ACCESSING LARGE SPATIAL DATA BASES VIA MICROCOMPUTER

Christopher G. Heivly
Office of The Geographer
Department of State
Washington, D.C. 20520

Timothy White
Social & Behavioral Sciences Lab
University of South Carolina
Columbia, S.C. 29208

ABSTRACT

The Office of The Geographer within the Department of State is currently changing its production mapping unit from a manually intensive system to a computer intensive system. At the core of the computer mapping system is a number of microcomputer workstations. The workstations consist of IBM-AT's with high resolution graphic boards and monitors, digitizers and AUTOCAD software.

This paper discusses the utilization of the World Data Bank II data within this microcomputer framework. This data is normally used in large mainframe graphic systems. However, the data has been reformatted and reconfigures for fast, efficient interactive use on the micro-workstations.

INTRODUCTION

The Office of The Geographer, Department of State is responsible for producing production quality maps and graphs for various bureaus within the State Department. These maps and graphs are produced for every day publications and some long term publications as well. The turnaround time for producing maps and other graphics on a timely basis is a problem that has plagued the office for some time. To rectify the time spent drafting maps in a manual mode, the office has implemented a PC based system designed to create maps for any part of the world.

Data

The move towards automation of map and graphic production began in February of 1986. The first priority of the office was to build a system that could access digital data of the entire world, place the data within a graphic framework for additional manipulation and annotation, and to have a mechanism to output the final maps in various formats. Data requirements were simple:

1. Coverage of the entire world for coastlines and international boundaries,
2. Data in latitude and longitude coordinates.

The primary use of the digital data is to provide a base map to which other important information can be added during the annotation process. Based on these simple requirements the CIA's World Data Bank files were selected. These files represent the entire world in a latitude/longitude format and representation for four areas of the world (North America, South America/Antarctica, Europe/Africa, and Asia). World Data Bank I includes:

1. Coastlines, Islands and Lakes, and
2. International boundaries.

The average digitized map scale was 1:12,000,000. The files contain approximately 110,000 points representing the two overlays. World Data Bank II includes:

1. Coastlines, Islands and Lakes,
2. International Boundaries,
3. Internal Administrative Boundaries,
4. Rivers,
5. Roads, and
6. Railroads.

The average digitized map was 1:3,000,000. The files contain approximately 10,000,000 points representing the six overlays (Coverage is not complete for features such as roads, railroads, and internal boundaries).

Hardware

The hardware for the map production system consists of two workstations. These configurations are:

WORKSTATION #1

1 - IBM-AT; 640K, (2) 20 meg hard disks, 1.2 meg floppy disk, 80287 math co-precessor.
1 - ARTIST 1+ graphics card; 1024x1024 resolution, 16 colors.
1 - GIGATEK 19" color monitor.
1 - CALCOMP 1043GT plotter; A-E size plots, 8 pen.
1 - GTCO digitizer; 36X48", 16 button.
1 - OKIDATA Microline 193 printer.
1 - ALLOY 9-Track tape drive.
1 - Bernoulli Box; (2) 20 meg removable cartridges.

WORKSTATION #2

1 - IBM-AT; 640K, 20 meg hard disk, 1.2 meg floppy disk, 80287 math co-processor.
1 - TECMAR graphics card; 640x350 resolution, 16 colors.
1 - BITEC 13" color monitor.
1 - Hewlett Packard plotter; A-B size, 6 pen.

Software

The software utilized on the production mapping workstation is AUTOCAD from AutoDesk, Inc. AUTOCAD provides a vector based graphics toolbox containing various functions to manipulate graphic features including a wide range of annotation capabilities. AUTOCAD also contains device drivers for a multitude of graphics boards, monitors, digitizers, plotters and printers.

MAP PRODUCTION SYSTEM

The map production system as designed contains four basic sections; (1)Data conversion, (2) Data query software including download capability from the 9-Track tape drive, (3) Projection and AUTOCAD conversion, and (4) interactive computer aided map design.

Data Conversion

World Data Bank files represent features of the world in a simple vector format. The direct access format for mainframe system using CAM/GS-CAM contains two types of files. The first is an index file containing line id number, feature rank code, number of points for the line feature, coordinate data file. Each index record is binary, unformatted and 32 bytes long. The second file type is a coordinate file containing 50 pairs of llatitude/longitude points in radians. Direct access records are binary, unformatted and 400 bytes long (50 coordinates per record * 2 words per coordinate * 4 bytes per word).

This original format is a simple but inefficient storage format which performs adequately for mainframe graphics systems. Access speed, and storage are not serious issues for a mainframe perspective. However, in a microcomputer environment this data structure does not optimally utilize existing disk space. Specifically, when the number of points in a linear feature is not an even multiple of 50, space will be wasted in the file. For example, given two features, the first with 163 points, the second containing 52 points, the following coordinate assignment would occur;

RECORD 1	---	coordinates	1- 50	for feature #1.
RECORD 2	---	coordinates	51-100	for feature #1.
RECORD 3	---	coordinates	101-150	for feature #1.
RECORD 4	---	coordinates	151-163	for feature #1.
RECORD 5	---	coordinates	1- 50	for feature #2.
RECORD 6	---	coordinates	51- 52	for feature #2.

In this example, only 26% of record 4 and 4% of record 6 is used. In other words, a total of 680 bytes of potential coordinate space is wasted.

In a microcomputer environment, the limitations of internal memory (RAM) and external storage (Disk space) made the processing of the original WDBII format very time consuming. Thus, an alternative data format was needed to alleviate the storage and access speed problems.

There were two basic operations performed on the WDBII files to enhance access speed and disk storage. The first operation attempted to compress the overall size of the files without eliminating any points. A revised direct access format was created by processing the sequential data structure one feature at a time. For each line output is generated for two data files; an index file, and a coordinate file. This is similar to the original CAM/GS-CAM direct access format. The new index record contains alternative information about each line feature. This information includes feature rank code, number of points representing the line, an index value indicating starting record position, an offset indicating the relative position within the record, and two coordinate pairs defining the feature window. Eliminated from the original format were the line id number. Each index record is binary and 32 bytes long. Additional disk access speed was gained by changing the block size of the coordinate and index files. Disk space on the IBM PC/XT/AT is allocated in clusters. A cluster is a group of disk sectors which varies in size from one to four kilobytes depending on the recording media (i.e. floppy disk, XT hard disk, or AT hard disk) and version of Disk Operating System (DOS). By formatting the WDBII data into records which correspond to cluster size results in the fastest and most efficient disk access possible on the AT. For the AT hard disk and DOS version 3.1, the cluster size is 2048 bytes. A new record size of 2048 bytes is created with a blocking factor of 256 coordinate pairs per record (1 coordinate pair/8 bytes per pair * 2048 bytes) for the coordinate file. Index records are grouped together in units of 64 to form direct access records. By changing the structure of both files to access cluster size records, the total number of records has been reduced and dramatic increases in access speed have been gained. For example, given a file with three features, the first with 514 points, the second with 102 points, and the third with 263 points, the following coordinate assignments would be made;

RECORD 1 ---	coordinates	1-256	for feature #1.
RECORD 2 ---	coordinates	257-512	for feature #1.
RECORD 3 ---	coordinates	513-514	for feature #1 and
	coordinates	1-102	for feature #2 and
	coordinates	1-152	for feature #3.
RECORD 4 ---	coordinates	153-263	for feature #3.

For the Office of The Geographer, this revised format is the most efficient and compact format for WDBII data files.

Quick Index Files

A second major need was to further decrease the time required to extract specific features for different geographic areas. The solution to the problem was the development of "quick" index files based on geographic area and feature rank code. The geographic quick index file consists of a 2 byte variable (16 bits) representing 16 areas of the world. Each area consists of a geographic window defined by 45 degrees of longitude and 90 degrees of latitude. The 16 areas correspond to the 16 bits in the 2 byte variable. The configurations are as follows:

AREA #1	180W/N	-	135W/N	-----	BIT 1
AREA #2	134W/N	-	90W/N	-----	BIT 2
AREA #3	89W/N	-	45W/N	-----	BIT 3
AREA #4	44W/N	-	0W/N	-----	BIT 4
AREA #5	1E/N	-	45E/N	-----	BIT 5
AREA #6	46E/N	-	90E/N	-----	BIT 6
AREA #7	91E/N	-	135E/N	-----	BIT 7
AREA #8	136E/N	-	180E/N	-----	BIT 8
AREA #9	180W/S	-	135W/S	-----	BIT 9
AREA #10	134W/S	-	90W/S	-----	BIT 10
AREA #11	89W/S	-	45W/S	-----	BIT 11
AREA #12	44W/S	-	0W/S	-----	BIT 12
AREA #13	1E/S	-	45E/S	-----	BIT 13
AREA #14	46E/S	-	90E/S	-----	BIT 14
AREA #15	91E/S	-	135E/S	-----	BIT 15
AREA #16	136E/S	-	180E/S	-----	BIT 16.

As each line was processed, corresponding bits were turned on based on the extremes of the latitude and longitude coordinates. This 2 byte variable can be quickly scanned and a determination made whether the line falls within the geographic window selected by the cartographer. The quick index file was also formatted to meet the cluster size of the AT. Thus, 1024 two byte variables reside on a record.

A similar version was made for the feature rank code. This code corresponds to a detailed ranking for each overlay. For example, in the railroad file there are subcategories for broad gauge, standard gauge, and narrow gauge railroads. The rank index file contains a one byte variable that corresponds to the individual feature rank codes. This variable can be scanned to determine whether the code matches the code(s) selected by the cartographer. Both the geographic and the feature rank index files were created to speed the data access process. By eliminating lines not falling within the predescribed window, the amount of data to be transferred to the PC is greatly reduced.

Data Query System

The data query system is the data base accessing module. There are three basic data queries that the cartographer can perform; (1) geographic window, (2) scale data base, and (3) feature/rank extraction.

The "geographic window" parameters contain the lat/long coordinates of the window selected by the cartographer. This allows the query system to eliminate data from the data base that is not needed for the specific application. The "scale data base" refers to either WDBI or WDBII. Depending on the application/scale of the final map product, one of the two data bases is chosen. For example, a map of Africa generated at 8.5"x11" does not require the detail of data within WDBII. Likewise, a map of El Salvador at 30"x40" does require the greatest detail possible from WDBII.

Projection & Conversion

After the appropriate data has been extracted they must be projected into a cartesian coordinate system and converted into the AUTOCAD data interchange format. In order to accomplish this the USGS General Cartographic Transformation Package (GCTP) was transferred into the PC-AT environment and linked to the data query subsystem. GCTP provides excellent forward and reverse projection transformations between 20 different map projections. The resultant XY coordinate strings are then written out as AUTOCAD polylines. These polylines are easily imported into AUTOCAD. AUTOCAD provides a full range of computer aided design functions that facilitate the final map production process.

CONCLUSION

The reformatting and reconfiguration of WDBII data files has provided the Office of The Geographer with a faster, more efficient data format from which microcomputer workstations can access "mainframe" type data. The ability for fast search and retrieval methods to access this data has greatly reduced the turnaround time for producing production quality maps. The fast access of digital data combined with the ability to transform the data into 20 different projections and work in an interactive CAD environment enables the cartographer to quickly and more accurately display geographic information.

Future developments include the polygonization of the WDBII files which involves adding topology to the data structure, accessing other world-based digital files within a similar framework, and an interactive projection driver to quickly view the results of the projection parameters to verify the geographic window and projection nuances.

REFERENCES

Software Documentation for GCTP: General Cartographic Transformation Package, U.S. Geological Survey, National Mapping Division.

GS-CAM: Geological Survey-Cartographic Automatic Mapping, U.S. Geological Survey, National Mapping Division.

CAM: Cartographic Automatic Mapping Program Documentation, Central Intelligence Agency.

AUTOCAD User Reference Manual.