



Integrating Quality Assurance into the GIS Project Life Cycle

Abstract

GIS databases are an ever-evolving entity. From their humble beginnings as paper maps, through the digital conversion process, to the data maintenance phase, GIS data never really stops changing. The key to developing and implementing a successful GIS project is a well-designed quality assurance (QA) plan that is integrated with both the data conversion and maintenance phases of the GIS project. The fundamentals of quality assurance never change; completeness, validity, logical consistency, physical consistency, referential integrity and positional accuracy are the cornerstones of the QA plan. This paper discusses the issues involved in the development and implementation of an integrated GIS quality assurance plan.

Introduction

Without data there would be no need for the computers, software and human resources that comprise GIS technology. Not just any data, but geographic data. And not just any geographic data, but data that is specific and reliable and that represents as closely as possible the spatial world we live in. The technology requires that the data be as clean, as healthy, as good as it can be. Neglecting that, the usefulness of the technology is short-lived. To maximize the quality of GIS databases there should exist a well-designed quality assurance plan that is strategically integrated with all facets of the GIS project.

Categories of Quality Assurance

All well-designed QA strategies have certain things in common. They must coexist within the processes that create and maintain the data. When they are not integrated within the procedures of the GIS project, they themselves can become an entry point for error. By definition, they must also incorporate key elements from the classic QA categories that are discussed below.

Completeness

Completeness is the adherence of the data to the database design. This means that all of the data conforms to a known standard for topology, table structure, precision, projection and other data-model specific requirements.

Validity

Validity is a measure of the attribute accuracy of the database. Each attribute must have a defined domain and range. The domain is the set of all legal values for the attribute. The range is the set of values within which the data must fall.

Logical Consistency

Logical consistency is a measure of the interaction between the values of two or more functionally related attributes. As the value of one attribute changes, to maintain consistency, so must the values of its functionally related attributes. An example would be the interaction between the attribute SLOPE and the attribute LANDUSE. If LANDUSE is "water", then SLOPE must be 0, any other value for SLOPE would be illogical.

Physical Consistency

Physical consistency is a measure of the topological correctness and geographic extent of the database. For example, the requirement that all electrical transformers in an electrical distribution database's GIS have annotation denoting phasing placed within fifteen feet of the transformer object is one that describes a physically consistent spatial requirement.



Referential Integrity

Referential integrity is a measure of the associativity of related tables based upon their primary and foreign key relationships. Primary and foreign keys must exist and they must associate sets of data in the tables given predefined rules for each table.

Positional Accuracy

Positional accuracy is a measure of how well each spatial object's position in the database matches reality. Positional error can be introduced via incorrect cartographic interpretation, through insufficient densification of vertices in line segments or through digital storage precision inadequacies, to name a few. These errors can be random, systematic and/or cumulative in nature. Positional accuracy must always be qualified, because after all, it is only just a map of reality.

The following section discusses the general stages of GIS database creation, from its start as an existing map product to its final stage as a seamless, continually maintained database. At each stage the integration of the QA plan within the process is discussed.

Map Preparation

Contrary to popular belief, not all maps are ready to be digitized in their current form. The first step in creating quality GIS databases from paper maps is map preparation, sometimes referred to as map scrub. Map preparation is the foundation for subsequent steps of the data conversion process and is the most cost-effective phase to detect and correct errors in the data.

Control Review

Establishing coordinate control for the database is the most important step in the data conversion process. Whether using benchmarks, corner tic marks or other surveyed locations, these must be visible and identifiable on the map source. Each control point should be reviewed to make sure it has a known real world location. Every dollar spent on coordinate control is worth at least two dollars that is spent later dealing with positional accuracy problems.

Edgematch Review

Edgematching is a critical component of the map preparation process. Edge features (those that cross as well as those that are near) must be reviewed with respect to logical and physical consistency requirements as well as noted for positional accuracy and duplication. The temporal factor must also be taken into account. If the ages of adjacent maps differ greatly, then there are bound to be edgematching errors between these maps. Cultural features are especially prone to this problem.

Primary Key Validation

Map information that will be converted to key information must be reviewed. An example would be a fire hydrant number that will be used as a key to relate to other hydrant information. Fire hydrant features in the GIS database without numbers or with duplicate numbers will eventually cause referential integrity errors.

Conflict Resolution

Another critical issue involves multi-source conflict resolution. If a layer in the GIS is to be compiled from multiple map sources, then there are bound to be conflicts between the original map data. An example of multi-source conflict resolution may be an electrical layer that is being compiled from two maps, an overhead map and an underground map. These two map series must be reviewed in conjunction with each other for duplicated features, conflicting positional locations and conflicting feature attributes. Think of this as "vertical edgematching".

Data Conversion

The act of creating digital data from paper map sources does not make the data more accurate, but actually introduces more or different error into the data. The goal of high quality data conversion is to limit the amount of random and systematic error introduced into the database. Random error will always be a part of any form of data, whether it is analog or digital. Random error can be reduced by tight controls and automated procedures for data entry. Systematic error, on the other hand, must be removed from the data conversion process. Systematic error usually stems from a procedural problem, which once corrected usually clears up the systematic error problem. The key to correcting both random and systematic error is a tightly integrated plan that checks both automatically and visually at various stages in the conversion cycle. A short feedback loop between the quality assurance and conversion teams speeds the correction of these problems.

RMS Errors

Registration of paper maps or film separates to a digitizing board, or to images with known coordinate locations introduces registration error. The amount of error is calculated by the Root Mean Square (RMS) method. Each feature digitized into the database will have an introduced error equivalent to the RMS error. The goal during registration is to minimize the RMS error as much as possible. Standards must be set and adhered to during the data conversion process. High RMS errors, in some cases, point to a systematic error such as poor scanner or digitizer calibration.

Visual QA

At various stages in the data conversion process visual QA must be performed. Visual QA is meant to detect not only random error such as a misspelled piece of text, but also systematic error such as an overall shift in the data caused by an unusually high RMS value. Existence and absence of data as well as positional accuracy can only be checked with a visual inspection. This visual inspection can be performed in two ways; hardcopy plots and on-screen views. The hard copy plotting of data is the best method for checking for missing features, misplaced features and registration to the original source. On-screen views are an excellent way to verify that edits to the database were made correctly, however they should not be a substitute for plotting.

Visual inspection should happen during initial data capture, at feature attribution, and then at final data delivery. At initial data capture the data should be inspected for missing or misplaced features, as well as alignment problems that could point to a systematic error. In either case each error type needs to be evaluated along with the process that created the data in order to determine the appropriate root cause and solution.

Automated QA

Visual inspection of GIS data is reinforced by automated QA methods. GIS databases can be automatically checked for adherence to database design, attribute accuracy, logical consistency and referential integrity.

Automated QA must occur in conjunction with visual inspection. The goal of the automated quality assurance is to quickly inspect very large amounts of data and report inconsistencies in the database that may not appear in the visual inspection process. Both random and systematic errors are detected using automated QA procedures. Once again the feedback loop has to be short in order to correct any flawed data conversion processes.

Data Acceptance

Defining acceptance criteria is probably one of the most troubling segment of the GIS project. Which errors are acceptable? Are certain errors weighted differently than others? What percentage of error constitutes a rejection of data? The answers to these questions are not always obvious and require knowledge of the data model and database design as well as the user needs and application requirements. Project schedule, budget and human resources all play a role in determining data acceptance.

Acceptance Criteria

Accepting data can be confusing without strict acceptance rules. One example of this confusion is a GIS data set with ten features, each feature containing ten attributes. If one of the features has one incorrect attribute, is the percentage of error 1% or 10% (10 x 1%)? If you subscribe to the theory that all of the attributes make a feature correct, then if one attribute is incorrect the entire feature is in error making the error percentage 10%. On the other hand, if only one attribute is incorrect for a feature, and it is treated as a minor error, then the error percentage is 1% where 1 out of a possible 100 attributes is incorrect.

For a minor attribute the 1% error may stand, but imagine if the attribute in error is a primary key where the value is not unique. This seemingly minor error cascades into the database where relationships to one or more tables may be jeopardized. Weighting attributes by their importance solves this problem. Each attribute should be reviewed to determine if it is a critical attribute and then weighted accordingly.

Additionally, the cartographic aspect of data acceptance should be considered. A feature's position, rotation and scaling must also be taken into account when calculating the percentage of error, not only its existence or absence.

Error Detection

Once the acceptable percentage of error and the weighting scheme have been chosen, methods of error detection should be established. The methods of error detection for data acceptance are the same as those employed during the data conversion phase. Check plots should be compared to the original sources and automated database checking tools should be applied to the delivered data. Very large databases may require random sampling for data acceptance.

Data Maintenance

The data maintenance stage of the project lifecycle begins once the database has been accepted. GIS data maintenance involves additions, deletions and updates to the database. These changes must be done in a tightly controlled environment in order to retain the database's integrity.

Controlling the Environment

Data-specific application programs generally are written to control the maintenance of a GIS database. These applications play two roles in the database update. First, they automate the database update process and second, they control the methods in which the database is updated. The benefit of this rigid control is that it provides the user with only one point of entry into the database, thus improving the consistency and security of the database. Maintenance applications rely upon a set of business rules, which define the features, their relationships to others and how they are to be updated. Maintenance applications are very dependent upon a static database design and a database that conforms to the design. The applications are usually supported by a database management system. This system is frequently composed of permanent and local (temporary) storage systems. Data is checked out from permanent storage into local storage for update and then posted back to the permanent storage to complete the update. In this environment, pre-posting QA checks are required to ensure database integrity. The data storage manager must maintain the database schema so that table structure and spatial data topologies are not destroyed. Automated validation of attribute values should also be a part of the pre-post QA. Visual check-plots are not out of the question for instances where large amounts of data are either added or removed.

Scheduled Validation

Scheduled database validation is also a must for large multi-user databases. This is similar to the 60,000-mile checkup for your car and can identify some very important and potentially costly errors. These errors may point to a lack of control during the database update process. Errors or last minute changes in business rules, bugs in the maintenance application or inconsistent editing methods can all be detected during scheduled validation. It is always cheaper to fix a bad process than to correct hundreds or even thousands of errors that may have been introduced into a database.



Is Quality Assurance Worth the Cost?

High quality data sets support high quality analysis. Software applications that interact with and manipulate GIS data make certain assumptions about the data to perform the jobs that they were designed to do. A classic method of spatial analysis, spatial overlay, requires good registration between layers, as well as high attribute accuracy and consistency. GIS analysis using poor data yields poor analysis.

High quality GIS databases facilitate sharing. Without some assurance of cleanliness it is hard to market or share data with others. The ability to advertise a high quality database and back it up with a solid quality statistics helps break down barriers to data sharing.

The decision making process that is supported by GIS is frequently liable for its decisions. The results of locational analysis such as flood plain evaluation or hazardous waste siting can be disastrous with poor data. Critical customer service information such as medical priority in facilities management or addressing in 911 databases can mean the difference between life and death.

In the past, issues of quality were often minimized because of the additional short-term costs associated with it. The average GIS database is very expensive to create and maintain. The additional burden that a well-designed QA plan makes on the GIS project budget is understandable and foreseeable. Conversely, the potential cost of problem analysis, application revision and data reconstruction that routinely accrue when QA is overlooked far outweigh the initial cost of a well designed and implemented QA plan. Protecting the organization's investment in its data is very important. Assuring the quality of expensive data is equivalent to insuring a home or business against catastrophe.