

## Topic Series 16 GIS Data Input, Storage and Editing

Previous lectures have been concerning with getting or developing the right data for the job. Much of the baseline data available for GIS already exists. **Let the buyer beware.** Some data are not suitable for applications due to the source scale or characteristics of the documents used to produce the data.

Recall that in terms of an **information system**, prior to input of data, you should have determined what information needs to be derived (**user needs**), how it will be developed (**planning**) and how it will be collected (**data collection**) before doing data input via the **Input Subsystem**.

Remember **GIS subsystems** include – **Input, Storage/Retrieval, Manipulation and Analysis, Reporting**

Also remember, the **information system** path includes: **user needs, planning, data collection, data storage, manipulation and analysis, output products, user action**)

### General Data Sources Include:

- Maps - the primary data source
- Tabular Data - from printed reports but more frequently digital (i.e. census data)
- Reports
- Field Data - stand inventory, wildlife census, GPS
- Experts - from discussion
- Remote Sensing - imagery from satellites, aircraft
- Digital Products

Much of the **baseline digital cartographic data** that goes into a GIS can be obtained from either **private companies** or from **government agencies**. It is usually much cheaper to acquire data this way than to acquire the equivalent maps in hard-copy form and hand digitize them on your own. I've heard examples of more than **80% of the cost** of starting up a GIS as being tied to manual digitizing. As the cost of computer hardware continues to fall, the proportional cost of manual digitizing is going up.

Some examples of data from government agencies are listed below.

**(overhead)**

<u>Data Type</u>	<u>Source</u>
Topography (DEM's)	USGS, DMA
Land Use / Cover	USGS, USFWS, NBS
Boundaries	USGS
Transportation	USGS
Hydrography	USGS

Census data  
Soils  
Imagery  
DRG (digital raster graphics)

USCB  
NRCS  
NASA, USGS, NOAA  
USGS

Numerous corporations provide the above data in **value-added products** ready to load directly into a GIS. Examples are **address data** with transportation, **enhanced imagery** and **image classification products**. Much of the data mentioned above can be obtained over the **internet**. Best source for data in MS is MARIS at **www.maris.state.ms.us**.

### The Data Entry Process

We go through a series of steps to insure the correct data are selected and we accurately convert the data to the system for our use.

**Plan** - the first step is to plan **what data are needed** and how the data will be entered and used. This actually starts during the data acquisition phase. You need to know the end use of the data so that it can be entered into the database in a way to provide efficient use.

**Enter** - this is accomplished by import of existing data, digitizing or scanning. Remember trade-offs of **accuracy versus time** to enter and use.

**Edit** - Data entry and development goes through edits and often several stages of quality control before it is released for general use.

**Geo-reference** - this is often associated with the data entry and edit procedures. You must put the data in a geographic reference system that is meaningful in your organization. For example, State plane vs. UTM.

**Conversion** - if not input in the form of final use, we sometimes must convert the data (vector/raster). For example, we may digitize data in vector format but intend its primary use in a raster modeling process.

**Construct database** - once the data have been entered, edited, etc. they are loaded into the database and made available for further processing.

**Attribute entry** - this is sometimes done in the entry/editing stage but can also be done in the database through tabular data entry procedures.

In developing the database, and during the data entry phases, it is often convenient to separate data from the layers (maps) into specific themes (i.e. roads, streams, forest). Although this may require a bit more up-front data input/handling, it gives much more flexibility to the options you subsequently have in doing data analysis (more possible combinations of themes). One short cut here in use of USGS quads is to purchase map separates.

## **Manual Data Entry** (Digitizing)

A large proportion of all data entered into a GIS is at some point, digitized through **manual encoding** techniques.

Manual digitizing is done with a device called a **digitizer**. This is pad or table with an **imbedded electronic grid** that senses the position of an encoding device such as a **puck**. The software communicates with the digitizer to determine the x, y coordinates of the puck. The usual way coordinates are encoded is by pushing a button on the puck at locations on a map. This tells the software to record that location. A series of locations is recorded by alternately moving the puck and pushing the encode button. Some software packages can also request the digitizer to send a continuous **stream** of coordinates. Then the software samples the stream at regular time intervals to get coordinates of features.

The usual way we **digitize features** is by using one button to signal the beginning point of an arc (or a point features) and another button to signal intermediate points.

Manual digitizing of curves and irregular features is not an exact science by any stretch. The person doing the encoding needs to be consistent to generate digital coordinates that represent the line-work of the map with the required accuracy of the database. Sharp curves should have more intermediate vertices than very gradual ones.

The digitized data are usually **attributed by keyboard entry** although there are ways to attach or relate pre-existing databases to your data.

## **Editing Digitized Data**

The most efficient way to enter data is, in general, to **encode the line work as rapidly as possible**, then edit the data. Some software has facilities to help in the editing process by **automatic identification of problems** (double lines, dangling lines, extra nodes, unlabelled polygons, etc.). Example problems include:

- **undershoot/overshoot** - arcs that do not meet (dangling nodes)
- **label errors** - polygons without labels or with more than one in conflict or wrong label.
- **double lines** (slivers at polygon boundaries or double digitized lines)
- **general digitizing errors** - inverted line work, gaps in coverage, slivers at adjoining polygons, etc.

## **Summary of Manual Digitizing**

(overhead)

- It is **labor intensive** - that is why I say you should first look for existing data.
- It is **slow, tedious**, and therefore can cause project delays
- Lots of **inaccuracies (must build in quality checks throughout)**

These all leads to the conclusion: If I can't buy, borrow the data, maybe I should consider automated encoding equipment (if affordable).

### **Automated Digitizing Technologies**

There are several alternative for automatic encoding of map data. Generally speaking, money buys accuracy. Simple desktop scanning equipment is OK for rough encoding but is not always accurate enough to meet project requirements. The following is a list of some types of equipment:

**Hand-held scanners** - these are advertised in PC magazines and are **generally not suited to GIS**. They can be used to produce an image of a document or manuscript that is cross-referenced to more accurate data. Limited in size of area you can effectively scan.

**Desktop scanners** - better than hand-held and getting better all the time. These come in regular page size and sometimes 11x17". These are useful for scanning small maps but are **not accurate enough for detailed mapping**. They use a lens system and operate in a fashion similar to a copy machine except that the output is sent to a digital file rather than to a page copy.

**Drum scanner** - best and also most expensive. These are used in large firms and government agencies. The map is fixed on a drum that rotates as a scan head systematically encodes the data.

**Video scanner** - these utilize a high-resolution video system to encode small pieces of the map manuscript incrementally. These have become an alternative to drum scanners and are generally cheaper but offer fairly good accuracy.

### **Problems with Automatic Digitizing**

Automatic digitizers must be monitored to get them to perform when they get to areas that produce conflicts in the vectorization process. Examples of problems are:

**Line Breaks** - maps developed for hardcopy cartographic presentation often have line breaks due to labels (highways and contour labels) or dangling contours.

**Annotation** - "o's" look like closed polygons. Other letters look like line work to a scanner.

**Direction** - the laser line follower must be told which direction to go when it gets to a point where line work branches in two or more directions from one location.

**Contrast** - maps may not have good enough contrast for a scanner to detect differences in polygon shades or linework. This must also be guided by an operator.

In summary, there are numerous things that must be taken into account with automatic encoding equipment and techniques:

- **Fast and accurate** - it can be fast and accurate (less bias than multiple human operators)
- Requires **human guidance**
- Often **expensive**
- **Mostly monochrome** (B&W) based so separations must be used.

### **Remote Sensing for GIS Data**

This is defined as the characterization of an object or phenomenon without physical contact. Imagery from aircraft (aerial photography, video, frame cameras) and satellites is often used both to **develop data for GIS** (by interpretation) or as an **image backdrop** to existing GIS themes.

The **EMS** (electromagnetic spectrum) is the basis for a large part of remote sensing. Our main area of interest is in the **visible and near-infrared wavelengths**. The majority of sensing devices utilize energy in this region to record information on our targets of interest.

#### **Example sensors are:**

- film camera (overhead)
- digital frame camera (overhead)
- multi-spectral scanner (overheads: TM and SPOT, AVHRR)

#### **Other sensors of note include:**

- radar (overhead)
- lidar (overhead)

**High-resolution satellite data** (1m pan, 4m MSS) will revolutionize GIS in that this will be available on a regular basis without users having to contract aerial photography missions. These can be obtained on the internet. (overhead Ikonos)

### **Image Analysis (getting data ready for GIS)**

In order to get information from the remote sensing imagery into a form that can be used in GIS, we **either interpret it manually and digitize** the line work, or we use automated interpretation procedures call **image processing** to develop **thematic maps** from the digital data.

### **Image processing involves multiple steps including:**

- **Preprocessing** - scan photography or import digital imagery, remove errors, rectify (includes **ortho-rectification** to remove relief distortions)
- **Enhancement** - for **initial exploration** of information content
- **Thematic analysis** (optional) - this is to convert data as an **image backdrop** or theme.
- **Classification** - **categorize the data** as specific land covers, vegetation types, etc.
- **Integration** - convert to GIS format, possible **vectorization** of classified raster product.

The **vectorization** of image data products can be inefficient in terms of time and computer storage constraints. Some systems can now handle vector and raster simultaneously so the conversion process is not needed. When we do modeling, we often convert our vector coverages to raster because **modeling in the raster domain** is usually more efficient.

### **GPS for Data Entry**

GPS can be used to collect **basic location data** for features. This can be done by:

- 1) using an inexpensive GPS unit and hand recording locations
- 2) using a GPS that can **log locations at intervals** and **download the log, or**
- 3) using combined data recorder with GPS to log location and also store attributes (in some cases, this involves a **notebook PC** as the recording device.).

The latter two types of units are becoming commonplace and can be used to efficiently enter data such as **new roads, stand boundaries, forest inventory plots and data.**

### **Applications in Natural Resources**

**(overhead)**

Establishment of stand or field boundaries (input to GIS)

Mapping location of new roads (GIS)

Site locations (RCW colonies, superior trees/stands, inventory plot location, soil sample locations)

Photo control (aircraft position in flight, control points for photo adjustments)

## Geo-referencing / Projections Revisited

Otherwise called registration. This is the process in GIS of registering input data (maps, images) coordinates (in inches or mm) to real world coordinates. **Geo-referencing and making changes to projections is a key stumbling block to getting data sets compatible.** In the data entry process, there are three ways to accomplish this (**overhead**):

1. **registration points** (tics in Arc/Info) with known real-world coordinates are located on the map sheet and **entered into the database in digitizer inches** along with the rest of the features. **The data are then transformed** (projected) into the real-world system by telling the computer to transform the tics to real-world. In doing so, the computer can also determine the translation of all other feature coordinates. (**post digitizer translation**).
2. the second way to do this is to **enter the tics as real-world coordinates** in the database prior to digitizing. In setting up to digitize, the operator tells the computer where the tics are on the map manuscript as it is attached to the digitizer. Then, as the operator enters new feature coordinates, the computer automatically translates from digitizer coordinates to the real world coordinates before storing the data being digitizer. (**coordinate conversion on the fly**)
3. Last way to to use a **pre-registered background data source** as the reference and digitize or encode features unto it directly on the computer screen. This is called **heads-up digitizing** and we will look at this in lab. (**overhead; stands on photo**)

The operator can change to a new coordinate system by telling the computer how to translate between projections by programmed algorithms that can take one known projection and change it to another.

**Projection:** recall that a projection is the mathematical representation of earth features to a system that represents them in two dimensions (2D or **Cartesian**).

**NOTE: Latitude and longitude is not a projection!!!!**

In **vector systems**, this involves translating the x, y coordinates of points, vertices, to new coordinates by solution of simultaneous equations. (Least squares regression).

In **raster systems**, it involves coordinate translation and re-sampling of the input grid to the output grid.

**Proper geo-referencing and projection conversions** are critical to data sets matching and is a large source of error in database development.

## File Conversion / Database Construction

After initial editing the resulting vector or raster files are attributed then entered into the database. Steps to do this are:

- a. convert data to topological structure
- b. design and attribute table
- c. add attributes to features
- d. merge data into the database

## Outputs for Edit Inspection

We have already talked about these but they are also data entry information (some outputs become new data in the database). Some outputs, particularly draft outputs, are used in the editing process to help the operator compare the features in the database to the original map manuscript. The most frequent way we subjectively evaluate data correctness is by bringing up multiple data sets on the computer screen and inspecting for correct coding and registration.

## Coverage Modification

Once the initial database is built or even during its construction, we often break up large coverages into more manageable pieces, either from a geographic stand point or from a computer storage/management standpoint. Coverages can be subdivided by **existing boundaries** or on the basis of a **geographic grid**.

**Tiling** is a GIS operation that involves organization of GIS data by logical **subunits**. This can be done by splitting the coverage but is more frequently planned into the original development of the database. Common approaches to tiling include: subdivision of the area by **map sheets** (7.5 minute quads), or by **political boundaries**. The GIS then maintains a library of all the tiles that represent the database. Query operations in these more complex systems can be designed to first look at the tile library and determine which tiles are relevant to the operation. Complex tile structures can be several layers deep (**hierarchical**) to provide a mechanism to more efficiently select the exact area of interest in database operations. An example would be to have the smallest tiles as 7.5 minute quads then aggregate them into 15-minute areas, then 30-minute, 1x1 minute etc. This is done at the expense of disc space.

## Other Editing Operations

Long after the database is built, there are frequent updates that involve modifications of the database. GIS provides the means to update (**add or delete**) information based on both spatial (graphical) or attribute manipulations. Some of these operations are interactive on the computer screen but they can be based on overlay of more than one coverage (**top of 202**).

We can also use the database to select features based on their attributes and then either edit or

delete them. These operations (**clumping/sieving**) may be performed to **reduce the complexity** of the data due to management considerations or to **simplify the data for plotting**.

### Metadata

(overheads)

Something that is very important in initial data entry is to **maintain good records** of the **source of data, how it was entered**, at what **precision, who** entered it, etc. All of this history of the processing record of the data is called **metadata** and is generally attached or closely associated with the database. (overheads: **Metadata is, What is Metadata?, Why Metadata?**)

### **Data Issues / Problems**

The final formal topic in our GIS implementation discussions is one that has been repeated several times in the course of this semester. The assumption here is that you are now implementing GIS somewhere, have gotten the infrastructure and are about to propagate data into the database. The **accuracy and consistency** of data can make or break a GIS. Problems arise with the **type of data, the resolution and scale, compatibilities, availability and cost**.

### Raster Data

Raster data can do a good job of representing the earth's surface given sufficient resolution is used. But, as we have seen, it can sometimes be problematic by poorly representing features with **linear components** or **discrete boundaries**.

**Gridding** - High resolution grids are often sufficient to represent the linear features. It is possible, however, to generate problems by either starting with a low resolution grid, or by going from high to low, then back to high resolution (in the latter case, **information is lost**). The problems are not just cosmetic, **area representation** is adversely affected when the analyst changes grid sizes.

**Accuracy issues** - Raster data used to represent features generally do so with uncertainty as to the exact location. **Point features** can be represented as raster, but in theory, can be inaccurate by as much as the distance of the **diagonal of the raster** used to represent the location. Line features **direction** and **true endpoints** are also misrepresented in raster structures as indicated on the right. Distance in raster on lower end systems may simply be in terms of pixel units which can be inaccurate if measured on the diagonal. High-end systems usually solve this by knowing the size of the cells and convert start and stop x, y to distance (simple Pythagorean calc.).

### Second-generation data

Here we have the problem of propagation of errors based on the number of times data are copied handled, reformatted, etc. This could be likened to the rumor. At every step in data reprocessing, there can be slight inconsistencies in the translation steps.

## Scale

You might recall that we discussed scale in terms of suitability for projects. Very small scale maps can be enlarged but the accuracy is still unsuitable for analysis of an area. An example might be to utilize a statewide soils map to make decisions for management of an individual stand. This is inappropriate because the level of detail in the soils data is not suitable for this level of management. The dominant soils for the polygon in the statewide map may not even be present in a 20-acre stand.

Note however, that for some media, scale can be deceptive. A coverage crunched down to fit the resolution of a computer monitor may seem less accurate than it really is. Zooming in can give the user a better perspective on the actual detail of the data.

An enlargement from a coverage based on small scale can be a disappointment while the original map scale may have been much better for the application. A trade off with scale (detail) is **disc space**. Coverages based on large-scale maps have more detail and thus, have many more vertices (or pixels) and consume a lot more disc space.

## Edge Matching

This can be fixed through combined manual and automated processes. The question is always, **which map is more correct** and if both had error, how do you balance it out. There may be **missing linework** on one side or **mislabeled** or misattributed **polygons and arcs**.

## Area and Scale Coverage

One major difficulty you will likely face is getting complete data for your entire study area. The example illustrates problems for a study of Australia. Closer to home, you may likely find that, for example, soils data (or even published surveys) do not exist for many rural areas in the region where you work. Therefore, making management decisions based on GIS alone could be hampered for lack of data.

Rule: the smallest scale of data dictates the accuracy of the project. "A chain is only as strong as its weakest link."

## Other Data Problems

### **Age of information**

**Credibility (too much)** as in oversimplified earthquake zones based on simple proximity and not taking into account different terrain, soils, etc.

**Attributes** - data you acquire may have problems due to keypunch error.

**Line work** - may be **missing** or in the **wrong place**. A lot of government data have been cleaned up by value-added vendors. Good example is roads and interchanges. Some polygons may be missing and the user does not detect.

**Scale-Resolution** - may be too little or too much detail for the application, especially when considered with all other project data.

**Quality** - may be inconsistent from data sets.

**Inconsistent classification schemes**

**All can add up to very inaccurate final products if taken in sum**

### Data Issues

**Accessibility** - who has access, how is it accessed

**Costs** - no set rules here. Some feel satellite data are overpriced, but this has to be balanced with the area of coverage and expected information to be obtained. Commercial users of data have to make a call as to how much benefit can be derived.

**Formats** - many types, not always compatible. Each new software release potentially and usually does generate some format problems.

**Standards** - there are no absolutes in standards but many are striving for common ground (**e.g. FGDC**). Standards would basically set the ground rules for accuracy, format, classification schemes, etc. This is a major issue in GIS circles. Part is being sorted through with oversight from federal agencies, universities, and major corporations.

### Bottom Line on Concerns

Don't let all of these organizational and data concerns scare you away from GIS. GIS is a constantly evolving thing that is continuously updated to provide the best information possible (remember back to circular flow diagram of an information system). Updates to the data and software/hardware environment are a fact of life. Proper planning and vigilance will maintain the system at a high level of accuracy and efficiency.