A Geographic Information System for Regional Planning

R. F. Tomlinson

Department of Forestry and Rural Development, Government of Canada

As a tool in its program of rural development, Canada is developing a computer-based information system for the storage and manipulation of map-based land data. The system and its capabilities are described.

Canada, like many countries, faces an immense problem in both understanding and guiding the development of its land, water, and human resources. One of the major agencies created specifically to implement policy to attack this problem is the Rural Development Branch of the Department of Forestry and Rural Development. A primary task facing this agency is to assemble social (demographic), economic, and land data for an integrated analysis to enable problems of rural development to be specified, development programs to be implemented, and their effectiveness evaluated.

Parallel with the gathering of data has been the development, by the Regional Planning Information Systems Division of the Branch, of interrelated computer-based information systems to handle and analyse the data. The Geographic Information System, for the storage and manipulation of land data is the most developed of these systems. Its design and development started in 1963, implementation began in 1965, and is now in its final stages; routine use is scheduled for September 1968. It is perhaps worthwhile to recount our progress with this system at this time.

Early in the life of the Branch (1962) a start was made with the gathering of some kinds of land data by the Canada Land Inventory. The data they collect is restricted to five types: the present use of the land, the capability of the land for agriculture, the capability of the land for forestry, the capability for recreation, and the capability for supporting wildlife. These data alone, if gathered in sufficient quantities for the summaries to be directly applicable to provincial and federal resource policy and regional planning, will generate an estimated 30,000 map sheets, at various scales. The Inventory has currently produced 7000 map sheets, of which 3000 have been prepared for computer input. The maps contain an average of 800 distinct areas on each sheet, and have been found to contain as many as 4000. Additionally, other types of maps covering watersheds, climate, geology, administrative boundaries, and land titles are generated by other agencies.

The need for a computer-based system, whereby map and related data can be stored in a form suitable for rapid measurement and comparison, is apparent as soon as the magnitude of the problem of handling large numbers of maps is appreciated. Lack of trained personnel makes it impossible to examine such large amounts of data manually in any sensible time, much less to provide a meaningful analysis of the content. A situation can be reached where the amount of data precludes its use. The end product of countless hours of survey can remain unused, with the result that administrators do not receive information necessary for a sound basis to decision making.

From the first, it was the intention to produce the maps generated for the Canada Land Inventory in such a way that their data could be related on a nation-wide basis by the geographic information system. This made it necessary to establish a common basis of data description. Classification systems were evolved for each type of data by discussions with the federal and provincial agencies concerned in the original survey, under the guidance of a federal coordinator. In each case, the classification systems were subject to trial in pilot areas in various regions of the country. Regional variations are incorporated into the classification system by development of ratings which recognize equivalent values. The classification systems vary from a relatively simple, one-letter code for present land use to a complex, multi-level description used for forestry.

The maps, essentially interpretations of existing data in terms of the classification system, are usually produced by the federal and provincial agencies most closely related to the collection of the original data (over 100 agencies are involved). The manuscript maps are sent to Ottawa to be edited and prepared for computer input.

The basic capability of the geographic information system is that it accepts and stores all types of location-specific information, that is, any information which can be related to an area, line, or point on a map. Information relating to land resources is most frequently location-specific in character. For example, census data (perhaps not usually thought of as location-specific) are collected from specific areas of land called enumeration areas, which are recorded on maps; a highway is a location-specific line; a campsite can be thought of as a location-specific point on a map.

The system can best be described as comprising two parts: the data bank and the set of procedures and methods for moving data into the bank, and for carrying out the manipulations, measurements. and comparisons of the data, once there. These two parts will be referred to as the 'data bank' and the 'information system', respectively. It is quite possible to have the entire geographic information system with full operating capability and have no data in the data bank. The amount of data which can be put into the data bank is infinite, as any number of magnetic tapes can be generated and stored. Additional data related to any area can be inserted at any time.

The system has the following capabilities:

It will accept maps containing data represented as areas or lines or points. The maps can be of any scale and on any map projection, and they can contain linear distortions. All of these characteristics will be adjusted to a standard format (normalized) when they are put in. Data relating to points only can be put in independently of maps. They are simply related to their latitude and longitude points.

The system compacts and stores information. The compaction is most efficient. For maps at a scale of 1 : 50,000 with an average density of information it is expected that a complete coverage of the farmed area of Canada (approximately 600 map sheets) can be recorded on two reels of magnetic tape.

The system can measure any data in the data bank. If the data have been inserted in the form of areas, then each area can be measured. For example, a soil map might be represented by different areas of different soils. The area of each patch of soil or the total area of any one type of soil can be calculated.

Similarly, the lengths of lines can be measured and the occurrences of points counted.

The region from which area, line, or point measurements are required can be limited in a variety of ways. Data can be retrieved within any boundary already described to the system. If, for example, a map of administrative region boundaries has been put into the data bank, measurements can be carried out within a specific administrative region. If a desired boundary has not already been described to the system it can, of course, be drawn on a clean sheet and inserted in the normal way, or if it is simple enough in shape to be described by a straight line joining points, then it is only necessary to put in the co-ordinate values of the points.

It will also be possible to limit retrieval by reference to any line or point already described to this system. The system can be asked, for example, to measure the area of patches of land crossed by the line of a highway or within a band of specified width along the highway, or to determine the areas suitable for sub-divisions within 20 miles of the center of a city.

A major system capability is comparison of two types of mapped data relating to the same area. Just as two maps can be manually overlaid to allow the relationships between the data to be examined, the system can overlay any two or more types of data to measure the exact amounts of each type of land in juxtaposition to the map or maps below.

This can be applied as a search capability, whereby a comparison of various types of information is made to find out where a selected set of characteristics occur together. For example, a request to find suitable landing sites for a helicopter would require an examination of the vegetation map to determine treeless areas, the topographic map to make sure that the area was flat, and the present land use map to make sure that the area was not populated. These three coverages would be compared to identify and describe all points having the desired characteristics.

A further extension of the search capability could result in a 'search in context'. A potential helicopter landing-site, for example, would be of limited value if, while being perfectly treeless, flat, and uninhabited, it occurred as an island in the middle of a swamp. The search routine can be instructed to ignore otherwise desirable sites if they do not occur in a desirable context.

Another search capability that can be implemented is referred to as the 'nearest neighbour search'. This would be employed when the limit of the search is not definite enough to be specified. The search command would simply request the nearest examples of the desired character to be located. A composite example of some of these capabilities might be an instruction to locate the nearest potash mine which is served by a main highway, north and south railroad connections, and is surrounded by a minimum of 10,000 sq miles of good farmland.

The system can produce information in two different forms. The commonest form is perhaps the normal printed alphabetical and numerical data produced on the regular computer printer. In addition to the printer will be a graphic plotter which, under the control of the system, produces a map showing the location of the desired areas, lines, or points which satisfy the request.

An inherent danger of information systems is that the data entered into the system may vary widely in reliability, but may be assumed to be equally reliable in subsequent multifactor assessments. The system can accept a reliability identifier with any type of information and can keep track of reliability tags so that degrees of reliability are printed out beside the answer to a request.

The advantages of information which is kept up to date, compared with data which have to accumulate for several years before it is economically desirable to reprint a map, are well known to users of map information. Data can easily be added to the system without waiting for large amounts of new data to accrue. Old coverage can be erased and replaced on the magnetic tapes or, if desired, both the old and the new coverage can be retained. New survey data at a more detailed scale can be incorporated with previous data at smaller scales, provided, of course, that the classification systems are compatible.

For many of the day-to-day information needs of administrators of land resource policy, simple forms exist to allow the administrator to initiate the request without the assistance of a computer programmer. Although more detailed assessments requiring the full flexibility and capability of the system would best be handled by someone acquainted with the data formats, a considerable amount of programming effort has been eliminated even at this level by use of programs already written and incorporated into the system. It is estimated that, with no previous computer knowledge, an administrator could be taught to complete normal form-originated requests in one week. Three weeks training and practice thereafter are expected to be necessary for the same administrator to handle more detailed requests. The unusual or very complex requests will need a programmer working in conjunction with the system librarian.

In many ways the system is self-monitoring. On accepting a request for information, the first response of the librarian will be to use the system's KWIC[1] index to check whether that particular request has been made before and, if so, to indicate where the answer is stored in the filing cabinet. If the request has already been partially answered, this also is determined. If the request requires new manipulation of data, the system indicates which tapes have the requisite data stored on them.

The tapes then are selected from the library, put on to the computer and the assessment is executed. An extension of this capability is to provide a cost estimate of the work, prior to processing, based on a preliminary analysis of the amount of data on the requested tapes. Such estimates will be necessary in more complex applications.

The system is independent of peripheral devices such as input scanners or output plotters. While the IBM cartographic scanner is now in use, in conjunction with a D-Mac X-Y digitizer, to convert graphic data to digital form, instrumentation is likely to be developed in the next two or three years to combine these functions.

The normalization step, which converts digitized graphic information to the format required by the data bank, is independent of the main system functions and can be changed accordingly.

The system is designed for use on the IBM System 360 Model 50, with 512 thousand bytes[2] of storage, 6 magnetic tape drives, and 3 magnetic disc drives under the control of the standard operating system. Greater operating efficiency is achieved if the System 360 Model 65 is used. The practical application of the data bank concept and the entire system capability is available by use of this general-purpose computer.

SYSTEM DESCRIPTION

Boundary data to be put into the data bank are traced (scribed) on to a clean

[1] KWIC — Key Word In Context document indexing and cross-referencing system based on computer sorting of key words in the title. Ref. IBM Publ. E20-8091.

[2] Byte — a unit of computer storage space made up of eight digits, or bits, in the binary system (using only 0 or 1). Each byte is capable of storing one letter, two decimal digits, or a binary value.

sheet from the source map (Fig. 1). The unique areas or 'map elements' are numbered on a transparent overlay and the corresponding classification is transcribed to a data sheet for punching into cards to be read by the computer.

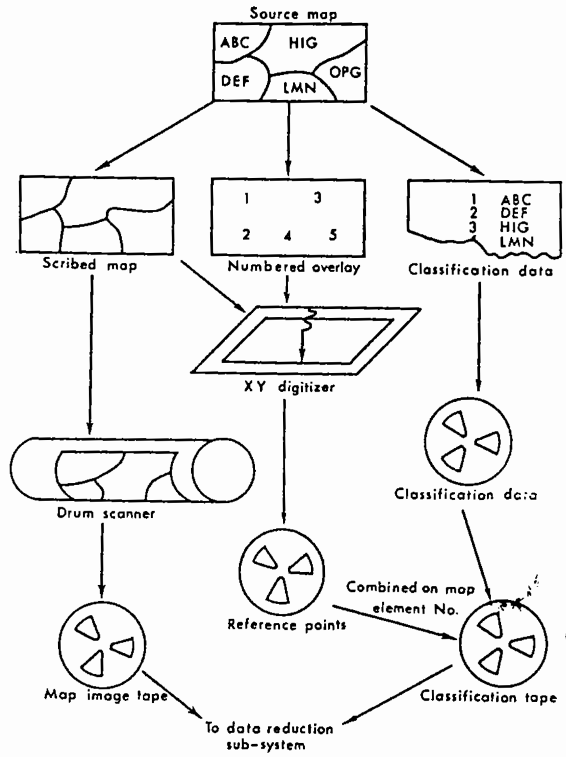

Fig. 1. Diagram showing flow of data preparation procedures.

The traced boundary sheet is placed on the drum scanner, and the scanning operation produces a digitized map of the boundaries on magnetic tape. The drum scanner was developed to meet Rural Development Branch requirements by the International Business Machines Company. The possible use of the drum scanning approach was first considered in 1963. The preliminary design criteria were established by the Rural Development Branch in 1964 and development work was contracted to the International Business Machines Company in 1965. The scanner consists of a cylindrical drum on which a map or chart can be mounted, and a movable carriage which slowly moves the scanning head across the front of the revolving drum. The scanning system consists of the scanning head proper, its associated electronics, and controls leading to a standard IBM magnetic tape drive.

The technique employed is to detect the changes in intensity of light reflected from black or white areas on the map or chart surface and to record this information as a series of binary bits written on magnetic tape. The scan head is a device utilizing fibre optics and is capable of scanning eight scan lines simultaneously. The drum scanner can accept a map up to 48 in. × 48 in. in size. A full-size map takes approximately 15 minutes to scan, including the time for mounting and dismounting it. Smaller sheets take a correspondingly shorter time.

It is not within the scope of this paper to give a detailed description of the drum scanner, though it is hoped that the engineering aspects will be covered in detail in a future paper. The format of the map-image data on tape is, however, pertinent to the discussion. One map-image record is produced for each 0·032 in. along the X-axis of a map sheet, and the height of each record area is 0·004 in. along the Y-axis. The 0·032 in. record, comprising one byte of computer storage, is divided into eight bits. Each bit thus represents an area or spot 0·004 in. wide. Lines drawn on the map are usually 0·008 in. wide. If the scan heads on the scanner identify 50% or more of a spot as part of a line, then a '1' bit is generated; otherwise a '0' bit is generated. A line in this manner is represented as a collection of bits which are usually either one, two, or three spots in width.

The traced boundary sheet with the transparent numbered overlay is placed on a D-Mac cartographic X-Y digitizer where the four reference corner points and the co-ordinates of one reference point per 'map face' are coded in digits. A map face is any one of the distinct areas that together make up the surface of the map. As noted before, information related to a face is considered to be homogeneously distributed within that face. The output from the X-Y digitizer is produced on magnetic tape by means of an NCR encoder; this will revert to punched cards if it is found that the error-edit capability of cards is needed. The classification data sheet is now also directly transcribed on to magnetic tape, though this may be taken back to punched card output. Classification data and the digitized reference points are combined on the basis of map face number to result in a classification tape.

Entering Data into the System

The basic approach to feeding map data into the system is to reconstruct a line segment, or the part of a line that lies between adjacent vertices, from the point comprising the scanned map image. These segments are then combined with the classification information to produce map faces which are a basic unit of storage.

The following are some of the steps in this input procedure. As a preliminary, the identification of the scanner and classification tapes, coverage and map identification, and similar data are put into the procedure which controls the flow in the subsequent update[1] operation. The classification tape is edited for data consistency and is changed into system format during this stage (Fig. 2).

The map-image tape then enters the main map-data reduction procedure. Since a 30-in. by 30-in. map generates over 56 million bits, occupying over 7 million bytes of computer storage on an IBM System 360, the data reduction of the map image is performed sequentially on smaller units known as 'sections'. The use of a square (or nearly square) section results in considerably longer lines being available from the map for processing at one time than would be the case if a long, thin rectangle were used. A computer with 512 thousand bytes of core storage can handle a section in the order of 1½ in. × 2½ in.

Each spot in the cloud of spots which make up the lines is assigned a 'V' value. This is a measure of the number of information-carrying spots surrounding it.

---

[1] Update — A computer procedure to combine new data being entered into the system with data already existing in the system. This may take the form of correcting, replacing, or deleting existing data, or inserting or adding new data.

This minimizes the effect of irrelevant bits and tends to pick out the center points along the line. The search follows the highest V values; it eliminates the redundant spots in the cloud.

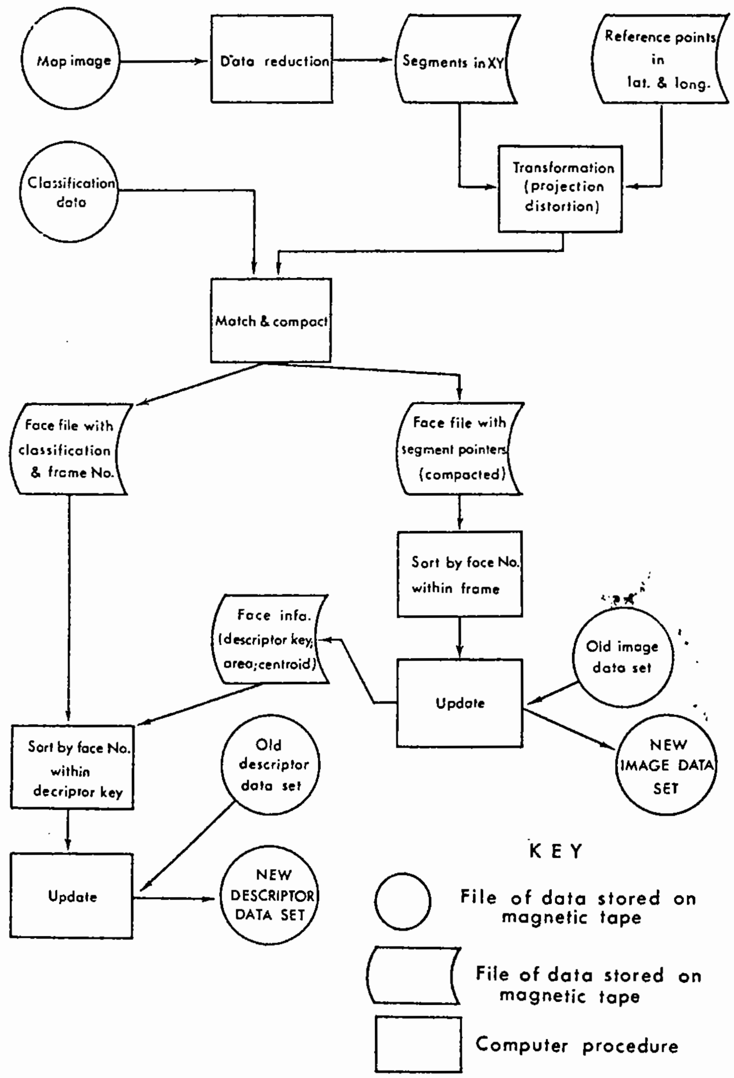

Fig. 2. Diagram showing sequence of file update operations.

The center points are coded to identify line intersections (or vertices) and the sense of direction of the line. Having thus located the points which comprise boundary lines, it is a simple task to record the X and Y co-ordinates of each point along a segment.

The system requires descriptive information to be related to map elements. One method of accomplishing this is to apply an identifying tag to both sides of the line. This tag also indicates in which direction the line was first followed, this being necessary if the sides of the line are to have a constant meaning.

The identifying tags are called 'system colors'. They are analogous to the colors in a political map. A sort-and-search of these colors enables the segments to be connected with each other, and hence faces to be assembled.

Using the reference points in latitude and longitude taken from the four corner points of the map, a transformation is carried out which locates the X-Y digitizer map-element reference points within the scanner. Map projections, which can vary from one map to another, are normalized. Calculations are made to correct for linear distortion and skewed orientation of the map on the scanner or digitizer. The transformed 'map-image data set' and the classification (or 'descriptor-data set') then are matched and compacted. During this match-and-compact operation, the map-image co-ordinates are recorded in terms of a standardized geodetic co-ordinate system. This allows a uniform base for storage and the subsequent measurement and overlay procedures.

The choice of a standard co-ordinate system was a major consideration. The eventual measurement needs (i.e. area, length, and centroid) required the chosen system to be locally cartesian. However, a co-ordinate system based on a projection can result in a system of regions, each with its own co-ordinate system. This problem is particularly pertinent when one considers an area as extensive as Canada.

Careful investigation indicated that a system comprised of the geodetic latitude and longitude had many advantages. The smallest division in the geodetic co-ordinate system used in the data bank is called a unit grid. It represents an angular displacement of $1/2^{24}$ degrees. This was derived quite empirically. Using a 4-byte unit, 1 byte allows a span of 128 degrees which is sufficient to encompass Canada. The remaining 3 bytes represent the possible subdivision of any one degree.

The theoretical resolution of the system is determined by the actual distance on the ground covered by this unit grid, which at 45 degrees latitude is just over 1/4 in. in the latitudinal (or X) direction. This is considered adequate for the data being put into the system.

Scale within the system is in terms of the unit grid distance. Factors from 2^0 to 2^{31} have been devised to provide coarser resolution.

To handle map information within the system, it is convenient to subdivide the co-ordinate system into regions called 'frames'. A frame has an equal angular displacement in the X and Y directions, hence is a square in the geodetic co-ordinate system.

A relatively simple calculation reveals that a map of average density (30 in. by 30 in., with 800 in. of boundary lines), will occupy 200,000 bytes of storage if no scale change or transformation is performed. With up to 30,000 maps envisaged as the primary content of the data bank, a compact notation for storage of co-ordinates was essential.

With a code based on direction change between co-ordinates and distance between co-ordinates, a sequence of simple codes can be used to describe co-ordinates. A sample line, requiring 864 bits for normal X-Y recording, occupies 76 bits in compact notation. If required, lines with regular patterns can be further compacted by storing the pattern with an indication of how many times the pattern is repeated.

In the match-and-compact phase, routines are carried out to calculate the area of each face, the centroid of face elements, and the length of line elements. In the same phase, an extensive error analysis is performed to ensure that the map is topologically correct. Errors found at this stage are documented by a series of error messages on the computer printer.

The match-and-compact operation produces two index files. The first of these is a face file with classification and frame number which, when sorted, is used in updating the descriptor-data set. The second is a face file with segment identifiers which is used to update the image-data set. Incorporated in the second file is the basic compact notation of co-ordinate data by frame number. The routine for updating the image data set provides the geodetic properties (area, centroid, and length) as required to update the descriptor-data set. Both of these update routines can produce error listings as new data are matched with data already in the data bank. Again, error correcting is carried out as an update to the primary map-data reduction phase.

The best approach to take with regard to error correction will only be found by trial with a working system. Given a high percentage of errors requiring reference back to source documents or even to field survey, the relatively expensive method using cathode-ray tube displays would add little, if anything, to the efficiency of the error-correction procedure. On the other hand, given a high percentage of errors of a strictly cartographic nature and not requiring reference to source documents, the cathode-ray tube approach, by which images displayed on the tube can be corrected by 'drawing' on it with a beam of light, would have considerable merit. Both approaches will be investigated during the system trials.

Data Bank Organization

The data bank is divided into classification data contained within the descriptor-data set and boundary data contained within the image-data set. Three levels of file organization are envisaged. These are: (1) consecutive, (2) regional, and (3) indexed. These file organizations, together with an unstructured or structured version of the classification data within the descriptor-data set, have been combined into six levels. Five of these will be possible within the present scope of the data bank.

Using the descriptor-data set as an example, the relationship between the various levels can be thought of as follows: Level 1 represents the basic descriptor-data set arranged by consecutive face number; Level 2 represents a sorted Level 1, grouped according to some selected characteristic or set of characteristics; Level 3 is the equivalent to Level 1 for a specific region or group of regions; Level 4 can be thought of as a Level 3 which has been structured by grouping the faces relating to a certain characteristic or set of characteristics; Level 6 is a Level 2 or 4 which is not only structured but has an index of its contents available to facilitate further search. Level 5 is not implemented as an indexed consecutive file is not an advantage.

In the descriptor-data set for each map element, there is a list of pointers to the frames containing relevant parts of the boundary information for that map element. The format of this key varies with the level of file organization, but in all cases, it serves to relate the image-data set to the descriptor-data set. The record formats of the various levels of descriptor-data set are illustrated below.

Level 1, 3

Record Type	Cover-age Number	Map Element	Geodetic Data	Factor Data	Frame List	Level 3 Region List

Level 2, 4, 6

Record Type	DDS Key Classification Data	Geodetic Data	Non-Key Factor Data	Frame List	Region No. List

Data Retrieval

As the boundary information is kept separate from the description information it is only necessary to use the boundary information if actual boundaries have to be compared or output. Otherwise, all retrieval can be done from the description information files. This leads to extremely efficient use of the data bank, as most requests will not require use of the boundaries.

A computer needs a detailed description of the location and organization of data within itself before it can bring it out or manipulate it. These detailed descriptions are themselves kept in computer storage and are indexed by key words. These key words have been made to be the normal words that would ordinarily describe the maps such as PRESENT LAND USE or AGRICULTURAL CAPABILITY. The use of such key words automatically generates computer programs that both describe the data and actually bring it out of the computer.

In the same way key words are used to describe the types of manipulation that can be carried out by the system. When data is to be retrieved the request is written by combining the key names of the data and the key words of the desired analysis. This results in a very powerful set of instructions being available that are also very flexible. This flexibility in data specification statements is made possible by use of the PL/1 language. Uncomplicated requests will be extremely simple to address to the computer. The more complex requests will necessitate a small program being written, but even this will be facilitated by the use of these key words which represent already written small programs.

Overlay Procedure

The overlay procedure of the system is the well-known function of putting one map over another and examining the resulting data relationships.

Firstly, the two maps in the data bank are brought to the same scale. Then a section of one map of a size that can be handled by the computer is brought into core and the corresponding section of the map being overlaid is similarly brought into core and superimposed on the first. This, in effect, creates a new map with new faces. The new faces are 're-colored' and identified as new homogeneous areas. The first description data set is then brought in and the proper description is applied to each of the new faces. The description data set from the overlay map is similarly brought in and applied to the new map faces. Each of the new faces has now got a double name, one from each of the original two maps. The process is thus one of creating one 'new' map from the two original maps being overlaid. The new map can then have its areas measured and summarized in the same way as any other map in the system. It is stored and kept in the system as if it were an original map coverage. Up to eight maps can be overlaid in the same operation but obviously this is not a limitation, as the results of two overlays can subsequently be themselves overlaid.

Data Control

Data control within the system is achieved by the system monitor. The system monitor accepts pertinent data on the history of map-data manipulation within the system at all times. Many of the responsibilities for system control in such an open-ended system must rest with the system librarian.

The librarian's responsibilities include deciding whether coverages are permanent or temporary, selecting the resolution at which boundary lines for various coverages need to be stored, and deciding the way in which the descriptor-data sets are filed for ease of retrieval and comparison. He is also responsible for providing the procedures which edit the classification data in the preliminary phase of the map-data reduction sub-system. He must tailor the key words that describe the different types of map and different types of manipulation to efficient, specifically applicable retrieval requirements. He is in control of the flow of individual maps within the system and, similarly he must evaluate the practicability of assessment requests, including the avoidance of duplicate assessments.

CONCLUSION

The Geographic Information System of the Rural Development Branch is still in an early stage of its development. Not all the procedures described have yet been fully implemented and at present rates of progress it will be several years before the data bank contains maps for any one type of information that cover the whole of the settled portion of Canada. The effectiveness of the system will of course depend as much on the quality of the data entered into the bank as on the capabilities for handling data. Nevertheless, the system is further advanced than any other major land data bank and contains several new concepts and techniques, especially those relating to the compact storage of boundary data and the rapid comparison of one map with another. Such a system is essential to effective rural planning in any country and offers for the first time the possibility of rapid and efficient geographical analysis which has application in any nation where the developing economy is concerned with the natural resources.

The author gratefully acknowledges the support and encouragement given this work by ARDA, particularly from A. T. Davidson, L. E. Pratt, and W. A. Benson, Chief, Canada Land Inventory, and from the International Business Machines Company of the IBM/Ottawa staff associated with the programming and computer applications of the project; specifically: Guy Morton, Frank Jankulak, Don Lever, Art Benjamin, Robert Kemeny, Peter Kingston, and Bruce Ferier.