Geodatabase Quality Control, now more important than ever.

Abstract

The ArcGIS geodatabase has unprecedented ability to share data within the enterprise through industry standard relational databases and outside of the enterprise with internet map services. Additionally the core ArcGIS software has streamlined data creation and maintenance, freeing up programming resources to develop more complex applications to model and analyze data. The value of the geodatabase is directly related to the quality of the data that resides in it, now more than ever quality data is paramount to the success and credibility of the enterprise. This paper discusses the methods and strategies for safeguarding the enterprise geodatabase to ensure that this valuable asset does not degrade over time.

File based data architecture

File base data architecture such as coverages, shapefiles and CAD files were the norm up until the emergence of ArcGIS 8x and 9x. These data formats are easy to use on an individual project basis; however enterprise usage of data in these formats required implementation of complex data managers. To solve the inherent limitations of the file based data architecture, tile level and feature level locking of the data had to be employed. This was just not the answer for the enterprise.

File based systems also created a data management nightmare. It was very easy to propagate several copies of a given dataset without knowing which the most up to date copy was.

Each department of division within the enterprise was responsible for their data and its quality. Data sharing across the enterprise was not a commonplace; therefore data errors might go unnoticed for long periods of time.

ArcGIS enterprise geodatabases utilize industry standard relational databases that control and administer the data across the enterprise without feature locking. Access, on some level, to the GIS data is now available to virtually every user in the enterprise.

This access allows for and unprecedented ability to share data across the enterprise, unfortunately data error is also more easily shared.

Reactive verses proactive approaches to quality control

There are two approaches to database quality control, reactive and proactive. The reactive approach is to correct errors as user transact with features and rows in the database. This approach gives the users and database administrators no clear understanding as to the total number of errors in the database at any given time. It also does not address the root cause of the data errors. The proactive approach takes the information from an initial

quality assessment and error analysis to develop an error correction plan. This approach not only, takes into account the data errors as they currently exist, but tries to determine the cause of each type of error. The goal of quality control is not only to detect and correct errors in the database, but to identify the causes of the errors to prevent future errors from occurring.

## Initial Quality Assessment and the Initial Quality Baseline

Once data has been loaded into the geodatabase an Initial Quality Assessment is prepared. This assessment points out existing errors in the data and establishes the initial quality baseline for the geodatabase. The assessment should include the total number of errors, which is useful for total error observation over time. The assessment should also include the number of errors by feature class and column within the feature class. Individual statistics will help troubleshoot the root causes of the errors over time.

Ongoing Database Audit and Scheduled Validation

Ongoing testing and database audits are compared to the initial quality baseline that was established. In Figure 1, below, the aggregate initial quality baseline is in dark blue. This is the known quality level at t_0. At time t_1 the total number of database errors is increasing. As the data is being edited and updated errors are not being trapped at the point of entry. This increase indicates processes that are not fully under control. At time t_2 the total aggregate error level has dropped. This indicates that overall processes are under control and now the database is realizing proactive data quality gains. At time t_3 error rates continue to drop, indicating the proactive approach to quality control is ensuring the database integrity.

Figure 1 Aggregate error over time
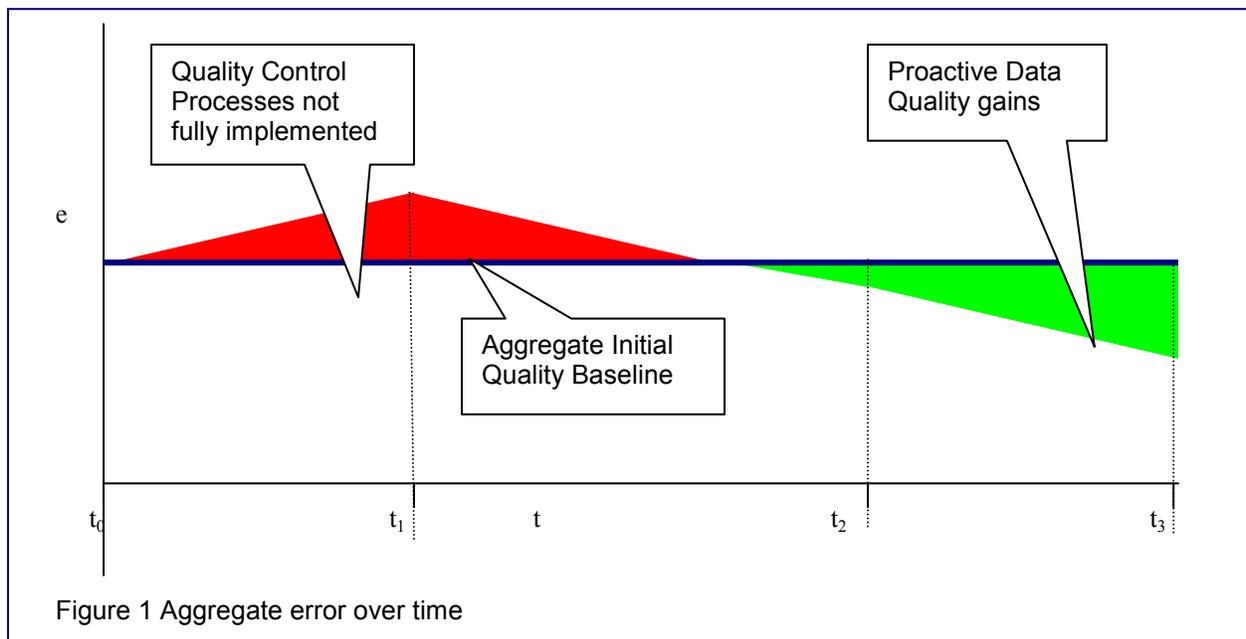
The initial quality baseline when aggregated helps give an understanding of the overall errors in the geodatabase. In order to have a detailed understanding of errors by feature class or by column, then the errors must be analyzed at the featureclass or column level. Figure 2 shows how the aggregate error totals decrease over time, while particular errors in columns increase over time.

Figure 2 Aggregate errors versus individual errors
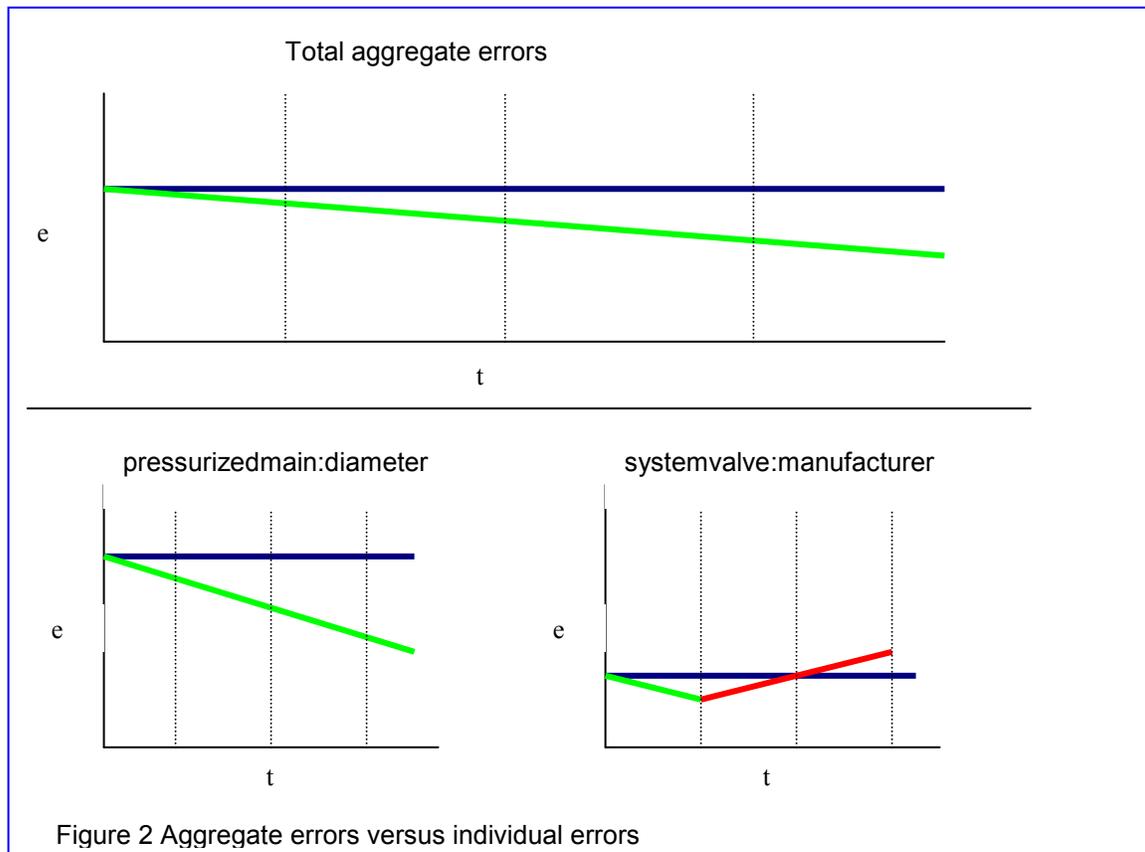
There are three causes to increasing error rates, application bugs, user errors and security gaps in data access. Application errors could include basic logic flaws or general design flaws. User errors generally indicate a need for training refreshers, or a need for tighter application control of data entry. Other errors can be attributed to security gaps where there is access to the underlying data outside the confines of the application environment. If there is access to the data outside the application environment, modifications to the data cannot be controlled and could cause wide-ranging errors in the database.

As the total aggregate errors are drop, it is critical to observe individual column or row error rates. A large decrease in errors for a particular tested column may dilute small, but significant, increases in errors for another tested column. The individual errors for pressurizedmain:diameter have a significant decrease, while the errors for

systemvalve:manufacturer have a slight increase. The net effect is a decrease in the total aggregate error; however the issue is that there is still a process that has not been brought under control. This problem usually occurs at the beginning of the error correction phase. Large numbers of data corrections may be made initially concealing small numbers of errors that are being added to the database. Therefore it is essential to perform routine scheduled validation of columns and rows in order to track individual error rates over time.

Conclusions

The past decade has seen some monumental changes in GIS. The way we store data, access data and share data will never be the same. The one constant component to GIS is that quality data makes for quality analysis. Just moving data into the enterprise geodatabase does not make it more accurate or give it more integrity. Only the highest quality data will yield the best analysis and foster the most user confidence. The credibility of GIS in the future required quality data.

Knowledge is the key. Finding and correcting every error is not an attainable goal. Understanding where, what and why about the errors that existing in the geodatabase is most important. Proactively addressing data errors with schedule validation and documenting them with metadata are keys to retaining the credibility and user confidence in the geodatabase.