognition capability of computers, perhaps through the use of array processing machines. These ideas will be explored further below.

It was suggested above that current digital data base systems might allow spatial relationships to be implicit in their data base structure and even, at some cost, subsequently calculated. The question arises as to whether the underlying structure of the existing commercially offered systems seriously inhibits or prohibits the implicit or even explicit definition of spatial relationships.

Most of the well-developed commercial data base management systems currently available assume that the relationships between entities can be described in structures based on network models, founded in graph theory, or on hierarchical models, which are a special case of a network model. Systems that utilize relational data structures based on the mathematical theory of relations are now being developed, but there are only a few reported instances to date of any such commercial systems being used to handle spatial data. From the limited evidence available, some comments can be made about the hierarchical structure. A data base management system employing hierarchical data structures was adopted for some of their files by the U. S. National Water Data Storage and Retrieval System.[1] The Groundwater Site Inventory File in that system currently contains inventory data describing the location, geohydrologic characteristics, construction and

1. Water Resources Division, U. S. Geological Survey. 1975. "WATSTORE - The U. S. Geological Survey's National Water Data Storage and Retrieval System" and "The National Water Data Storage and Retrieval System of the U. S. Geological Survey Users Guide" U. S. Geological Survey, Reston, Va.

production histories, and field measurements for approximately 250,000 groundwater sites. A total of 370 million bits of data are stored. The entities are points with related aspatial attributes. The locational information is minimal, consisting only of coordinates for the point locations of the grounswater sites. The locational data are essentially treated as aspatial values amenable to plotting, contouring, and straightforward forms of statistical analysis. The data are indeed spatial, but little is actually demanded in terms of spatial query. The file is, however, an example of a large volume of point data being handled by an institution with the aid of a data base management system, in a manner that fills the immediate needs of the institution.

In contrast, an attempt was recently made to use the same approach of hierarchical structure to handle data that described the boundaries and attributes of oil leases off the coasts of Mexico and California.[1] It was found that the hierarchical concept does not allow the definition of graphic entities other than points (and presumably dendritic patterns). Links in the hierarchical model are implicit; they do not have to be labeled, but between any two record types there can be at most one link. This can, to some extent, be overcome by using two or more hierarchies in concert, but only at the cost of data duplication. The hierarchical structure prohibits the asking of questions that involve items from disjoint records. This implies that definition of spatial relationships within a hierarchical structure is cumbersome in many cases and impossible in others, and calculation of spatial relationships inherent in the data is severely inhibited.

Data structures based on the network model arrange data in one or more interconnected graphs. Record types are used to represent the entities, and the "links" are used to specify the relationships between sets of entities. The network at once offers more flexibility than the hierarchical approach, but it imposes the burden of specifying every "link." Phillips[1] moved to a network structure for the representation of the oil lease boundaries mentioned above, but found that the data volume incurred by specifying the linkages between every node used to define the graphic polygon boundaries was prohibitive. An alternative schema was

1. Phillips, R. 1977. "A Query Language for a Network Data Base with Graphic Entities" University of Michigan, Ann Arbor, Mich.

devised based on a simplified block structure of the oil lease boundaries. This allowed a limited set of queries to be developed and the system was improved. The lesson that seems to come out of this experience is that present data base management systems employing a network approach are useful for small, simple sets of spatial data but are cumbersome for the storage and query of most of the data types common to topographic maps, for example.

Relational data management systems are a more recent development. They allow the results of formal relations theory to be applied to problem solution, but as yet none are known to have been applied to the task of handling spatial data.

The three approaches have been compared[1] in general terms, but not in terms of their capability to handle spatial data. There appear to be problems inherent in adapting some of the existing data base management systems to handle large volumes of spatial data. There is no clear understanding of the relative applicability of the various types of data structure inherent in existing data base management systems to the problem of specifying spatial relationships. Also, as mentioned earlier, there are substantial methodological deficiencies in defining spatial relationships themselves.

Many of the questions raised so far could be regarded merely as interesting areas for academic study, except that answers to them are needed before any sensible plans can be laid for making large volumes of digital spatial data economically amenable to query. In the interim, such volumes of digital spatial data are accumulating in numerous agencies.

There is a tendency to use the data base management systems that have already been acquired and supported by an agency, simply because they exist. Similarly, data tend to be stored in archival formats, which are related more closely to the method of producing the data than of using them, because it is assumed that the user can perform the necessary reorganization of such "pure" data. There is an appealing logic in this approach. The user of

1. Date, C. J. 1975. "Relational Data Base Systems: A Tutorial" Proceedings, 4th International Symposium on Computers and Information Science. Plenum Publishing Corporation. pp. 37-54.

the data can presumably specify the types of query more clearly than can the data gathering organization. If the user reorganizes the data, he clearly has an interest in organizing them efficiently. The fallacy in the approach is that for many types of multi-purpose data, for example, the LANDSAT digital imagery, there are many more users than there are patterns of enquiry, and each user is faced with the task of reorganizing archival data. Repeated efforts, for example, must have been expended by innumerable users in many research centers to re-orient the LANDSAT data spatially and stretch them numerically to overlay a standard topographic map. This surely has placed a substantial burden on the use of the data and is typical of the multiplication of overhead costs that occurs when data are provided to many users in forms that are not amenable to query. Clearly a trade-off is possible between a distributed responsibility for data organization and the centralized provision of data organized for efficient query. That trade-off, however, can occur only when the agencies concerned have a much better understanding of the relationships between data structure and query. Unfortunately, the volume of data that exists is already large and there is a commitment to further growth. When large commitments of funds and staff have been made in building a specific data organization, it is difficult to reverse the process. There will be a natural tendency to try to work with the data bases that have already been created, rather than to reorganize them. This will limit the number of queries that can be economically answered from the data, and therein lies the constraint that poor data structures place on future investigation.

In summary, the underlying methodological problems and difficulties of using modern methods for handling spatial data are as follows:

1) There is no widely accepted and clearly defined set of spatial relationships between entities.

2) There are no clearly identified categories of spatial query which can be specified in terms of the operations that they require to be performed on spatial data.

3) It is not clear whether the use of modern data base management systems is inhibited because of the present imprecisions outlined in 1) and 2) above, or whether the relationships and

queries are adequately defined from a user's standpoint and it is the technology of data base management systems that is inadequate.

4) There is no clear understanding of the relative applicability of the various data structures inherent in existing data base management systems to the task of recording spatial relationships.

5) There is little understanding of the relationship between the need for explicit definition of spatial relationships in digital data base management systems, and the use of display to permit human observation and recognition of relationships.

6) There is little understanding of the relationship between the need to specify spatial relationships explicitly for data base management systems and the calculative capacities of present and future computers.

7) There appears to be no competent source of advice within the profession of geography or elsewhere which can provide answers to these questions at this time.

The remainder of this paper will suggest several initial steps that can be taken to address these problems, that in turn may lead to lines of investigation with potential for their solution.

Proposed initial investigations

The following approaches are mentioned sequentially, but they must be regarded as interrelated and interdependent topics. They may draw on activities previously considered in one or more disciplines, which, because they were not considered in context with problems of more general concern, were perhaps thought to be esoteric within the host discipline and thus attracted less attention than they deserved. It is important to recognize the relationships between the various aspects of the work and bring them together.

1) A clearer definition of spatial relationships can be drawn from work already accomplished in the field of picture processing and

pattern recognition. Schwebel and McCormick,[1] as part of work on scene analysis, provided a focal point for developing a taxonomy of spatial relationships. They examined one axiomatic characterization of such relations, namely, how stable are they under various mathematical operations on the related entities. If one wishes to structure a data base to handle different types of spatial relations, perhaps the most crucial aspect is the ability to handle mathematically different types of relations and it may not be necessary to worry about every possible semantic interpretation of what the relation represents physically, as long as the syntactic properties can be captured.

Very useful contributions to the semantic problems inherent in defining spatial relationships have been made in recent years in the fields of linguistics and cognitive psychology. Clark, Carpenter and Just,[2] and Clark and Chase,[3] among others, have examined the semantics of reasoning and the process of comparing sentences with pictures. Workers in the fields of architecture and structural engineering, notably Winston[4] in the MIT project MAC, have made a systematic attempt to define spatial relations as expressed in words defining scene descriptions. Workers in picture processing, in particular Freeman[5] and Haar,[6] under the guidance of Azriel Rosenfeld at the University of Maryland, have described mathematical and computational expressions which can be used to embody the semantic content of these terms. They have also experimented with constructing maps from relational scene descriptions so encoded.

1. Schwebel, J. C. and McCormick, B. H. 1970. "Consistent Properties of Composite Formation under a Binary Relation" Information Sciences, 2, 179-209. 2. Clark, H. H., Carpenter, P. A. and Just, M. A. "Semantics and Perception" in Visual Information Processing, W. G. Chase, Ed., Ch. 7, 311-381. 3. Clark, H. H. and Chase, W. G. 1972. "On the Process of Comparing Sentences Against Pictures" Cognitive Psychology, 3, 472-517. 4. Winston, P. 1970. "Learning Structural Descriptions from Examples" MIT Project MAC, TR-76. 5. Freeman, J. 1973. "The Modelling of Spatial Relations" Report TR-281, GJ-32258X, Computer Science Center, University of Maryland, College Park. 6. Haar, R. 1976. "Computational Models of Spatial Relations" Report TR-478 and 1977. "Generating Spatial Layouts from Distance and Bearing Information" Report TR-528, MCS-76-23763, Computer Science Center, University of Maryland, College Park.

The extension of this line of investigation could lead to an internally consistent classification of types of spatial relationship, expressed in mathematical terms, which in turn could lead to a more precise definition of specific spatial relations between entities. The relationships between data base structures and types of spatial relationships could then be more clearly examined.

2) When one approaches the task of defining categories of spatial query, one finds that the epistemological foundation in geography is less than firm. Gale[1] takes the view that although questions themselves may be taken simply as the primary realizations of an inquiring mind, needing no further justification, what does require definition is a) the language chosen for description and inference, b) the specific axioms or assumptions, and c) the method of judging whether the elements of a theory satisfy the specific inquiry. These, in turn, are not completely free choices; they are functions of the subject of the question, the kind of theory under consideration, and operational concerns (for example, information processing). The types of questions, the overall pattern of enquiry, are to a considerable extent dependent on the criteria that the investigator sets up to determine whether the answers are reasonable and satisfying. The characteristics of query are thus directly related to our experience or perception of the world in which we live and, using Lowenthal's[2] phrase, to the "geographical imagination" in providing concepts and principles upon which to build a common geographical epistemology. In any thorough examination of the nature of geographical questions, it is hard to avoid some discussion of these issues, yet this interface between methodology and philosophy is difficult to write about without either a careful exploration of the meaning of experience itself or the making of substantial presuppositions. I suspect that the same may be true in other natural and social sciences.

It is possible that the next increment of progress toward the definition of types of spatial question, and the types of operations on spatial data, may be made initially through empirical investigation.

1. Gale, S. 1975. "Simplicity, Again, Isn't That Simple" Geographical Analysis, VII, 4, 451-457. 2. Lowenthal, D. 1961. "Geography, Experience and Imagination: Towards a Geographical Epistemology" Annals of the Association of American Geographers, 51, 241-260.

One can approach the task by working either from operations to classes of query, or from classes of query to operations. A useful step would be a thorough examination of a series of existing systems of spatial data handling, to identify clearly and in detail the types of query that have been recognized to be needed, the types of operations that have been used in responding to these queries, and the types of data structure already found useful. The work could proceed from recent descriptions of several geographic information systems undertaken by the IGU Commission on Geographical Data Sensing and Processing.[1] However, it would require a considerably deeper examination of the systems than has been undertaken thus far.

As a parallel step, it would be useful to bring together some of the lists of operations that have been developed by various workers. Tomlinson[2] differentiates between logical operations and physical (computer) operations and between logical operations and "data manipulations." A logical operation is described as a change in data value, a comparison, or a movement of a data element. A change in data value, for example, can be accomplished by any of the physical operations of addition, subtraction, multiplication, or division, either singly or in sequence (multiplication and division are in themselves a sequence of the physical operations of addition or negative addition). The result of the sequence of physical operations, however, is a logical operation, the change in data value. The ranking of data manipulations is based on the increasing number of logical operations they contain. The simplest capability is basic data retrieval, a single logical "Move" operation. The second level is the result of two logical operations, and includes data manipulations such as summary, elimination of linear distortion, classification change, selective search, scale change, projection change, or measurement of straight-line distances between points. Data summary, for example, is achieved by the combination of the logical operations "Move" (to acquire data), and "Change in Data Value." Five other higher levels are recognized, each containing manipulations that require increasing numbers of

1. "Computer Software for Spatial Data Handling," "Computer Handling of Geographical Data," and "Second Interim Report on Digital Spatial Data Handling in the U. S. Geological Survey"
2. Tomlinson, R. F., ed. "Environment Information Systems", Proc. Unesco/IGU First Symposium on Geographical Information Systems, Ottawa, 1970.

logical operations. Existing geographic information systems and the types of query that they can handle are classified with respect to the rank of the manipulative capability, volume of data, and type of location identifier employed. Recent work by Peuquet[1] has established a list of operations (data manipulations) involved in handling spatial data in raster structures. Goodchild[2] has briefly listed operations (data manipulations) relating to the types of spatial entity (points, lines, areas) being handled. It would be interesting to see how the nature of the operations (data manipulations) varies with the type of logical schema being addressed and with the type of entity being handled. It would also be valuable to see if there is any correlation between levels of operations (data manipulations) and types of query.

Little work has been done on defining the categories of question inherent in any specific pattern of enquiry addressing earth data. In simplistic terms, there are two types of query. The first is satisfied by a descriptive answer to the questions asking what, where, and when (past or present). The second is satisfied by an explanatory answer to the questions how and why (past or present) and, by extension, what, where, and when in the future. The second is central to scientific inquiry and subsumes the first. It is possible that most stores of data are structured to service only the first type of query. Gale[3] has suggested a four-part partition of kinds of enquiry which may form the basis for identifying the categories of query that can be handled by existing data banks. He recognizes a) descriptive, b) normative, c) strategic, and d) evaluative types of query. Descriptive questions are essentially the first type of query described above. Normative questions are concerned with what ought to be. They contain assumptions about goals, which in turn influence how goals should be measured and the structure and content of data banks needed to facilitate such measurement. Strategic questions are somewhat different. They concern the organization of rules that govern the behavior of the entities, the way in which the data may be used. They require that

1. Peuquet, D. 1977."Raster Data Handling in Geographic Information Systems" Harvard University Symposium on Topological Data Structures for Geographic Information Systems , Oct.
2. Goodchild, M. J. 1977. "Geographical Data Elements" Mimeo.
3. Gale, S. 1977. Personal Communication.

data be organized in a manner that reflects the constitution or framework of the institution (discipline), so that answers can be derived that are acceptable in terms of the role of the institution (discipline) and that can be translated into practice. Evaluative questions concern the measurement of relative performance, the monitoring of activities, and the learning based on that experience. It would be interesting to determine whether the queries asked of existing data banks can be seen to fall into these categories, and whether these categories of query demand mutually exclusive logical schemas of data. Similarly, it would be interesting to find out whether the categories of query each employ essentially the same suite of operations for data manipulation, or whether there are any significant subsets of operations related to different categories of query.

3) The applicability of the data models in existing data base management systems to the problems of handling spatial data was called into question earlier in this paper. They obviously cannot be used for all kinds of spatial data, but that does not mean that they are not appropriate for particular environments. Drawing a line between realistic possibilities and wild expectations is very important at this stage of their development. If an application can be handled by an existing data base management system, it is very costly and unwise to "re-invent the wheel" in a specialized spatial data system. On the other hand, the introduction of a particular data base management system without proper analysis can increase the cost of data manipulations and hamper future applications.

Given a clearer understanding of spatial relationships and types of query from the previous lines of investigation, it may be possible to start to identify logical schemas that can usefully be employed to organize spatial data for various types of query. The ease or difficulty of fitting such logical schemas to the hierarchical, network, and relational approaches of existing data base management systems might then be assessed. The resulting effect on accuracy and geographic resolution could be evaluated with respect to the consistency that should be true for the data. Data language sketches of each type of operation (data manipulation) could be generated and the ease or difficulty of relational operations performed on the data within each approach could be assessed.

This line of investigation can be extended to an evaluation of

specific existing commercial data base management systems. For each system to be assessed, a data definition language program could be written for each logical schema. A data manipulation language program could similarly be written for the suite of operations (data manipulations) associated with each logical schema. On the basis of these programs, the existing systems could be compared with respect to their facility for handling each logical schema and type of operation. As a first step, the evaluation could be based on the ease of programming, from which can be extrapolated comparative costs and performance. Subsequent evaluation could be based on benchmark tests. The primary objective of this line of investigation is to establish which applications can be handled well by existing systems. However, where substantial problems are observed in fitting certain logical schemas and types of operations to the existing approaches, recommendations for new data models and data languages could be made in a more specific way. Criteria for establishing a new type of data base management system might be one outcome of this line of investigation.

4) There remains the uncertainty of whether our present difficulties arise not so much from our current ability to define the characteristics of spatial relationships and types of query as from an inadequate level of development of data base management systems. It is reasonable to ask whether there are lines of investigation that would provide data base management systems with more flexibility, perhaps with the ability to handle fuzzy definitions, and perhaps with the capacity to reorganize their own data structures in response to frequently used types of query.

One line of investigation can perhaps proceed from work on cognitive structures being undertaken in the fields of linguistics,[1] cognitive psychology,[2] and artificial intelligence.[3] Hays[1] suggests that human information structures can be viewed as a series of layered networks (systems) where a construction in one becomes an elementary unit in the processes of the next. In this way a human can proceed from the sensory monitoring system to abstract ideas through a series of networks, each actually describing the

1. Hays, D. "Cognitive Structures" Unpublished mimeo. 2. See for example Olrich Neisser, "Cognitive Psychology." 3. See for example Stewart Dreyfus, "Artifical Intelligence."

next at a different level of resolution, abstraction, type of thought, and degree of belief. Fuzzy concepts are handled at different levels from those of ordinary concepts. As pattern matching is implicit between networks, an inferential capability is provided. The human intellectual process can thus be thought of as regulated iteration between the networks, and creative thought results from the association of concepts in different networks. This type of structure may be useful for the design of data organizations to allow answers to the questions of how and why. It is probably already possible to model digital simulations of the networks, but the relationships between those networks may contain such complex computations that a much more powerful computer architecture than that currently available would be needed to simulate them elegantly.

In existing data base management systems, any substantial change in the logical schema requires a total rebuilding of the system. As there is a direct relationship between the design of the logical schema and the type of questions asked, a major change in the nature of questions implies substantial redesign. However, work is being performed by Merton and Fry at the Data Translation Project at the University of Michigan on the dynamic restructuring of data bases. Techniques now exist that allow a system to recognize the nature of frequently made queries and automatically to establish related files that contain the data for adequate responses. It would be interesting to determine whether these techniques could be incorporated in a data base management system either to reduce the necessity of defining all types of queries at the outset of schema design or to improve access to data thereafter. Recent developments in the Data Translation Project also include the concept of an "aggregate schema." This is not the same as a CODASYL subschema capability, but actually allows separate schemas to be merged so that the user can retrieve according to the combination of two or more individual schemas or views of the data. This resembles Hays' view of human coordination of concepts from separate networks and may represent the next step in improvement of the design of data structures.

5) The path tracing concept, inherent in all existing data base management systems, is designed to maximize the efficient use of existing computers, which are sequential processing machines. This concept may indeed be a fundamental limitation of current

technology. An important line of investigation must be the effect of replacing path tracing with pattern analysis, and sequential processes with array processes, both as separate steps and in combination.

One of the few areas where the human mind is demonstrably more efficient than existing computers is in pattern recognition and picture processing. It is not known whether the human subsequently processes such data on a raster basis, but at least the visual sensory input originates as a raster of excited retinal cells from which measurements and comparisons are made. Peuquet[1] has asserted that raster processing of spatial data has substantial benefits over sequential processing, and she has closely examined a suite of raster-based operations and the algorithms associated with them. In Hays' concept of human data structures, the networks are effectively low-order patterns and the relationships between them are pattern-matching operations.

Undoubtedly the area of pattern analysis is the one in which the human is effective, provided that the patterns are behaviorally established in the mind and the pattern can be mentally accommodated.

The limitations of human channel capacity and experience possibly constrain the degree to which pattern analysis can be used as a surrogate for explicit definition of spatial relationships in a digital data base management system. The outcome of that inquiry will probably depend on the availability of array processing computers and the replacement of human pattern analysis with machine pattern analysis.

Array processing computers are already in existence. The Goodyear STARAN machine and the CDC STAR machine are string processors which act as array processors. Complete array processors are under development in the United Kingdom and in North America. Their capabilities have only minimally been applied to the problems addressed in this paper. but the possibilities are substantial. One of the reasons why better data organization is

1. Peuquet, D. 1977. "Raster Data Handling in Geographic Information Systems" Harvard University Symposium on Topological Data Structures for Geographic Information Systems, Oct.

required is that large data volumes are expensive to handle on existing computers. There is every reason to suggest that those at the forefront of computer architecture design have the opportunity to examine thoroughly the current difficulties in organizing spatial data for economical query and to contribute to the answers.

The use of array processors with substantially improved memory capacity and improved ability to move data between memory and processing capacity may lead, with other lines of investigation, to the design of new data base management systems for spatial data. These, in turn, may prescribe new languages of dimensionality for the adequate description of spatial phenomena.

Certainly there is a pressing need for the current difficulties to be resolved by concerted research effort in several fields if there is to be any sensible planning of ways by which large volumes of earth data can be made economically amenable to query.

DR. MARBLE: Thank you, Roger.

Our next presentation is by Dr. Richard Phillips, of the University of Michigan, and Dr. John Sibert, from the Los Alamos Scientific Laboratory, who are going to discuss a cartographic query system for management of off-shore oil leases, in which they try to implement some of the ideas that Roger has talked about. The paper will be presented essentially in two parts, and both authors will speak.

A CARTOGRAPHIC QUERY SYSTEM FOR MANAGEMENT OF OFF-SHORE OIL LEASES

A CARTOGRAPHIC QUERY SYSTEM FOR MANAGEMENT OF OFF-SHORE OIL LEASES

DR. MARBLE: Thank you, Roger. Our next presentation is by Dr. Richard Phillips of the University of Michigan, and Dr. John Sibert, from the Los Alamos Scientific Laboratory, who are going to discuss a cartographic query system for management of off-shore oil leases, in which they try to implement some of the ideas that Roger has talked about. The paper will be presented essentially in two parts, and both authors will speak.

DR. RICHARD PHILLIPS: Could I begin with the first slide, please, which is the title of the paper that Duane Marble just quoted to you. I want to use that as a lead-in to my remarks, because it is an unfortunate choice of title. It unfortunately connotes a specialization of application which is really not present in the system we are going to describe. There are two terms in the title I do not like. "Query system" has an implication of a fairly static repository of data, about which the user can only ask a series of structured questions; that is not the case. Also, the fact that I have put in the specific application, that is, off-shore oil leases, leaves the mistaken impression that we have tailor-made the system to handle only that type of data base; it is quite general. Dr. Tomlinson talked about so many geographic information systems that have died for lack of use--probably in many cases because they have been developed for a specific application or have been doomed to failure for a variety of other reasons. In fact, the system that we are going to describe is indeed a generalized data base management system, just of the type that Roger was talking about.

If I could just perhaps recount a couple of the things that he has said, and just remind you what the term "generalized data base management system" has come to mean today. First of all, we are talking about a very large collection of data. Now, large can take on a great many meanings, but we are talking about such a large collection of data that we cannot build a single specialized way of handling that data and expect it to work on any data selection. It also implies that we are never going to have a main memory-resident data set. We are going to have to develop techniques we can use to efficiently extract data from relatively slow secondary storage devices. Also there has to be a rich query language. The user should be able to form a variety of fairly sophisticated criteria for extracting data from the data base, but, more than that, a generalized data base management system must allow the user the capability of modification; he should be able to delete items that are in the data base, add new items, alter the items that are presently there--all contingent upon some security overseer who is deciding who can do which of those

particular operations. Then, in addition, we usually think of a reporting capability being present in any generalized data base management system. This can be a tabular report or, as we will see, we can consider the system we are going to talk about today as having a graphical reporting capability. Dr. Sibert will tell you in a moment that the final product of this generalized system in many cases is a thematic map which summarized the queries that the user has asked of the system. Could I have the next slide which summarizes a couple of points about data base management systems. A data base management system is, after all, a collection of entities--and I will try to stay away from the data base jargon. I do not want to bore you with that. But we generally talk about a collection of entities, and these could be employees in an employee data base, they could be cities in a population data base--anything that we can consider as being important in the collection of data with which we are working. The attributes that are associated with these entities are usually scalar attributes--and here I simply mean a single value associated with perhaps a city and its population or a city and its name.

What is important in this system that we are discussing today is what I call graphical attributes. This term connotes both geometry and topology. Graphical attributes are important in the system that we are describing today because I need to have the capability of not only asking for a query based upon the scalar attributes, but I want to issue queries based upon adjacency; for example, show me two geometric entities or cartographic entities that adjoin one another, or give me the results of a query that is based upon a neighborhood or a vicinity consideration. And therefore I need to have these graphical attributes associated with this data base management system as well. Computer graphics plays a very important role in this system we are describing. It is by no means an afterthought. It does not take a back seat to traditional querying and produce after-the-fact graphical results as a postprocessing step. It plays a vital role in the system in several ways. First of all, and probably most obvious, computer graphics gives us this important window into a complex data set that allows the user to not only print a series of numbers that result from a complicated query, but also to take a cross section, a cut through the data base, and show a two-dimensional profile of the data base. In addition to the obvious display of data that is there, we also talk about the use of computer graphics for input operations. I have already given an idea of that in that we want to be able to use graphical attributes like adjacency and nearness. We want to be able to use those kinds of concepts as part of an input query, so graphics plays an important role in the input to this data base in forming queries. One of the most

important uses of computer graphics will be discussed by Dr. Sibert, and that is thematic map production.

Roger talked about data models and the familiar terms hierarchical, network, and relational. When we set out to build this data base management system, we tried to re-invent as few wheels as possible. In so doing, we tried to use the machinery that data base people had developed over the years, to make use of the tremendous amount of money and effort and research that has been expended in data base management systems. We searched for the most obvious data models that would suit our needs. Roger mentioned hierarchical, the tree type of approach. A natural application of that would be in describing an employee data base in a company or in describing the structure of an airplane--a fuselage with wings connected to it, and a tail connected to the fuselage, and so on. The generality that one can achieve with a hierarchic or tree structured data base is certainly not enough for the type of data base management system that we sought to develop, primarily because it does not provide the capability of expressing these topological relationships that we want to use both in querying and in display of the information.

A network system on the other hand does permit these relationships. If you are familiar with graph theory you will think of a network model of a data base as simply the entities, these generalized quantities that I have talked about earlier, all connected by arbitrary paths. The paths imply relationships among the data entities. Each of the entities then has its attributes. If we can construct an arbitrarily complex relationship among all of the entities, we should be able to achieve the generality we are looking for. In fact, we did decide to use a network approach. Roger mentioned the relational system where one does not impose any structure on the data; the developer of the data base does not even think of relationships among data. In fact, he allows the user to simply express all of the queries in terms of set operations, where he forms unions and intersections of sets of entities which have certain things in common. The network system proved to be the most flexible for our application. Then we started to look at the commercially available systems. Roger already mentioned some of the shortcomings of commercial systems.

Generally, a data base management system consists of three major software modules. There will be the data description module, where the developer of the data base can, with a fairly simple language, express the relationships among all of the entities that are going to be in the data base. Then there is the data manipulation language which, once the data base is built, is used to do the actual extraction of information from the data base, to do the

modifications that the user subsequently requests. How does the user do that? Well, the third major software component is a query language module, and this is where the user can fit together the criteria that he wants to use to extract information, to make modifications, and to generate reports. The system that was used is called ADBMS, by the way. It was developed at The University of Michigan and is an acronym for A Data Base Management System. It has the attribute of being written entirely in FORTRAN. The standard network model, which was proposed by the Committee on Data Systems Languages, which has come to be known as the Data Base Task Group, developed a COBOL interface. Anyone who has ever tried to do any programming in COBOL, which is other than business-oriented, knows the difficulties in trying to do things like draw a square, for example, or express any other sorts of geometric relations.

ADBMS does have the advantage that it has a FORTRAN interface, but it does not have a query language associated with it. The rest of my remarks will deal with the major work that we did on this project, and that is developing the query language module which could use the data manipulation language and the data definition language which is inherent in ADBMS.

Figure 1 shows what is called a data base schema. I again will refrain from dwelling on the specific application that was the impetus for the development of this system, but I just wanted to show you a graphical representation of a schema. This is what this particular data base looks like to the developer of the data base. Each of the rectangles represents one of the entities that can be in the data base. We have color coded these just to distinguish the graphical entities, which are in green, from the ones that have only scalar attributes. This is an oil lease data base. If anyone is interested, we can talk about the details later. But the lines joining each of the rectangles represent the network configuration, the implied relationship that exists between each of these entities in the data base.

The important thing that had to be done in developing the specific system that we are describing today is to develop a query language that would totally shield the user from this structure; he should not have to know anything about the implied relationship among each of these entities. He should only know that all of these entities are present for him to interrogate, and he should be able to form a query which involves any of these entities and any of the attributes of these entities without regard for how tortous the path may be to get from one of the entities to the other. He should be able to blithely say, "find," and state his query, and have the system do all the work for him. And that is really what

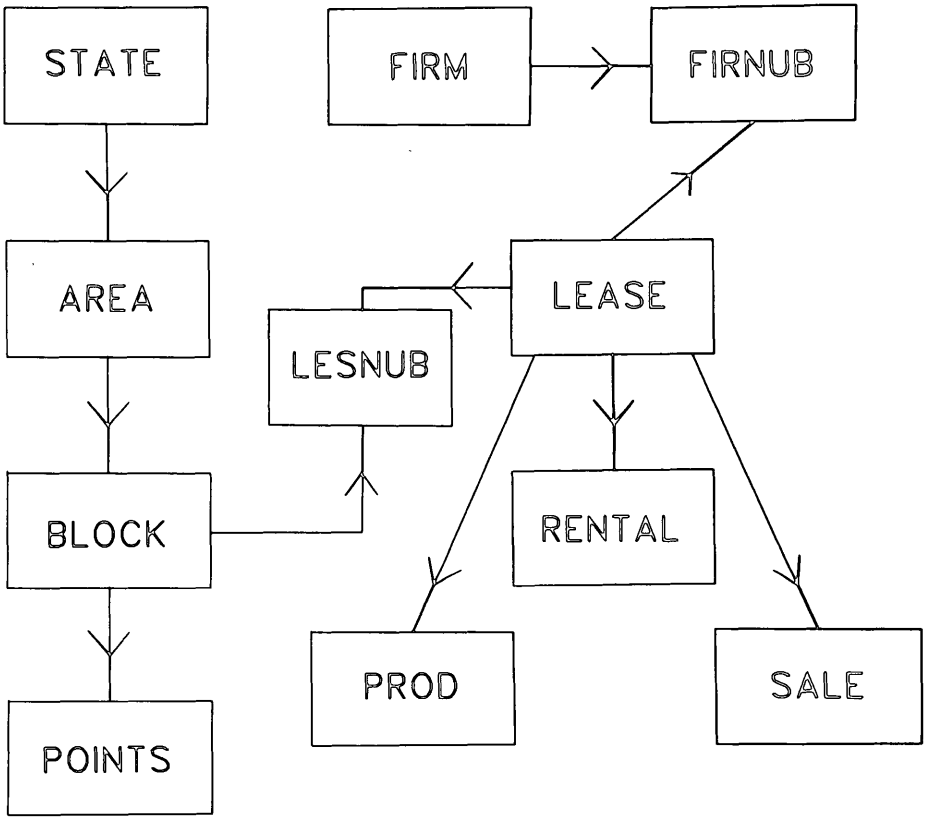

Figure 1. Data Base Schema

we set out to do and I think we have accomplished. In describing the query language, I will just briefly make a couple of points. The idea, as it always is, is to make it English-like so that it is easy for an untrained user to learn the language. More than that, it should be easy for someone who knows it well to abbreviate it. All of these attributes have been built in.

Path finding is the operation I just described of actually being able to find the way around the schema based upon the query that is specified by the user. Generalized accessing simply means the

system had to be able to, without regard for how many attributes each of these entities had, find all of the occurrences of them that were implied by the user's query. As a query example, a user might say FIND LEASE WITH SUM (OILP) > SUM (GASP). In place of the word "lease" you can simply substitute any of the other entities found in the schema in Figure 1.

SUM happens to be a function that operates on one of the attributes, oil production, for one of the entities. You can replace SUM with any other function that seems reasonable to you--square root, average, or whatever. We are saying find all the leases in this system with the sum of the oil production greater than the sum of the gas production. Once that is done, we would like to graph the results. Just to show you what the graphical output would look like, consider the next slide. The user will eventually have the capability of completely defining the graph background in order to produce, if he wishes, report-quality output. He will be able to represent the data on the graph with histograms, bar graphs, regressions, or whatever he wishes. This slide is an example of a graph that has been produced by this system as it stands today. It shows the oil production, for a particular lease that had previously been selected, as a function of year of production. I want now to turn this over to Dr. Sibert who will talk about one of the most important aspects of this system, that which allows us to produce thematic maps.

DR. JOHN L. SIBERT: Thank you Dick. I am going to have to be a little bit more specific about our system because the maps do not make too much sense otherwise. They are specific to our data base application. I would like to begin with a little background about the geographic definition of the outer continental shelf leasing survey. This is similar to the public land survey, with which most of you, I am sure, are familiar. In fact, in most cases the public land survey has been extended into the water to include the outer continental shelf area. It is organized in several levels of agglomeration. The largest level is called an area. It is analogous to a country in size. The basic unit of the survey is called a block, and is approximately equal to a quarter township in the standard public land survey. In addition, for each block in all of the areas--currently we have Louisiana and Texas offshore areas--we have stored the latitude and longitude of the block's corners. In most cases that is four corners. Some of the blocks are irregular in shape and have more than four. Here is a sort of pseudomap portraying one of these areas (slide not available). You will note that one of the squares or blocks is shaded in red, so this is the approximate size relationship between the block and an area. The coordinates we have, again, represent the corners of all of the blocks.

When we wish to produce a map we assume that we have already retrieved a set of leases according to the sort of criteria Dr. Phillips was talking about. Normally, for this particular application, what we want to map are the leases since we are interested in managing the off-shore leases. Obviously, the system is much more general than that. The first step then is to link from these leases to the blocks in the survey which contain the leases, and then linking from each block to the coordinates so we have a complete definition of the blocks. Finally, we must decode the legal description, because when tracts are leased originally they are defined by a legal description, and, in many cases, the description is straightforward because it is simply block number N, area number M.

What I just described is illustrated in Figure 1. After finding the lease record, you will notice that there is a link through a thing called lesnub to the block record. The nubs are artificial records which allow us to do many-to-many linkages. So there is a link from the lease, to the lesnub, to the block, and we can link backwards up from block all the way to state and downward from block to points. We need the lesnub because some leases are a little bit more complicated in their legal description and, in fact, consist of more than one block or parts of more than one block or more than one part of one block. The description is given in simple English in a manner that is probably familiar to most of you; for example, the southeast quarter of the northwest quarter of block number N.

In order to protray the leases accurately, we must be able to store that legal description in the data base, link it to the appropriate lease, and then decode the description so that we can draw only the part of the block that actually is included in the lease. As an illustration, in order to decode the above description we first bisect the north and west sides of the block. It is easy to compute, and gives us the northwest quarter. We repeat the operation for the southeast quarter of the northwest quarter. Having done that for all of the leases we are interested in mapping, we are then ready to portray them. We have several different mechanisms for doing this. Before I describe them in more detail I thought I would mention the hardware we use.

The data base management system is currently resident on a CDC 6600, and (for anybody who is interested in that kind of detail) occupies approximately 140 K octal words of core storage. The data base itself resides on a disk. We are currently involved in building a new version of the data base with data for '75 and '76. The version we have now only has data through '74. We expect that quite soon the size of the data base itself will surpass a

megaword. The output devices we use are: 1. For immediate display, a Tektronix 4000 series cathode ray storage tube display. I imagine you are all familiar with them. There are quite a few over in the vendors' area. This allows us to protray a map or a graph immediately on the screen during the retrieval process so we can look at it, get a good idea of what we have. We can also modify the display by adding additional information. 2. We have as part of our task the production of relatively high quality color output. For this we need a somewhat different hardware device, the I.I.I. FR-80 with 35-mm color camera. This device is basically a PDP-15 minicomputer which drives a high-resolution, fast-phosphor CRT. A 35-mm camera with program controlled filters is mounted over the CRT. By changing filters on the camera, redrawing the map or other graphic on the CRT screen, and multiple exposing the film, it is possible to produce a variety of colors on the output.

We allow several mapping options, the most popular are called new, old, lease, and area. New erases the screen or advances the film before the map is drawn, so it is pretty obvious what it does. Old adds additional material to the map that has already been drawn. The lease type map portrays only the leases themselves, while the area type map draws in the survey lines as a sort of background grid system. I have several examples of these maps. The first is a lease type map. You can see across the top of the picture the coastline of Texas, then Louisiana. The leases are portrayed as little red filled-in squares. These happen to be all the leases that had produced anything before 1975 in that area. As you can see, it would be very difficult from this type of map to identify a specific lease.

Figure 2 is an area type map which while retrieving the blocks also retrieves the areas which contain them, and portrays all of them as background. This is a much more useful form because now we can, particularly when we make a hard copy and look at it a little more carefully, identify specific leases and determine their location. However, if you want to portray more information about the scalar attributes of the data it becomes necessary to view a smaller area at a larger scale.

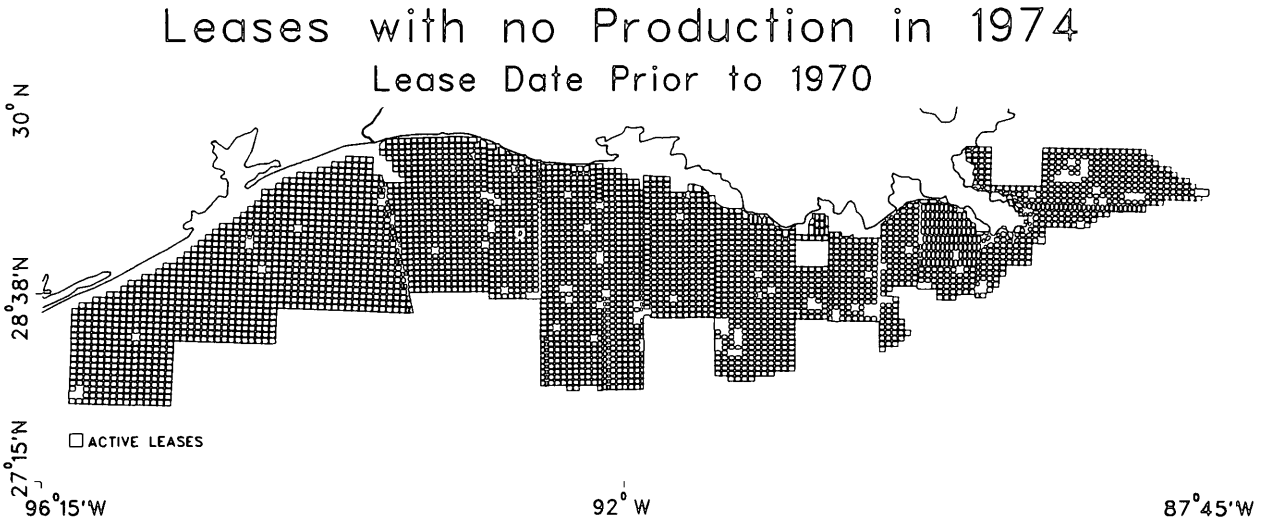

Figure 2. Area Map

DR. MARBLE: Thank you. We have a third member of our panel who is not going to make a formal presentation, but who is going to comment in part on the presentation of the other speakers in the light of his own experience, and this is Mr. Robin Fegeas of the Geography Program at the U.S. Geological Survey. Robin has been heavily associated with the development of the computerized land use mapping system, and we will have some comments from him. Robin?

MR. ROBIN FEGEAS: You must forgive me if my remarks do not seem well prepared. I discovered just about ten minutes before the session what my role was going to be. But, sitting here I had a few questions which have been bothering me for some time since we are just beginning to get into thinking about managing a large data base, a land use data base, for the entire country.

Just to go down some of these questions. The first: At what level of development are data base management systems? Are they adequate? Can we make good decisions right now? We have been told the hierarchical model cannot hold the relationships we need. The relational model is still very theoretical; no good examples have been yet brought into the market operationally. The network model, which Dick Phillips and John just discussed, seems promising, as their presentation showed, but it still puts a burden, namely in having to specify your relationships ahead of time, well-defined, and any relationships you might want to impose later on creates large overhead in update. Basically, the data base has to be restructured completely.

The second question I had was just how much of an overhead are we going to pay for using data base management systems? The objective, that of data independence, is a very admirable one, one which should allow us to use our data for many more purposes than if data independence did not exist. But, still, we have to pay a price in efficiency and, perhaps, in computer storage. I have been told at our computer center in Reston, Virginia, we just had installed a commercial data base management system, System 2000--which the Water Resources Division will be using. The computer center people are complaining that just this one application of a data base management system will drain the resources of the hardware and software there so that it will preclude other users.

Of course, this brings in another question: What role do minicomputers play in this data base management scheme? More and more, minicomputers are being used and will be used, and I think this is the wave of the future. So, where do they come? It appears the systems as they are thought of today to use a lot of resources which are not available on even large minis.

Another question I have is one of management. The use of a data base management system requires a new position, a new expertise, namely that of a data base administrator, one who can oversee all the requirements of the different people who will be using the data base, and then structure that data base accordingly so that all users can use it. At the Geological Survey we just were presented with this problem, and a position description has just gone out. We are going to get a data base administrator, perhaps. It is unclear as to what his powers will be, but all in the field or most in the field now agree that this position should be a very powerful one. In the present bureaucratic structure when you introduce a new powerful position, it is difficult. That consideration will have to be made. That concludes my questions.

DR. MARBLE: Your raised a number of interesting questions, Robin, which do need to be addressed. The particularly interesting one, this question of the organizational changes that come about when you start talking about the management of large quantities of data. It represents a change in view on the part of the organization. Many organizations, governmental and private, have used and accumulated large stores of data, most of which has been oriented toward individual users. I have my data, you have your data. Occasionally you may want to borrow mine and use it, and that is all right, and I will tell you about it, perhaps. But when you start recognizing that data within an organization--and I use the Geological Survey as an example, since it is a data-oriented organization--constitutes just as much of a resource to the organization as an individual company's buildings and trucks and aircraft, then you have to start worrying about how this utility is to be managed for the best good of the organization. It becomes necessary to institute an administrative structure. This notion of a data base administrator who has certain powers over the data, does not own the data any more than the bank manager owns your money, but he does have certain powers to regulate the way in which it is used. This is largely to prevent people from falling over themselves, and one side inadvertently changing portions of the data base while another side is trying to use it, of insisting on consistent definitions of data elements, things of this sort.

Robin mentioned the potential problem of resource use with a data base management system. I do not think that we should sit down and say that the data base management system is utilizing such a large volume of computer resources; the problem is that we have a lot of data. A case in point, the use of System 2000 by the Water Resources Division; there are very large data volumes involved and a large number of users, many of whom wish to access the information in an interactive fashion. This places a load on any computer facility that has to be dealt with in one fashion or another. If the facility is operating at or near capacity, even small additions

in demand can, of course, have great impacts. How about questions or queries from the audience?

MR. MITCH MODELESKI: Mitch Modelski from ESRI. Roger, I do know of one experiment where a relational data base management system was applied to geographic data, and that is geoquell. Are you familiar with this? Geoquell is a front end to GIRAS, which is a relational data base management system currently operational at Berkeley on a DEC 11/70, running under UNEX. This data base management system is written in C, a language developed at Bell Labs. I would like to contrast this particular experiment with another program with which I am familiar to demonstrate just a couple of examples. GIRAS was asked to draw 90 simple polygons once, and it took about five minutes to do so with a Tektronix scope. Many of the people that ran that experinent felt that it was a failure.

I would like to contrast this particular experiment with ARITHMACON, a program that Marv White is currently developing for the Census Bureau. Many people will say that the comparison is unfair because ARITHMACON is running on a PDP10 under Macro-11 Assembler, and an 11/70 can't hold a stick to that particular machine. However, I think the important difference lies in the way the data was modeled. When I examined the data structure of geoquell, it turned out that they were storing nickel records, namely, from coorinates to coordinates and right polygons, but no left polygons. ARITHMACON explicitly stores not only the from and to, left and right, but orders these with a set of relations that involve boundary and co-boundary across all possible combinations -- say, the boundary of a line, the boundary of an area, co-boundary of a point, the co-boundary of a line, and so forth. The type of queries that can be processed with ARITHMACON far exceed the types of queries that can be processed by the generalized data base management system whose geographic front end was simply developed after the fact to demonstrate that the marginal cost of an application would be lower, given a generalized data base management system. But the particular data model that was used in GIRAS was not appropriate to geographic data where the graph, to me, is the ultimate thing we have to be careful about. So, in closing, I guess I might just comment and say that we might be able to store all the relations for some of the data, and some of the relations for all of the data, but not all of the relations for all of the data. (Laughter.)

DR. MARBLE: Thank you, Mitch. (Applause) One of the

people involved in the study group that Roger mentioned was a computer scientist specializing in data base management systems, Dennis Tsichritzis of the University of Toronto. About a year and a half ago Dennis, Donna Pequet and I were talking about these systems. There was no real question in our minds that if we were going to handle really large cartographic data bases they were going to have to have to be handled in an efficient manner, and the data base management system approach is the way this is done in most areas. After some discussion, Dennis made a remark that I think is still quite pertinent. He said that in dealing with spatial data one of two things must be the case. Either you people (geographers and cartographers) really do not know what you are talking about, that your statements about spatial relationships and the things you are trying to do with all these points and lines and areas are poorly put together, and that if you sit down and try and think about it properly, you will be able to place the things you are doing within the context of existing data base management systems and they will work for you. Or, on the other hand, you may actually have something new, which from the standpoint of people working in computer science and data base management systems would be most exciting because, he said, we are getting awfully tired of yet another airline reservation system. (Laughter.)

Part of the work of the IGU study group has been to try and develop some insights into this area. I think that one conclusion that has come out of the work is that it is probably the latter case rather than the former, and that there are indeed some unique characteristics of spatial data. For example, within the concept of a data base management system it is the entities that are considered to have attributes and not the relationships, whereas in spatial data we frequently have attributes attached to the relationships themselves, such as distance. So we may very well be working in an area which is going to provide a great deal of interesting development for people in computer science as well as cartography, geography, and other areas in the earth sciences. Are there any other comments?

DR. AANGEENBRUG: I want to pursue this comment of yours about data management from a policy point of veiw. I would say, to venture a guess, that if you got a powerful data base manager in USGS somewhere, that unless there were complete cooperation from the top down in defining that job, and that job was made as unpowerful as possible,

it would not work. Data is not owned by a single person in a single agency. If you create somebody that has too much power -- and I am standing here having been such a person in the university, controlling the entire budget with a small computer, so to speak. It was the wrong kind of thing to do for a university, and if the system didn't crash, the power position did.

It seems to me that the function of a data base manager could be that of, say, a catalog librarian rather than someone who you have to get past. Because in many of the large agencies -- I will not name some that I used to work at -- it is rather difficult if the computer division controls the operational aspects of the division. After all, that is not why it exists. It is a facilitative thing. So I would suggest perhaps not at this conference but another, maybe in public administration, if you pursue this question, do not create a very powerful data base manager, would be my advice. A very capable technical one, yes. They should not have a very high rating, because primarily what you are dealing with is a conceptual model. The decision to standardize every chunk of information within the vast divisions of, say, USGS as a management-administrative decision must be made by the policy makers first. Otherwise it won't work. They will not let it.

DR. MARBLE: The concept of the data base administrator is one which is somewhat strange to many people. I will not try and elaborate on it here. There is an excellent book dealing with notions of data base administration as well as a very good shorter discussion in James Martin's book on Principles of Data Base Management. It is a position which is strangely structured administratively since the person is not just a librarian. The post combines the duties of a librarian, a technical standards committee, and a number of other things as well. Sid?

MR. SID WITTICK: Before I make a comment, I would like to ask Dick whether my impressions are accurate, that the application that you just described is operational and has indeed been successful on existing data sets.

DR. PHILLIPS: Oh, yes, it is operational. There are some planned enhancements to it, but it has been operational for six, eight months.

MR. WITTICK: That was my impression. I think it makes the point very, very well, that there are areas where we have been very successful, and it is usually when we try to take on a reasonable size task, a reasonable size vol-

ume of data in a temporal framework. I think we are guilty as a set of professionals in some instances of trying to operate at too large a scale. I think even in terms of very large problems posed by the USGS's and the Census's of the world, I wonder if in their manual systems, and, indeed, these are systems as well as any others, we can demand the same standards that we are demanding of the computer systems we are trying to create? For example, I wonder whether all the maps that exist within these mapping agencies currently are all consistent and uniform and standard in terms of their ability to be manipulated? They evolve through time, and I think we should start trying to design systems that we know are going to have to be replaced, data structures we know are going to be replaced, but such that we can save money while we try to do it.

DR. MARBLE: Sid, three large, burly representatives of the Topographic Division will be waiting for you near the exhibit area. (Laughter.)

MR. BOB RANDELL: I am Bob Randell from the University of Saskatchewan, but do not expect me to be as erudite as Dr. Boyle. I am just a biologist, but I am very much reminded of a situation that exists in one of Lewis Carroll's works where a country prepares a map which is a one-to-one representation, and then makes it illegal to unfold the map because it cuts out the sun. Now, how big does a data base get before it gets bigger than the original? (Laughter.) (Applause.) To what extent can you, especially now we have space platforms -- how much data do you need to store that you cannot obtain, say, at the next ERTS path, especially when a lot of these data are now available locally if you have just a small radio station? I know radio amateurs who can process ERTS signals.

DR. MARBLE: That is an interesting question. It was posed, I think, for the first time in print about 13 or 14 years ago in a joint paper presented to the American Institute of Astronautics by William L. Garrison, and it contains an illustration showing a rough globe of the earth entitled "World's Largest Data Bank." This is indeed a question one has to really address: "How much data does one want to retain? What are the necessary things to retain? What is the balance between current operational needs and long-term archival structures? We could very easily end up drowning in data, particularly as the direct digital data capture techniques increase in efficiency and the volumes generated from them escallate. Are we going to keep everything forever? That is a policy

decision, and one that tends to fall under the area of data base administration within an organization. There are a lot of questions in this area that are at the present time relatively unanswerable. But we are going to have to find operational answers to them within the next few years or find ourselves neck deep in difficulties.

MR. KEN PYLE: I am Ken Pyle, and I am from San Diego County. I would like to address the problem of a one-to-one relationship. Because if you look at what a county does in the way of mapping, with records kept currently in a totally disjointed, uncomprehensible and very often conflicting style, then I think you begin to realize that it is desirable to establish a base which does in fact represent a one-to-one relationship with the ground. If you are familiar with mapping at the city and county level throughout the country, I think you recognize that most mapping occurs in this disjointed fashion. There are little groups of drafting technicians squirreled away here and there and everywhere competing against one another in many cases, surely contributing to job security, but nevertheless representing a tremendous duplication and loss of the tax dollar. Now, we can very well go on and create automated systems in the very same fashion -- particularly with minicomputers coming through so quickly. But, what we will have again is a series of automated programs for each specific or specialized use within the city and county, none of which are compatible with one another, and all of which represent a tremendous waste of the tax dollar. Now, if we are going to actually resolve this problem and create a system that will be a cost-effective use of the tax dollar, then we better get everything together, put it on a one-to-one basis so it represents the real world, not a map world, and make sure that it can be used by all the necessary users.

From a manual standpoint you know this is almost possible today. We could create a series of large-scale maps with multiple overlays. In fact, in San Diego County we have an intermediate scale, a regional scale in which we have done that. We have 180, 200 overlays to a base map which you can put together any way you want. We have 50 people in my section doing mapping. We do not do any maps for ourselves. We do them for others. What we have to look at is the fact that, first of all, we need that one-to-one relationship because of the administrative work that we have to do on a daily basis that engenders the information in maps, and, number two, which is heartening to me because I see the trend happening here, that we have

to recognize that maps are merely a graphic display of the data that we wish to show, that we wish to use. We do not make maps for the purpose ofmaking maps. We make maps to show information. Consequently, we are talking really about data base management with the capability of a graphic display that comes out in the form of maps as well as others. From our standpoint we are working on the basis of one-to-one relationship. We are building our data base based on ground calculations, engineering calculations being direct input into the system. I think if we do not work that way, all we are going to do is build our own specialized little system that will serve my department's purposes, but certainly none of the other 50 departments in our county, because they will all be getting their own. So I am in favor of a one-to-one relationship.

DR. MARBLE: I think you are using the term one-to-one relationship in a non-standard and confusing sense. In a cartographic operation we tend to interpret the form of Lewis Carroll's map as one-to-one. Mr. Carroll mentions another map, in The Hunting of the Snark, which references yet another solution to our data problem. I believe it was the Bellman's map which unrolled contained absolutely nothing. Somewhere between these two extremes we must reach a balance. We have come to the end of our scheduled time. Thank you.