LaurelHillGIS
Geodatabase Integrity Solutions℠

Preparing for Data Acceptance: Are you ready?

Abstract

Receiving large amounts of GIS data can be complex, not just from a procedural standpoint, but also from the staffing and training perspective. An efficient, streamlined data acceptance process promotes a more stable application development environment and speeds database implementation. This paper describes the planning and preparation an organization should consider to ensure a successful data acceptance phase for a GIS project. These preparations include scheduling, acceptance criteria, client/vendor communication, problem tracking and resolution, database design configuration management and system administration.

Introduction

Data acceptance can be a daunting endeavor. From large projects with multiple deliveries containing 30 or 40 maps to smaller projects with 1 or 2 deliveries of 5 to 10 maps, good acceptance planning is vital to project success. Data acceptance is not only a technical process of reviewing digital data and comparing the results to known criteria, but also an administrative process linked directly to the project contract.

Advanced Planning

Data acceptance generally occurs after the request for proposal has been delivered to the vendors, the data conversion vendor has been selected, the project schedule has been set and the staff has been allocated. Planning for data acceptance however, must begin well in advance to ensure project success. In most cases, the data conversion project is just the beginning of further GIS activities. Application development and training may proceed in parallel to data conversion with the expectation that the database will be completed at a certain time. These expectations put pressure on the data conversion process making it the critical path within the organization. If data conversion is the critical path within the organization, then data acceptance becomes the critical path for data conversion; without accepted data, none of the other planned projects can get started.

Acceptance Criteria

Developing acceptance criteria is oftentimes the most difficult part of the data acceptance process. Errors occur in every data conversion effort. It is important to eliminate systematic errors and minimize random errors. Small random errors are tolerated in different amounts in most GIS databases. Spelling mistakes, an occasional misplaced or omitted feature, an incorrectly coded attribute, are small random errors. The intent of the quality assurance process is to limit these errors. On the other hand, systematic errors should be identified and corrected as soon as possible and should not be allowed to effect the health of the database. Examples of systematic errors are poor registration, poor edgematching, data that does not adhere to the database design, topological inconsistencies and consistent sets of invalid or out-of-bounds values. Systematic errors usually point to a flaw in the data conversion process.

Prioritizing the data

The first step in creating acceptance criteria is to prioritize the importance of the data. The most important aspects of the database should be the most heavily monitored in the data acceptance process. Criteria like horizontal control, adherence to the database design schema and conversion specification may take a high priority. These are foundations of the GIS database and little error should be tolerated. Conformance to the physical data model, database completeness, cartographic feature placement, feature connectivity, feature proximity and referential integrity might be the next priority. This data, though important, may contain errors due to interpretation. Attribute accuracy in terms of ranges, domains and logical consistency might be the final priority. Attribution errors should be monitored for systematic errors; however, random errors in this category are likely.

Error Detection Methods

Once the data has been prioritized, the method for detecting errors must be determined. In general, there are three error detection methods: automated applications, semi-automated applications and visual inspection.

Automated quality control applications detect data that does not conform to the database design or physical data model as well as problems with referential integrity and attribute accuracy. These applications can test the data in bulk and report the results. Automated testing should be used in conjunction with semi-automated and visual methods since it will only detect a portion of the potential errors in a GIS database.

Semi-automated applications test the data with production applications to determine how well the data will perform under real world conditions. For instance, an electrical utility might run connectivity tracing or network feeder management applications on the data to detect complex conversion errors. Semi-automated methods can detect feature connectivity errors, feature proximity errors and data that do not conform to the conversion specification. The semi-automated methods may output reports for review or may programmatically visit each error for the technician to visually inspect.

Visual inspection includes comparing hardcopy plots with original source material and on-screen viewing of the data. Visual methods will detect horizontal control problems, absence of data, cartographic feature placement, vertical integration and data that do not conform to the conversion specification. This is a labor-intensive process that may require a random sampling of plots in order to effectively inspect large amounts of incoming data.

If creating GIS data was compared to using a word processing application, then automated testing would be the spell checking program, semi-automated testing would be a grammar checking program and the visual inspection would be the final proof-reading. All three steps are necessary to create high quality documents as well as GIS data.

Error Categorization

Once the methods for detecting errors have been determined, the next step is to categorize the errors. Certain types of errors will be pass/fail. This means that if an error is detected then the data fails the acceptance. The data conversion vendor should not deliver any data that does not meet the pass/fail requirements. The highest priority data should have pass/fail requirements.

Other types of errors will be quantifiable; the number of errors can be counted and compared to the number of features or attributes tested. Categorizing the errors as quantifiable will allow for percentages of error to be calculated. Allowable error percentages should be part of a contractual agreement between client and vendor.

Other types of errors may be too subjective to be quantifiable. It is important that both the client and the vendor have good communication when deciding acceptance or rejection of this error category.

Error Calculation and Weighting

Calculating data errors and weighting the results must be done with care. The complexity of the database design dictates the type of error calculation method that should be employed. Below are two examples of simple calculation methods and one example of a weighted error calculation method. Each is geared toward a specific database design and acceptance philosophy.

The Feature Acceptance Criteria calculation method entails counting all the features equally. An error in any counted attribute fails its feature and is counted as one error. A counted attribute is one that is monitored for accuracy. If a feature has ten counted attributes and four were not filled correctly, then the number of errors represented by that feature is one. Therefore, the total number of possible errors for a group of features is the total number of features. This method is feature-centric. It favors simple database designs and the determination of rejection/acceptance is straightforward and easily calculated.

The Compound Acceptance Criteria calculation method counts all the errors in all the counted attributes equally. A counted attribute is one that is monitored for accuracy. An error in a value of a counted attribute is counted as one error. If a feature has ten counted attributes and four were not filled correctly, then the number of errors represented by that feature is four. The total number of possible errors for a group of features is therefore the total number of features multiplied by the counted attributes per feature. This method favors large, complex database designs that have many attributes of relatively equal importance.

The Attribute Weighted Acceptance Criteria calculation method counts the errors in counted attributes differently based upon a predetermined weighting scheme. A counted attribute is one that is monitored for accuracy. Each attribute is given a weight that reflects its relative importance. The importance of an attribute can be measured in several ways. It may be required for an application to run correctly or to ensure accurate analysis, or it may be necessary for legal reasons. An error in a value for a weighted attribute is counted relative to its weight. This method favors large, complex database designs as well as those that contain a mix of attributes of varying importance.

Acceptance Management

To keep the project moving forward acceptance of data must be closely managed. Error information must be fed back to the vendor in a timely fashion so that updates to procedures and processes can be made for future data deliveries. Most likely there is an agreed upon turnaround time for the data to be accepted or rejected, usually two to four weeks.

Staffing

In order to keep the acceptance process on schedule there must be adequate staffing. For a small project, this might be just one person. For a larger project there may be a dedicated visual quality assurance group and an automated quality assurance group consisting of several technicians per group. If there are not enough technicians to review the data then acceptance will slow and there is a greater possibility that data will have to be accepted without review. Accepting data without proper review is never a good situation and should be avoided when possible.

Training

Having a staff to review the data is one thing, having a well-trained staff that truly understands the data acceptance process is another. The acceptance technicians must know the physical database design, the data conversion specification and the acceptance procedures. They should be proficient in the organization's GIS software. If they are reviewing hardcopy plots, they must be able to identify data errors, prioritize and classify them into acceptance categories. If they are running automated quality assurance applications, they must have an understanding of how the programs work and what is the expected outcome is for each type of test. This may seem like a lot to ask, but keep in mind that once the data is accepted, the responsibility for the data quality has shifted from the vendor to the client. It may be very difficult or impossible to have the data conversion vendor correct a data problem once the data is accepted. The only recourse is for the client to correct the data and make sure that future deliveries are correct.

Dealing with Rejection

The decision to reject data is often a painful one. The conversion vendor's schedule will be affected, the acceptance schedule as well as the overall project schedule may be affected. It is critical that both client and vendor understand the acceptance criteria and how errors are to be calculated. Well-defined acceptance criteria will make clear the determination to accept or reject the data. Strict early enforcement of the acceptance criteria will eliminate late project data rejection and keep the overall data quality high. Conversely, a lack of criteria enforcement will lower the quality standard of future deliveries to the level of the data that is accepted.

Database Design Configuration Management

One of the most costly mistakes that can be made on a project is changing the database design during the data conversion process. Once the pilot project has been completed and the conversion vendor starts into production

conversion, the database design should remain as static as possible. The database design should freeze in terms of the schema or the physical data layout. This means that layer names, table names and table definitions should not change. The conversion specification should also remain as static as possible. The conversion specification details exactly how the vendor should capture the data. If these specifications change, the processes for capturing the data must also change. The net effect of these changes will be a cost in either schedule or dollars. The one aspect of the database design that should be flexible however, is the valid values for attributes. At the beginning of the project, all of the discrete values for an attribute may not be known. As the vendor converts the data, they will often run into new values that were unforeseen at the beginning of the project. Adding new valid values for an attribute may be more cost effective than changing the database design.

Problem Tracking and Resolution

In the data conversion process, there are a lot of unknowns. We try to classify all of the source data so that it fits nicely into the database design and conversion specification. Oftentimes during the conversion process, the vendor comes across situations where the source(s) do not fit into the database design. There may be illegible data sources or missing source coverage for a particular geographic area. A Problem Action and Resolution (PAR) system must be developed to deal with each of these issues in an organized and timely manner.

The response to the vendor must be as quick as possible so that its production line is not suspended waiting for instructions. A specific response time should be determined and agreed upon by both the client and the conversion vendor, in most cases 24 or 48 hours. In order to accomplish this, a system must be put into place to track the PARs as they come in, evaluate the situation, find the solution and respond to the vendor. A standard form for problem descriptions should be adopted. This form can then be faxed between both parties to facilitate the problem/resolution cycle. Another option is an Internet system that is accessible by either party at any time of the day.

It is imperative that there be a detailed log of the vendor/client interaction. Employee turnover during the life of the project is inevitable. The loss of client or vendor staff should not throw the project into chaos. From the client standpoint, problem tracking is a critical path issue where the entire project schedule could be jeopardized because of untimely responses to vendor questions.

Data Acceptance System

How is all of the incoming data to going to be managed? Is there enough disk space to deal with these large amounts of data? Will the network be able support data acceptance traffic? Are there enough licenses of the GIS and automated acceptance software? These are the kinds of questions that the client should be asking when planning for large-scale data acceptance. The data conversion vendors ramp up projects so that they can create a data conversion factory for the project. The client should plan to ramp up the same way creating a data acceptance factory so that once large amounts of data start to arrive, the system can handle the data flow.

Computer systems must be dedicated to accepting data so that the process is not impeded. The computer network must be able to withstand access of large datasets without slowing the entire organization. The computer system directory structure must be organized so that at all times the acceptance technicians know exactly where each data delivery is located. Additionally, a unique delivery naming convention should also be adopted to minimize confusion and make locating specific delivery easy.

The original data must be carefully archived so it can be referenced later as acceptance proceeds. It is also important to be able to pinpoint errors in the database once it is in the maintenance mode. Errors found during maintenance are either caused by client side interaction with maintenance applications or are conversion errors that were not detected during the acceptance phase. Knowing the root cause of database errors can speed the correction of either the maintenance applications or the acceptance process.

Piloting the Plan

Once all the planning is complete, these assumptions must actually be tested. No project goes exactly according to plan but, by testing and evaluating the plan, revisions can be made and expectations can be managed.

Too much data

The data conversion vendor may deliver more data than can be reasonably reviewed within the agreed upon timeframe. This amount of data might be exactly what the overall project schedule calls for in order to finish on time. In this case, either the conversion vendor must slow the deliveries or more people or streamlined procedures must be put into place to handle the data throughput. Conversely, the overall project schedule may need to be revised in order to accommodate the existing staffing levels.

Not enough data

Another situation might occur where not enough data is delivered from the vendor. The entire project schedule may be jeopardized. In this case, no amount of staff changes is going to make the data come in any faster. The earlier in the project that this is known, the better. New projections can be made based on the current data flow. Other projects that are relying on completed data by a certain date can revise their schedules based upon these new projections.

Process Tracking

To correctly balance staffing levels with incoming data, it is necessary to track each step in the acceptance process. Tracking systems can take many forms: manual systems where each step is checked off on a form, automated systems that use spreadsheet technology, geographic systems that utilize maps and databases.

Regardless of the system that is used, it is important to clearly document the amount of labor and computer time that each step takes in order to identify bottlenecks and critical path processes. By knowing exactly how long it takes to review and accept the data, accurate projections can be made for future deliveries.

Quality data is the goal

Receiving quality data from the conversion vendor is the goal of data acceptance. Setting detailed, enforceable acceptance criteria will enable the process. It is imperative to track data processes, staff levels, training levels and schedules. Balancing all of these factors while creating good communication channels can sometimes be overwhelming. Planning for the acceptance process is key to receiving quality data in a timely fashion.