UNIT 66 - DATABASE CREATION

- EXAM AND DISCUSSION QUESTIONS
- NOTES

This unit is the longest one included in the Curriculum. It will be impossible to cover all this material in one lecture, but there is no clear break at which to split this cleanly. Some of the material is technical and some of it management oriented. You will have to decide what to omit, if you need to, based on your students' interests and educational backgrounds.

UNIT 66 - DATABASE CREATION

A. INTRODUCTION

- an FRS establishes:
  - the products to be generated by the system
  - the data needed to generate the products
  - the functions which must operate on the data
- working from the outline provided by the FRS, the data base design and creation process begins
  - this unit examines the management and planning issues involved in the physical creation of the database
- note that specific implementation details will not be reviewed as these are highly dependant on the particular GIS used
- emphasis is on databases for resource management applications
  - databases for facilities management are often extensions of existing non-geographic databases
    - depend too much on specifics of systems
- the key individual involved at this stage is the Database Manager or Coordinator who is responsible for:
  - definition of database contents and its "external views"
    - see Unit 43 for a discussion of the different "views" of a database
  - maintenance and update control
  - day-to-day operation, particularly if database is distributed over a network

B. DATABASE DESIGN

- provides a comprehensive framework for the database
- allows the database to be viewed in its entirety so that interaction between elements can be evaluated
- permits identification of potential problems and design alternatives
- without a good database design, there may be

irrelevant data that will not be used
- omitted data
- no update potential
- inappropriate representation of entities
- lack of integration between various parts of the database

- unsupported applications
- major additional costs to revise the database

Stages in database design

- recall from Unit 10, that steps in database design are:

1. Conceptual

   - software and hardware independent
   - describes and defines included entities and spatial objects

2. Logical
   - software specific but hardware independent
   - determined by database management system (discussed in Unit 43)

3. Physical
   - both hardware and software specific
   - related to issues of file structure, memory size and access requirements
   - this unit focuses mainly on this last stage

C. ISSUES IN DATABASE CREATION

- what storage media to use?
  - how large is the database?
  - how much can be stored online? what access speed is required for what parts of the database?
  - how should the database be laid out on the various media?
  - what growth should be allowed for in acquiring storage devices?

- how will the database change over time?
  - will new attributes be added?
  - will the number of features stored increase?

- how should the data be partitioned - both geographically and thematically?
  - is source data partitioned?
  - will products be partitioned?

- what security is needed?
  - who should be able to redefine schema - new attributes, new objects, new object classes?
  - who should be able to edit and update?

- should the database be distributed or centralized?
  - if distributed, how will it be partitioned between hosts?

- how should the database be documented?
  - who is responsible for maintaining standards of definition? standards of format? accuracy? should documentation include access to the compiler of the data?

- how should database creation be scheduled?
  - where will the data come from?

  - who determines product priorities?
  - who is responsible for scheduling data availability?

- the following sections address some of these questions

D. KEY HARDWARE PARAMETERS

Volume

- databases for GIS applications range from a few megabytes (a small resource management project) to terabytes
  - a small raster-based project using IDRISI, 100 by 200 cells, 50 layers might require 10 Mbytes database on a PC/AT
  - a mid-sized vector-based project for a National Forest using ARC/INFO might require 300 Mbytes
  - a national, archival database might reach many hundreds of Gbytes
  - the spatial database represented by the currently accumulated imagery of Landsat is order 1013 bytes

Access speed

overhead - Storage media

- data which can be accessed in order 1 second is said to be "on-line"
  - to be on-line, data must be stored on fixed or removable disk
  - relative to other forms of permanent storage, disk costs are high, and there is an effective upper limit of order 100 Gbytes for on-line storage when using common magnetic disk technology

- "archival" data (data which is comparatively stable through time) can be stored off-line until needed
  - only extracts will be on-line for analysis at any one time
  - archival systems incur additional time to mount media on hardware
  - access time to extract subsets from archival data once mounted is order 1 minute

- archival media:
  - magnetic tape
  - removable disk
  - CD-ROM

        - no ability to edit data once written - this is acceptable for many types of geographical data
        - copies are very cheap
    - optical WORM (Write Once Read Many)
    - "video" tape

- automatic multiple storage and access systems increase capacity and decrease access time
    - magnetic tape stores can be automated, raising effective capacity to 1 Tbyte (order 10,000 tapes)
        - order 10,000 tapes is also an effective upper limit to the size of a (conventional, manual mount) tape library
    - optical WORM libraries can be automated much more easily using "jukebox" technology - automatic selection and mounting of platter
    - devices to mount cassette tapes automatically are also available

Network configuration

- should database be centralized or distributed?

- there are two answers: 1. all departments share one common database, or 2. parts of the database exist on different workstations in an integrated network
        - each department responsible for maintaining its own share of the database
        - optimizes use of expertise

- with modern technology (e.g. NFS (Network File System)) user may be unaware of actual location of data being used
    - some workstations may be "diskless", owning no part of the database

- distributed databases require careful attention to responsibilities, standards, scheduling of updates

E. DATABASE REDEFINITION

- in some applications, all files, attributes, objects can be anticipated when the database is defined
    - e.g. systems for facilities management typically do not allow redefinition of the database structure by user

- other applications, particularly those involving analysis, require ability to define new objects, attributes
    - this capability is generally important in resource management applications

- important to determine who is allowed to change the database definitions
    - database administrator only?
    - project manager only?
    - any user?

F. TILES AND LAYERS

- many spatial databases are partitioned internally
  - partitions may be defined spatially (like map sheets) or thematically or both

- the term tile is often used to refer to a geographical partition of a database, and layer to a thematic partition

Reasons for partitioning

- capacity of storage devices
  - may limit the amount of data that the system can handle as one physical unit

- update
  - easier to update one partition (e.g. map sheet) at a time

- access speed
  - may be faster if only one partition is accessed at a time

- distribution
  - easier to copy a partition than to extract part of a larger database
    - e.g. US Bureau of the Census chose to partition its TIGER files by county for distribution based on user needs
    - e.g. US Geological Survey partitions digital cartographic data by 1:100,000 map sheet

- user needs
  - users need certain combinations of geographical area and theme more commonly than others
  - illustrated by the conventional arrangement of topographic and thematic map series
    - e.g. soils information is not normally shown on standard topographic maps
  - the best source of usage patterns is conventional cartographic products because their traditions have been established through continual usage and improvement

"Seamless" databases

- despite the presence of partitioning, system designers may choose to hide partitions from the user and present a homogeneous, seamless view of the database
  - e.g. are systems available to automatically mosaic Landsat scenes, so users can work independently of normal scene boundaries

- in seamless databases, the data must be fully edgematched
  - parts of an object which span geographical partitions must be logically related
    - features which extend across tile boundaries must have identical geographic coordinates and attributes at adjacent edges
    - every object must have an ID which is unique over the whole database

- the term Map Librarian is commonly applied to systems which remove partitions from the user's view of the database

Organizing data into layers

- the source documents (maps) generally determine the initial thematic division of the data into layers
  - these initial layers need not coincide with the way the data are structured internally
    - e.g. the application may consider lakes and streams as one layer while the data structure may see them as two different objects - polygons and lines

- several distinct layers may be available from the same map sheet
  - e.g. topographic maps may provide contours, lakes and streams (hydrography), roads
  - the Database Manager may choose to store these as different thematic partitions in the database

- when deciding how to partition the data by theme, need to consider:

- data relationships
  - which types of data have relationships that need to be stored in the database
  - these will need to be on the same layer or stored in such a way that relationships between them can be quickly determined

- functional requirements
  - what sets of data tend to be used together during the creation of products
    - it may be more efficient to store these on one layer

- user requirements
  - how diverse will the users requirements be
  - more diversity may require more layers to allow flexibility

- updates
  - data which needs to be updated frequently should be isolated

- common features
  - features which are common to many coverages, such as shorelines and rivers, may be stored separately then used to create other coverages that incorporate these lines as boundaries

- internal organization of layers depends on the system chosen
  - CAD systems treat each class of object as a separate layer
  - many raster systems treat each attribute as a separate layer, although objects may have many attributes
  - some newer GIS designs avoid the concept of layers entirely, storing all classes of objects and their interrelationships together

Selecting tile configurations

- tiles may cover the same area throughout the database, or they may have variable sizes

- fixed size tiles are:
  - generally inefficient in terms of storage since some tiles will have lots of data and others very little
  - good when data volume changes through time since it is not necessary to restructure tiles with updates

- variable size tiles are:
  - efficient in term of storage
  - difficult to restructure if new data is added

- boundaries may be: overhead - Tiling Variations
  - regular
    - e.g. based on map sheet boundaries
  - free-form
    - e.g. based on political or administrative boundaries, watersheds, major features like roads or rivers

- tile sizes and boundaries can be chosen based on:
  - areal extent of common queries or products
  - scale needed in output
  - balance between getting the largest areal coverage possible and speed of processing

- practically speaking, in most databases, partitions correspond to conventional map sheet boundaries, e.g. 7.5 minute quadrangles

- products will likely be created one tile at a time
  - e.g. a forest manager wants maps of timber inventory at a scale of 1:24,000
  - the size of plots is limited by the plotter itself, and by physical constraints on handling and storage
  - it makes sense to generate timber inventory maps in 7.5 minute quadrangles
  - since data will be input from quadrangles, why not tile the entire database in quadrangles as well?
  - however, a Map Librarian will be needed when small- scale products have to be generated using many tiles at once

G. DATA CONVERSION

- the process of data input to create the database is often called data conversion
  - involves the conversion of data from various sources to the format required by the selected system

- previously have examined the different ways of inputting data and various data sources
  - consideration of these options is critical in planning for database creation
  - Unit 7 discusses several issues related to integrating data from different sources

- often there are several alternative sources and input methods available for a single dataset

Database requirements

- need to consider database requirements in terms of:
    - scale
    - accuracy
    - scheduling priorities
    - cost

- scale
    - FRS specifies the scale required for output
    - will determine the largest scale that is required for datasets
        - may not need to go to added expense and time to input data at larger scales

- accuracy
    - required accuracy will determine the quality of input necessary and the amount of data that may be created
        - e.g. coarse scanning or digitizing versus very careful and detailed digitizing
        - e.g. field data collection versus satellite image interpretation

- scheduling priorities
    - some datasets will be critical in the development of later datasets and early products
        - these may justify expensive input methods or the purchase of existing sources

- alternatives for creating the database include:
    - obtaining and converting existing digital data
    - manual or automated input from maps and field sources
    - contracting data conversion to consultants

In-house conversion

- data entry is labor intensive and time consuming
    - some GIS vendors assist in the conversion effort, and there are a number of companies which specialize in conversion
    - some agencies do their conversion in-house, but there is a reluctance to do so, in many cases, as the added personnel may not be needed once the initial conversion is complete

- advantages of in-house conversion
    - agency personnel, who are familiar with the "ground truth" and unique situations of the areas of interest, are able to supervise the conversion effort
        - this can be important for unanticipated situations in which general rules cannot be uniformly applied
    - auxiliary maps and data are available if needed for interpretation
        - if the maps are sent out for digitizing, what you send is all you get
    - in-house validity checks can be made more easily

- disadvantages of in-house conversion
  - additional equipment and personnel need to be added to the project plan
  - long-term commitment to full-time employees can be expensive

H. SCHEDULING DATABASE CREATION

- database creation is a time-consuming and expensive operation which must be phased over several years of operation
  - the total cost of database creation will likely exceed the costs of hardware and software by a factor of four or five
    - e.g. over a 5 year period, of a total $5 million cost of a typical GIS project for resource management, $4 million went to data collection and entry, only $1 million to hardware, software, administration, application development

- since the benefits of the system derive from its products
  - database creation must be scheduled so the system can produce a limited number of products as quickly as possible
  - however, benefits will not be realized at the full rate until the database creation is complete

- need to know the complexity of data on each input source document to forecast data input workload
  - e.g. numbers of points, polygons, lengths of lines, number of characters of attribute

Scheduling issues

- to generate a tile of a product, the required data layers for the correct tile must have been input

- to determine the order in which datasets must be input, must rank products based on: 1. perceived benefit 2. cost of necessary input
    - highest ranked are those with high benefit, low cost of necessary input
    - lowest ranked are low benefit, high data input cost
    - some layers may be used by several products - once input, the cost to other products is nil

- the promotional benefit of a product is highest for a single tile, decreases for subsequent tiles
  - a single tile of a product can be used to "sell" the system, draw attention to its possibilities
  - high priority needs to be given to generating a product which can "sell" the system within each department or to each type of user

- need to know the payoffs between 1. producing a single tile of a new product and 2. producing further tiles of an existing product

- determining priorities under the constraint of data input capacity is a delicate operation

for the Database Manager
- many layers of data may not exist, may have to be compiled from air photos or field observation
- the schedule for data input will have to accommodate availability of data as well as product priorities

I. EXAMPLE - FLATHEAD NATIONAL FOREST DATABASE

Background

- Flathead NF located in Northwestern Montana on west slope of Continental Divide
  - adjacent to Glacier National Park
  - headquarters in Kalispell, MT

- total area within Forest boundary is 2,628,705 acres (1,063,822 ha)
  - Forest area spread over 133 1:24,000 (7.5 minute) quadrangles

- resource management responsibilities include: timber fisheries wildlife water soils recreation minerals wilderness areas rangeland fire plus maintenance of Forest infrastructure (engineering)

- substantial investment in use of Landsat imagery for forest inventory and management, using VICAR image processing software

- FRS conducted in 1984/5, planning period extended to 1991
  - important to note that this plan considers the needs of Flathead NF in isolation
    - may not be compatible with the national needs of the Forest Service or the national policy developed under the National GIS Plan
    - may conflict with emerging concepts of service- wide Corporate Information (see Unit 71)

Examples of products

- FRS identifies 55 information products

  handout - Examples of products (2 pages)

  - extracted from a study by Tomlinson Associates, Inc.

Proposed database contents

- total of 58 input datasets required

  - total database volume estimated at 1 Gbyte

- 12 already in digital form in VICAR image processing system, running on mainframe outside Forest
  - 3 interpreted from Landsat, available in raster form
  - 9 digitized in vector form, rasterized by VICARS

- 2 are large attribute files (forest stand attributes, transportation data for roads) maintained as System/2000 database outside Forest

- all remaining datasets must be derived from non-digital maps or tables
  - map scales range from 1:24,000 to 1:250,000
  - datasets vary in complexity
  - number of map sheets varies depending on scale

Example dataset characteristics

- see page 2 of handout

Tiling

- 1:24,000 7.5 minute quadrangle dominates both input and output requirements
  - therefore, makes sense to use quadrangles as tiles in the database if it must be tiled (depends on system chosen)
  - could use aggregations of 7.5 minute quadrangles, e.g. 15 minute quadrangles

Database creation plan

- needed to:
  - assign input data to object types, layers
  - determine which relationships to store in the database
  - determine naming conventions for files, attributes

- scheduled input of data from 3,162 individual map sheets over 6 years
  - need to allow for updates as well as initial data input
    - some layers updated on a regular basis - e.g. timber harvesting
    - some irregularly - e.g. forest fires

System specific issues

- preferred arrangement is a centralized database at Forest headquarters, access from workstations across the Forest

- implementation plan is based on scheduled generation of products

  - system design provides little access to the database in query mode
  - therefore product generation can be batched and data need be online only during product generation
  - however 1 Gbyte is easily accommodated online

Schedule

- database creation schedule determines ability to generate products
  - FRS calls for generation of 4,513 individual map products and 3,871 list products in same 6 years
  - digitizing need will be heaviest at beginning of period

    - ability to produce will be highest at end

- input phasing:
    - roads, PLSS section boundaries - all input in year 1
    - lakes and streams - phased over years 2-4
    - forest stands - phased over years 2-4
    - harvest areas - input to begin in year 4

- over the 6 years of database creation there will be increasing output, diminishing input

REFERENCES

ACSM-ASPRS GIMS Committee, 1989. "Multi-purpose Geographic Database Guidelines for Local Governments," ASCM Bulletin, Number 121:42-50. Provides a general outline for the consideration of scale and content for municipal GIS databases.

Calkins, H.W., and D.F. Marble, 1987. "The transition to automated production cartography: design of the master cartographic database," The American Cartographer 14:105- 119. Stresses the need for rigorous database design and illustrates the use of the entity-relationship model for spatial databases.

Nyerges, T.L., 1989. "Schema integration analysis for the development of GIS databases," International Journal of Geographical Information Systems 3:152-83. Describes methods for analyzing the differences and similarities between two or more databases

Nyerges, T.L., and K.J. Dueker, 1988. "Geographic Information Systems in Transportation," US Department of Transportation, Washington, DC. Report describes the potential use of GIS in State Transportation offices and the types of data and functionality that would be required.

Tomlinson Associates Inc., 1985. Advanced geographic information systems workloads analysis: Flathead National Forest individual forest report. Report to US Forest Service.

EXAM AND DISCUSSION QUESTIONS

1. Discuss the advantages and disadvantages of distributed vs. centralized databases for resource management agencies.

2. Most applications of GIS for resource management to date have been developed around a dedicated host, rather than sharing space on a general-purpose mainframe. In fact the development of minicomputers in the late 1970s gave a tremendous boost to GIS because it made dedicated computers available at costs which were within the range of many agencies. Why do GIS developers prefer a dedicated host? Are these arguments compatible with the idea of a database distributed over a network?

3. Describe the functions of a Database Manager for a GIS installation.

4. Despite the importance of design in GIS databases, there is remarkably little literature on the topic at this time. Why?

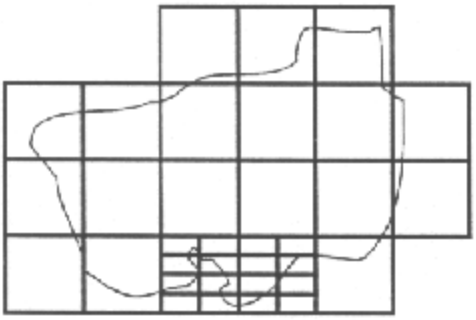
5. The Caribou product for the Forest Service calls for information on elevation and aspect. What are the different sources and data models which could be used to provide this information? What effects might each choice have on the generation of the product?

6. Suppose the Forest Service determined that its GIS would be implemented in query rather than product mode, i.e. it would train its managers to work directly with the database, instead of defining products for the system to generate. How would you go about identifying the input and output requirements of a query based system? What other items would be important in system planning?

Last Updated: August 30, 1997.

UNIT 66 IMAGES

1: Free-Form based on administrative boundaries



2: Rectangular Grid based on map sheets

3: Free-Form based on watersheds