UNIT 15: LABELING

Written by Barry Bitters, Lockheed Martin Information Systems, Imaging/GIS S pecialist

---

INDEX

CONTEXT

When geographic information is stored in a geographic information system (GIS), each geographic feature will have two types of data stored:

1. Locational information in the form of explicit geographic coordinates.

2. Descriptive information in the form of project unique standardized descriptive attribute values.

For example, an outline of an area feature defined by a series of coordinate pairs will be stored to show the boundary of a quarry. In a separate location values for depth, material quarried and owner's name and address might be stored as attribute information.

A point coordinate would be stored for each building point feature with values for length, width, height, orientation, and primary use separately stored as attribute information.

For geographic features to be properly described, the descriptive attribute information must be explicitly associated with the locational information. In most modern GISs this is accomplished by storing locational information in one file and descriptive attribute data in another data file. In real-time, a relational link is used to associate the attribute data with the locational data. In modern GISs, this link between locational information and the associated attribute data is automatic and totally invisible to the operator.

Prior to using a GIS though, a GIS technician or a clerk must physically enter all the attribute information into the data file. This process is termed the GIS labeling operation. The term is an outgrowth of the computer assisted design/drafting (CAD) process of labeling features with annotation.

The GIS data labeling process is the most time consuming process in database generation. It is also the most scrutinized in most contracted database production efforts. Capturing feature positional information is relatively simple and in most projects a minor portion of the data capture effort. However, in most GIS database generation, classifying each feature and then identifying specific feature attributes constitutes the major data capture effort.

Labeling, the most time consuming task in the database generation process, can be a smoothly orchestrated set of processes or it can be a nightmare of convoluted, error-prone processing. Usually, the organized, well thought-out labeling operation will generate the best (most accurate) data, while the poorly planned, haphazard labeling operations will be plagued with errors and continual rework to repair errors. It is therefore imperative that prior to imbarking on any digitizing project a through, well thought out database generation design be created.

! Design time at the beginning of a project will prevent a lot of correction time at the end.

In this unit we will discuss the planning and execution of labeling operations with the primary intent of creating a quality finished product. Before we do this though, we must first have an idea of what might be involved in a GIS labeling operation. Although a relatively simple labeling operation, the example below has caused many problems throughout the GIS data capture community. We will use this example throughout the unit to illustrate what to do and what not to do when planning or executing a large or small labeling operation.

Example Application

An ongoing data conversion effort at most mapping organizations is to convert hard-copy map data into a soft-copy format. The U.S. Geological Survey has, for several years subcontracted the data conversion of hard-copy 1:24,000 scale topographic quadrangles of the United States into Digital Line Graph (DLG) data sets. The generation of DLGs from hard-copy maps emphasizes the capture of precise locational information and only a limited set of descriptive attributes. (For a more detailed description of the DLG format see the Digital Line Graph Data User's Guide.)

In a formal DLG data capture effort, all line, point and area features symbolized on the map are digitized as either an area feature, a line feature, a node (point feature along a line), or a point feature (in DLG parlance a degenerate line feature.) All features are classified based on the standard USGS map symbology. This is the basis for the assignment of descriptive attribute codes.

There are no limits on the number of attribute codes that can be assigned to each feature. However, there will always be at least one attribute code for each feature and generally no more than twenty attribute codes assigned.

Each DLG attribute code is composed of two distinct numeric fields: (1) a three digit major code, the first two digits of which identify the data category to which the feature belongs, and (2) a four digit minor code which more specifically identifies the feature or further describes some characteristic about a feature.

Major Code Description. The first two digits of the major code uniquely identify the data category to which the element belongs. The table below lists the major codes and the categories they represent.

02	Hypsography
05	Hydrography
07	Vegetative Surface Features
08	Nonvegetative Surface Features
09	Boundaries
15	Survey Controls and Markers
17	Roads and Trails
18	Railroads
19	Pipelines, Transmission Lines & Misc. Trans. Features
20	Manmade Features
30	U.S. Public Lands Survey System

TABLE 1. Digital Line Graph Major Codes.

The third digit of the major code is used to designate the type of minor code in two ways:

- **If it is zero, the associated minor code numbers represent a description or classification of a specific feature.**
- **If it is not zero, the associated minor code numbers have special interpretations as a parameter.**

For example, the major code 020 is a hypsographic major code and would be associated with a minor code that further distinguishes the type of feature that is being described. The major code 021 is another form of a hypsographic major code and would be associated with a minor a code that is a parameter used to describe some aspect of a feature of a particular classification. Both of these major codes are used when describing feature that are associated with the portrayal of elevation information. This includes all forms of contour lines and spot elevations.

<u>Minor Code Description.</u> Each DLG minor code is a four digit code, that when associated with its corresponding major attribute code expands on the description of a particular feature. The first digit of the minor code is usually zero. If it is not a zero, the digit is used as a modifier to provide additional information about a feature. The remaining three digits are normally used to classify specific features. The type of element described by a particular code can be determined from the table below:

Type of Descriptive Element	*Range of Values*
Node	0001 - 0099
Area	0100 - 0199
Line	0200 - 0299
Point	0300 - 0399
General Purpose	0400 - 0499
Descriptive	0600 - 0699

TABLE 2. Broad Classes of Digital Line Graph Minor Codes.

As an example the following DLG attribute major and minor attributes could be used to describe a single hypsographic point (degenerate line) feature:

020 0300, 020 0611, 020 0614, 020 0610, 021 0125, 020 0605

These values would translate as follows:

Major & Minor Code	*Translated Value*
020-0300	A point feature, specifically a spot elevation
020-0611	A descriptive attribute, specifically a depression
020-0614	A descriptive attribute, specifically a best estimate identifier
020-0610	A descriptive attribute, specifically an approximation identifier
021-0125	A parameter attribute, specifically an explicit elevation value = 125
020-0605	A descriptive attribute, specifically an explicit fractional portion of an elevation value = 0.5

TABLE 3. Translation of Example Hypsographic Point Feature.

This set of codes would describe a spot elevation at one of the lowest elevations within a depression with an approximate/best estimated elevation of 125.5 (unit of measure is defined in the file header and would usually be meters.)

This example illustrates the type of labeling that will be associated with each feature in a DLG data set. When a typical 1:24,000 scale topographic quadrangle can contain thousands of different features, how do these attribute values get assigned? There are some semi-automated attribute assignment capabilities available in the private sector, however, those that do exist are well guarded secrets. Most DLG attribute assignment is done manually. An operator seeks-out each individual feature or a group of like features, and translates the hard-copy symbology and the context of the surrounding features into a set of major and minor attribute values. This process can be performed as the linework is being digitized or as a separate interactive operation.

As with any digitizing operation, the time efficient digitizing of DLGs demands that an efficient software-driven set of processes be available to allow an analyst to rapidly, correctly, and easily assign attributes in an economical and profitable manner.

Only thru a thorough and conscientious planning effort prior to each digitizing operation, can economical and profitable labeling procedures be developed. Sloppy, poorly thought-out procedures will usually produce sloppy work. In industry, or in the commercial sector, when sloppy work is performed, then re-work is required. Re-working of a job is usually lost profit.

...and what about the analysts? Sloppy, poorly though-out labeling procedures will often result in increased operator fatigue. This can then be a cause of a significant increase in errors in the labeling process.

Learning Outcome

So first lets face the facts, labeling is a boring, often menial task. It is almost as boring as eternally performing hand digitization on a digitizing table (tracing features with a digitizing puck/cursor.) The labeling process, assigning attribute values to all features in a data set, is the most time consuming process in data capture. It is also the most critical and the most difficult process in which to maintain low levels of error.

However, there are certain concepts and generally accepted procedures that will foster an efficient, economical and operator friendly environment for labeling geographic data. This unit will discuss procedures that can streamline data labeling operations. Built-in quality control measures, procedures designed to insure an accurate finished product, will also be discussed.

Design concepts and generally accepted procedures are available to reduce operator fatigue and to reduce the potential for production of large amounts of erroneous data. There is

however, no universal replacement for good management practices. Managers will still have to oversee productivity and through normal management practices maintain acceptable productivity levels - productivity levels at an acceptable error rate.

This unit will provide an introduction to some of these design concepts and procedures that allow the increase in labeling productivity and assist in maintain low error rates.

Preparatory Units

None

Awareness

Learning Objectives

1. Introduce the student to Labeling Processes.
2. Introduce the student to Interactive Attribute Assignment.
3. Introduce the student to Attribute Quality Control Operations.
4. Introduce the student to the Planning of Labeling Operations.

Vocabulary

- Data Dictionary - A document that defines valid attributes for a digitizing project.
- Default Attribute - Derived, computed, or assumed attribute values that are assigned in lieu of valid attributes.
- Digitizing Specification - A document that in detail describes the steps that are involved in a digitizing effort. This document usually contains a full data dictionary plus detailed instructions to digitizing analysts on how to capture and attribute data.
- Graphical Users Interface (GUI) - a set of software tools to customize a computer screen graphics to aid in the creation of valid geographic data.
- Project Prototyping - The process of performing a project dry-run as an integral part of the project planning phase. The basic idea is to perform all steps in a production effort to validate the project design.

Unit Concepts

1. Data Capture Approaches - A Question of Software Design Versus Time and Cost

In the commercial data capture arena there will always exist a balancing act between software development and production time and cost. There will never be enough time to produce the perfect digitizing software and there most certainly will never be enough money. It then becomes the managers responsibility to allow the development of the best possible software

within the constraints of budget and time. In a routine production environment, time is not as critical as in a crisis production effort. (But aren't all production efforts managed like crises?)

Often, when a crisis production effort occurs time will not be available for even minimal formal software support. But managers at all levels must remenber that automation allows for the efficient execution of repetitive tasks and labeling operations are nothing more than a series of repetitive tasks - *select a feature, display its attributes, change the attributes*.

! REMEMBER: Over an entire project, the elimination of a single repetitive operator keystroke can represent a significant reduction in operator effort over the entire project.

It is therefore an essential portion of the database generation process to initially assemble a set of software tools to aid in the efficient creation of valid geographic data. These software tools, often termed a graphical user's interface (GUI - pronounced "gooy"), are the interface between the operator and complex commands that operate a computer system. Labeling operations can be performed without a user's interface, by performing command line entry at the keyboard.

However, performing GIS operations from the command line without any operator interface should be used only for very short digitizing efforts or for prototyping purposes. An operator's interface should be designed for all long term labeling efforts. The complexity of this interface is dependent on available time, money and software development assets. Four broad types of labeling software should be considered for each labeling effort:

A. Attribute Conversion Routines for Imported Data. Using attribute data assigned by some other reputable organization saves manhours, therefore, whenever possible use attribute information that comes with existing data. Economy of time dictates that the shortest time-span between the start and end of a project is dictated by the least number of processes that have to be performed. If reliable attributed data is available use it. Do not visit each and every feature and re-attribute unless the validity of the new data is in question. Perform a software conversion of the attributes to reformat them to the current project's specifications. Then only perform a quality control check rather than a full labeling session and a quality control session.

   When digitizing from a map, go so far as to even preserve the hard-copy feature color as initial attribute information. Although it might be thrown out later in the digitizing process, in some instances it might be usable. For instance, most brown features on a map are associated with hypsography. If all brown line features captured from a map are segregated into a single file, for most USGS map sheets, all non-hydrographic contour lines would be stored separate from other data.

   ! Be careful, however, because in some instances USGS has used brown linework to symbolize ephemeral (non-perennial) streams on arid alluvial fans and pediments.

   One way to preserve pre-existing attribute information is by importing existing data through software filters which reformat or convert existing attribute data to the finished format of the project. Usually, these filters will have to be custom designed for each job. They will be developed based on a set of attribute conversion rules defined in the

project data dictionary. If time does not allow creation of a project data dictionary, the attribute filtering software can be created on-the-fly. In crisis mode this is often the case.

B. Automatic Label Point Generation Routines. Locational information on area features is usually composed of a set of coordinates defining the area feature boundary and the coordinates of a label point, the entity that stores the attribute data. Prior to attributing area features a check should be performed to insure that all area features contain a corresponding label point. If the label point is not present, it is impossible to assign attributes to that area feature.

To insure that attributes can be assigned to all area features, a software routine should be available to automatically add label points to all area feature that do not contain them. After new label points are installed in the database, a check should be performed to insure that valid polygon topology exists.

This label point routine will normally not need updating from one project to the next.

C. Default Attributes Assignment Routines. There are times when a discrete attribute value can not be determined for a feature or an attribute can be calculated or derived based on other attribute values. These derived, computed, or assumed attribute values are called **default attribute values**.

During the planning phase a data dictionary should be developed. This document defines those attributes that will be used in the database. It also identifies the range of valid values that can be stored for each field. This document is the basis for all labeling within a data generation project. The excerpt of a data dictionary shown below is an example of the type of information that is shown in a data dictionary. Note that for each feature classification named on the left the following information is provided:

- Allowable Feature Types
- Classification Code
- Default Attributes
- Optional Attributes

In addition, elsewhere in the data dictionary, maximum and minimum values are defined for each possible attribute variable.

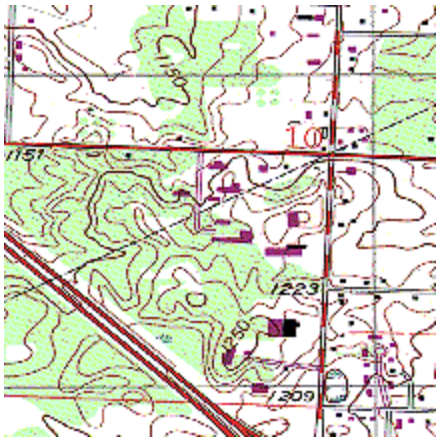
HYDROGRAPHY

Name	Feature Type	DLG Feature Classification	Default Attribute	Name	Optional Attribute	Name
Upper Origin of Stream	N	050 0001	---------		050 0300	Spring
Upper Origin of Stream at Water Body	N	050 0002	---------		---------	
Sink, Channel no Longer Evident	N	050 0003	---------		---------	
Stream Entering Body of Water	N	050 0004	---------		---------	
Stream Exiting Body of Water	N	050 0005	---------		---------	

Alkali Flats	A	050 0100	058 0000	Estimate	---------	
Reservoir	A	050 0101	053 0____ 050 0200 050 0201	Angle Shoreline Manmade Shoreline	05n _____ 050 0619	Elevation Integer Interpolated
Covered Reservoir	A	050 0102	053 0____ 050 0201	Angle Manmade Shoreline	05n _____ 050 0619	Elevation Integer Interpolated
Spring	P	050 0300	--------- 050 0615	 Hot or Mineral	050 0001	Head of Stream
Nonflowing Well	P	050 0301	---------		---------	
Flowing well	P	050 0302	---------		---------	
Cistern	P	050 0306	---------		---------	
Rapids	ANP	050 0400	---------		---------	
Falls	ANP	050 0401	---------		---------	
Gaging Stations	ANP	050 0403	---------		---------	
Pumping Stations	ANP	050 0404	053 0---	Angle	---------	
Water Intake	ANP	050 0405	--------		---------	
Canal Lock or Sluice Gate	ANP	050 0407	--------		---------	
Spillway	ANP	050 0408	--------		05n ---- 050 621-9 050 0619	Elevation Integer Elevation Decimal Interpolated
Gate (Flood, Tidal, Head, or Check)	ANP	050 0409	--------		---------	
Lake or Pond	ANP	050 0421	--------		05n --- 050 621-9	Elevation Integer Elevation Decimal

TABLE 4. An Excerpt from a Digital Line Graph Data Dictionary

Using the data dictionary as a guide, software should be developed that will automatically generate acceptable attribute values based on assumptions from other attributes or other information. These default values will not necessarily be the final value to be stored, but could later be change based on operator's interactive attribute assignment.



As an example, when scanning or digitizing linework in DLG production, feature detail is initially separated based on the map print colors. All brown line features are initially stored as hypsographic data. All black line features are stored as transportation features and all blue line features are stored as hydrographic data. This same color separation process is also performed for point and area features. To preserve the color designation for each feature, the classification of the most likely feature class is assigned as a default value. The classification that should be assigned is that feature classification that will most often be encountered in the source materials. For brown features the default feature classification will be contour line, for black line features - generally light-duty road hard or improved surface, and for blue line features - perennial stream.

By initially classifying each feature based on its map color, the majority of the features

within each color will have been properly classified. Generally, the predominant hypsographic feature on a map would be the standard contour line. Since all brown line work is initially classified as a contour line, the only features that must be reclassified are those features that are not contour lines. An analyst can perform these using interactive attributing procedures. Each contour line must still be assigned an elevation value, but this can be done using grouped classification routines - special routines that allow selecting and attributing a set of lines.

What does this approach save? The major portion of the map features of each color will automatically be attributed with a value derived based on the original feature color. Only by exception will features have to selected and attributes reassigned. This can be completed by having the analyst perform a visual inspection of the map to identify those features that are not members of the default feature classification. Then interactively, attributes are assigned to the non-conforming features.

D. Interactive Attribute Assignment Routines. Generally, interactive attributing is used most often in database generation. In this process, the analyst must select each and every feature and manually change the attributes. Although this process is time consuming, there are situations in the database generation process where there is no other way to add attribute information.

2. Data Capture Emphasis - Attributes vs. Location

The types of labeling software that are used in a digitizing operation are dictated by time, money and also by the data capture emphasis. There are two broad classes of digitizing emphasis that dictate how a database will be created. When location information is emphasized there is often a minimal set of attribute values that are to be entered for each record in the database. In this situation when positional information is emphasized and minimal attribute data is required, feature digitizing and feature attribute assignment are generally performed at the same time.

However, when an extensive set of feature attributes is involved, emphasis would be shifted to creating an accurate set of feature attribute data. In this case the database is often initially populated with features and then a separate process is performed to add attribute information to newly digitized existing features.

Generally there are five processes that are used to adding attribute information to features. The five labeling processes are:

1. Attributes Inherited from Parent Data
2. Attributes Entered Manually Into a Single Menu
3. Attributes Entered Manually into Scripted Windows
4. Scripted Manual Entry
5. Relational Database Development

3.Attributes Inherited From Parent Data Using attribute data assigned by some other organization saves man-hours. Whenever using source materials that contain salvageable attribute information, an attribute reliability evaluation should be performed. If the attribute

data contained in the source data is deemed worth salvaging, then every effort should be made to import not only the location information for each feature, but also to import as much attribute information as is usable. This will often require software development time because attributing rules are rarely the same from one data set to the next. Once software is developed, these processes usually require no operator intervention, only possibly a quality control session to insure that no erroneous data has been introduced.

Pros	*Cons*
Preserves Prior Work Effort	Requires Extensive Software Support
Requires Minimal Analyst Processing	Usually Requires Data Dictionary
Allows for Rapid Database Generation	Assumes Prior Work was Correct
	Requires Detailed Quality Control

4.Attributes Entered Manually into Single Menu.

Pros	*Cons*
Interactive Editing Process Can Be Performed While Digitizing	Very Labor Intensive Often Requires Software Support

5.Attributes Entered Manually into Scripted Windows.

This form of processing is used when all features must be visited and attribute values must be entered in an interactive environment. This is most often performed with custom-made, a simple or elaborate windowing system and is often the processing choice when attributes are assigned as linework is digitized. The manual attribute assignment process will usually be a "heads-up" process whereby the analyst will display a coverage of lines, points or polygons linework on the screen. Those features that do not have any attributes assigned to them will automatically be highlighted on the screen. When software development time is available the system will automatically center and zoom in on each feature identified as requiring attribute assignment. The analyst will then select features to be attributed from those highlighted on the screen. As a minimum, three methods of analyst feature selection should be available:

1. Selecting single features.
2. Selecting multiple features.
3. Outlining an area and automatically having all features inside the area selected.
4.

Pros	*Cons*
Interactive Editing Process Can Be Performed While Digitizing	Very Labor Intensive Often Requires Software Support

6.Relational Database Development

Pros	*Cons*

Example Application

Two broad categories of attributes will be addressed in this software. The first, the primary attribute, will be used to classify a specific point, line, area or node feature type. The second broad category is the secondary attributes. Appendix A lists all attribute codes by the DLG factor. These are further divided into groups based on the type of feature primitive: either point, node, line, or area. The analyst will be given a list of features based on the current DLG factor and the coverage being worked. The operator will then choose the feature code from a list, for those features that were selected.

Based on the primary code chosen, a set of feature code specific prompts will lead the operator through the optional secondary attribute assignment process. Figure 9 shows the general concept of the assignment of secondary attributes. Appendix B lists all primary attributes and the associated mandatory and optional attributes. (Some mandatory attributes will also require prompting. If they appear in Appendix B with hyphens in any positions, they also must be prompted.)

The three general forms of secondary attribute prompts to be used are:

A. Yes/No answers without data entry.

- Example: Is feature earthen (yes/no)?
  - If response is no, an attribute is not required.
  - If response is yes, an attribute is automatically added to the attribute file.

B. Yes/No answers with data entry. This type of attribute prompt In Appendices A and B these codes are shown with dashes in the spaces requiring analyst input of data.

- Example: Is feature coincident (Yes/No)?
  - If response is no, an attribute is not required.
  - If response is yes, an additional prompt must be used to obtain information from analyst.
  - For example:
    - What is the two digit major code of the coincident feature (99)?
    - There should also be the option that allows the analyst to enter a yes response followed by a space and then the required code. This will allow the experienced analyst to bypass the second prompt and move to the next operation.

C. Automatic assignment. Some features will have mandatory attribute values that will automatically be assigned based on the feature classification that is chosen. These values are identified during the planning phase and are explicitly defined in the data dictionary. Some of these will also prompt the analyst for additional information

Adding Attributes Interactively When attributes must be entered interactively, an iterative process should be created that will allow an analyst to repetitively select single or multiple features and assign attribute codes to them. The analyst must first select features of the same type from the screen and then assign a feature classification. The choosing of a particular feature type will then activate a set of feature specific questions to prompt the analyst through the mandatory and optional attributing process. The analyst should not have to bother with the

numeric attribute codes. There are to many to remember and numeric designations are easy to transpose, thereby causing the entry of false data. Therefore, all attribute assignment should be based on short verbal prompts, whenever possible.

This attributing process must be performed for each class of primitive individually. For example, if an analyst is working on the polygon coverage for vegetation s/he will first attribute all arcs, then change the EDITFEATURE to labels and attribute all labels.

Competency

QC of Attribute Data - Cheap Assurance of Quality

Maintaining a quality product is the most important part of any production effort. Quality production is the basis of an organization's reputation and as a consequence the basis for all future work. Quality control processing is what insures that a finished product is acceptable to the customer. For this reason, the customer must be involved in the definition of what precisely constitutes an acceptable finished product.

Quality control processing takes place throughout the life cycle of every project. However, it is usually financially impractical to have explicit quality control phases after every step in a production effort. An efficiently designed production effort will have a minimum number of phases that are devoted purely to quality control. However, quality control will be a part of every phase because when ever possible each subsequent step should serve as a quality control check of all previous processing. This approach to quality control is more cost effective in that minimal time is dedicated to explicit quality control tasks.

However, as a last phase in most production efforts an explicit quality control phase should be present. This phase can be an interactive process where analysts perform manual checks to insure the validity of the data. A more cost effective alternative to interactive quality control would be an automated batch process that would insure the validity of each and every feature and its descriptive attributes. As this automatic evaluation finds errors, a listing of those records that are potential in error is created and used for subsequent interactive editing.

Automatic error checking algorithms are not cheap. In most cases, with each new project, new algorithms must be created since source data, any intermediate processing, and finished products will vary from job to job. For this reason, software support usually will be required before the start of full production. When developing software to perform attribute data quality control, consideration should be given to the following factors:??

- All required features have been digitized.
- Features are stored as the proper feature primitive - point, line, area, or node.
- Features are properly classified.
- All required attribute values are present.
- All attribute values are within their allowable ranges.
- After edits are performed, proper topologic relationships are established.

In commercial digitizing, contractual agreements will usually contain explicit allowable error

rates. These are usually expressed as the percent of error free features. Generally, contracted data capture will be performed so that no less than 90% or 95% of all features are error free.

During quality control processing, statistical sampling is often used to evaluate the overall validity of attribute data. A statistical sampling should be checked at varying intervals during the quality control process to see that each data set is within the allowable error rate. If at any of the intervals the data set is within the contracted accuracy tolerance, the QC process can be discontinued. However, if the error rate is greater than the allowable tolerance then quality control review of the data set must continue. Data quality control continues until at one statistically sampled interval the error rate is within the allowable range.

What constitutes an error free finished product? This is dependent on the database design, however, some general conditions can be defined which will universally constitute bad data. Automatic attribute checks can be made over the entire database to isolate features that have attribute discrepancies. Examples of discrepancies that can be identified by automatic data checking include:

- Search for Blank Attribute Fields.
- Search for Out of Range Attribute Values.
- Insure that Derived and Calculated Attributes are Correct.
- Search for Incompatible Adjoining Features.
- Insure All Area Features Have Label Points.
- Insure Each Coverage Has Valid Topology.
- Search for Other Project Unique Data Inconsistences.

The design of the entire quality control process should reflect only that information that the customer has defined in the acceptance test procedure as being important. A set of customer approved attribute validation rules, usually found in the digitizing specification, is used in this quality control check.

There will be instances when time and cost will preclude the creation of automatic quality control procedures. If this is the case, then some form of manual quality control will have to be performed. In most projects, time will usually not be available to visit each and every feature and check for erroneous attribute values. Searches performed using logical querying can be used to isolate known repetitive errors, however, this type of error checking should only be used as a means of isolating errors that can not be detected in automatic error checks. To fulfill most contractual requirements, final automatic error searching is essential.

This search of the database should not have to be performed in an interactive mode. In fact, for best results automatic attribute checks should be performed in a batch mode. The final results of this automatic attribute check should be a list of features that have been flagged as having potential errors.

After automatic attribute checks have been completed, error statistics should be computed for each separate data set. This statistical analysis should compute an error rate based on the total number of features in the data set and the number of records with potential errors detected during the automatic attribute check. If this statistical analysis indicates that the data set is within the customer's range of allowable error, the data set can be considered as finished. If,

however, the statistical analysis indicates that the error rate is higher than acceptable, an interactive quality control session must be performed.

During this interactive quality control session the analyst should be guided from one potentially erroneous feature to the next. The analyst should be provided a view of the vicinity around each potential error and a view of the feature's attributes. In this way the analyst can determine whether an error is present and subsequently be permitted to make appropriate changes.

When contractually bound to a statistical error rate, and interactive QC is necessary, the interactive session should be performed until the error rate approaches and surpasses the contractually acceptable rate. During processing, at the point in time when the computed error becomes less than the contractual error rate, the interactive edit session can be stopped since all contractual obligations concerning acceptable errors have been met.

Remember that your customer will be performing the same sort of quality control checks prior to product acceptance. Insure that the statistical computations performed during batch and interactive quality control are the same as those that will be performed during your customers acceptance testing. If these statistical procedures are not the same then computed error rates may not match.

Mastery

Planning Labeling Operations

Throughout this unit it has been stated that organized, well thought-out labeling operation are a direct result of the effort put into initial project planning. This effort consists of a formal design process that entails a thorough analysis of the current requirement, available assets, and other ongoing production efforts. What should be addressed when planning an impending GIS labeling operation? The following factors should be considered during the label planning process:

- Define Requirement
- Define Assets - Personnel, Hardware, Software, Time
- Define Processing Flow
- Develop Data Dictionary
- Develop Digitizing Specification
- Define Default Attributes
- Prototype all Proposed Label Processing

Project prototyping is probably the single most important step in the overall labeling process. The basic concept is to perform all steps in the data conversion process to validate the project design. Always perform dry-runs to determine the feasibility of the planned set of processes. In this way a set of prototype objectives, defined prior to the dry-run can be validated. Prototype objectives that should be accomplished include the evaluation of:

- Processing Design

- Process Flow
- Suitability of Source Materials
- Attribute Data Conversion Accuracy
- Feature Positional Accuracy
- Quality Control Procedures

This validation should not only be used internally, but the results should also be shared with the customer. Presenting the results of the project prototyping dry-run to the customer will ensure that there are no misunderstandings between the producer and the end-user. It will also insure that the customer has effective data acceptance procedures because the customer will use the results of the dry-run to evaluate their acceptance test procedures.

Often, project prototyping is performed before the award of the contract to assist in developing a bid package. When prototyping is performed prior to the award of the contract, either the customer has overseen the prototype production operation, or the results of the dry-run are submitted in the bid package for customer scrutiny.

Even when a pilot project has been performed and a project concept has been validated, do not be adverse to change during the subsequent production effort. Remain flexibility and allow for change in the project design. The source materials used in prototyping are often not a representative sample in format or content. Often an over simplistic view of the project was used to guide the project design and the dry-run was not complex enough to identify actual design problems.

Remember, the amount of effort that goes into project design has a direct bearing on the quality of the project that is produced. If the feature attributes in a finished database are bad, it is often dificult to salvage anything of value.

Resources

Montgomery, G. E. and Harold C. Schuch. 1993. *GIS Data Conversion Handbook*. GIS World, Fort Collins, CO.

Created: May 14, 1997. Last updated: October 5, 1998.