STANDARDISATION AND CENTRALISATION OF DATA – DATA QUALITY AND DATA QUANTITY

David R. Jenkins

Terra Search Pty Ltd, P.O.Box 981, Castletown Queensland 4812
drj.ts@bigpond.com

ABSTRACT

Digital data management in the Minerals Industry is still in the developmental stages. Many data management procedures currently in place are more suited to hard copy reports than digital files. Data management practices in our industry need to change in response to the major changes in communications and digital technology occurring throughout the world. Standardisation of data in terms of the way data is gathered, collated and stored allows rapid integration and analysis on any scale. Standardisation on a global basis is a difficult task due to the diversity of attributes that are collected. A system needs to be comprehensive, flexible and user friendly to allow standardisation.

The attributes collected also need to be standardised to ensure completeness of datasets. The codes used should be as consistent as possible and codes should not be concatenated to reduce the number of fields in a file.

Operating system standardisation, while less important, must allow free movement of data between software packages. Media for storage should be easily readable and preferable online.

Duplication of data should be avoided as much as possible. Centralising data on an office level removes the risk of having multiple copies of the data without knowing which if any is the most up to date version. Centralisation of data on a wider scale should only occur when the data is needed centrally. Maintaining of two versions of a database is inefficient if the data is only used in one of the locations. A more elegant solution is to have the databases linked on line to allow querying and retrieval when required. This requires consistency between systems in each location.

A standardised data structure will solve many of the long-term data problems while also streamlining current data gathering. Centralisation on a project level is critical. Broader centralisation of a group's data may be desirable but will only be cost effective if standardisation of data management has occurred on the same scale.

INTRODUCTION

The exploration industry is a group of intensive data gatherers and the data collected is the major asset of an exploration company - at least until the stage of resource calculations. The nature of data has undergone a major change over the last 15 years with the advent of the PC and, more recently, advances in GIS, plotting and database software, data has been moving from hard copy to digital at an increasing rate. Similarly the amount of data readily available to explorers has expanded significantly over time. The issues of how this data is managed have often been seen as having minor importance. While that may have been a reasonable position 15 years ago it is now a critical decision effecting the overall efficiency of a company. Most inefficiencies and costs are created when data is lost. This loss may result from not recording an important attribute when collecting the data through to the complete loss of a system with inadequate backups. Of all the issues effecting data, standardisation and centralisation have the greatest effect on data management.

Information and the communication of information are becoming more and more powerful tools available to a corporation. This is reflected in the strong growth and preeminance of telecommunications company and media conglomerates in Australia over the last few years. Both sectors are undertaking massive programs to ensure their effective operation in what is becoming known as the information age. The benefits to our industry may not be as obvious or profound as they are for these sectors those groups who can take the most advantage from the new technologies will lead our industry.

It has been said that we are in the midst of a revolution as significant as the industrial revolution – the information revolution. The keys to benefiting from the revolution will be the adaptation of new technologies to suit the mining industry and the conversion of information into knowledge. No longer will the gold assay of a stream sediment sample with its northing and easting be sufficient. You will also require a third dimension but also analytical details, the geological data, the environmental data, a digital terrain model and the structural regime of an area. A model requiring this amount of detail requires high quality data and those who recognise this will be collecting and storing these attributes now. The fact that you do not currently use a particular attribute should not stop it being recorded as long as it can be recorded and stored at minimal cost.

As more data is collected from each sample or observation point the need for comprehensive storage methods becomes more acute. Any data storage system must encompass the need to have a standard methodology for the storage and retrieval of data and the flexibility to cope with all future data requirements.

Standardisation of data can be divided into two aspects – format and content. Both aspects are equally important in ensuring the usefulness of the data in the short and long term. Format issues include the data structure - both in terms of directories and file structures, and data storage - in terms of media and operating system. Content issues include the language or codes used for storage and the methodology for data collection.

Historically the mineral industry's record of storing data is varied. While there are excellent procedures in place for documenting and archiving exploration activity in hard copy, storage of digital data has been more ad hoc. Examining the general methods of these examples can give insights into the best data management procedures.

The centralisation of the hard copy reports at the various States' Mines Departments has been key to the effectiveness of the hard copy system. Standardisation has only played a part in the cataloguing of the reports. The reports themselves have a wide variety of styles and content. This has not detracted greatly from the quality of the data due to the way the data is accessed and we have found that over 90% of historical data is still retrievable to a high degree of accuracy.

Digital Data is of a vastly different nature to hard copy reports. There is an enormous variety in the format and content of digital data, as there is with hardcopy. In the case of digital data, however, the centralisation of files is less advantageous than having detailed knowledge of the file formats (or a standard file format) and content. Without this knowledge the breadth of application of this data and the ease of retrieval is severely affected.

In our experience retrieval of digital data from companies in a useful form would only represent around 50% of active projects and less than 20% of archived projects. Often a digital file will contain only partial data omitting attributes less critical to first pass analysis but crucial for full and detailed analysis. Quality control is also an issue with little information available in digital format on the analytical details, collection methods and processes completed to ensure the quality of the data.

The reasons for the disparity in the amount of useful data retrieved from reports compared to digital files is that the data management systems for hard copy systems are being applied to digital systems. This is clearly inappropriate with consistent data loss from digital practices. Format and content changes in digital files both between projects, and over time, is the main cause of this loss. Often data managers have failed to recognise the need to change data management practices to reflect the changing nature of data and data analysis.

STANDARDISATION

Standardisation of data format in terms of operating system and software compatibility is critical when bringing together multiple datasets. At the data gathering stage it is consistency in content that will improve data management. Content issues include whether particular attributes are collected consistently, inconsistently or not at all, the codes being used and the consistent coding of like attributes.

Standardisation should be made as automated as possible. Systems that shepherd the users towards consistency increase data quality. While no system can stop inconsistent coding of the same lithology, a system can identify non-standard codes, missing fields and missing data without relying on individual geologists or data gatherers. Once the data is standardised then centralisation becomes not only easier but also more useful.

Standardisation is the key requirement of good data management. This has been recognised for a long time by many participants in the computing and telecommunications industries. Examples of standardisation in these areas include Morse Code, ASCII and the Internet Protocol. Languages and dictionaries are other earlier examples. For exploration and mining data there are four main areas which effect standardisation. These are:

- Data Collection.
- Data Structure.
- Codes (languages) used.
- Storage Media (including software and hardware).

The standardisation of data collection methods should include ensuring that all relevant and readily collectable attributes are consistently collated by using templates and logging guides. Also efforts should be made to ensure that the coding remains consistent for similar lithologies between drillholes, drill programs and geologists.

Standardisation of storage requires the storage of all attributes not just some. A Sample S, with attributes S_1 to S_n requires a system that can cope with n variables. If n is fixed then this is a straightforward process if n is variable, as it always is in a geological environment, then the data management system needs to be able to cope with that change. If n is very large then you also need to ensure the management system is built in a way that the attributes are efficiently stored. The more comprehensive a system is the more complicated it is. This complexity requires investment in software and methodologies to make the system easy to use. Comprehensive systems are of no use if they are incomprehensible. The data storage should be in a form the user can quickly and easily retrieve and understand the key information.

The way in which we store each attribute is also important. Fragmentation of attributes to their smallest elements is desirable in many instances. The storage of data as elements rather than concatenations allows more flexible use of the information. Consider an example of storing the results, as a single field, of tossing two coins - the possibilities are: HH; HT; TH; TT. A lookup table

describing each of these notations would thus need to hold four entries. Storing the results in two fields each with possible entries of H or T requires just two entries in a lookup table. This example shows by splitting the information the storage of library entries and thus the overheads for validation is reduced by 50%. This facilitates standardisation of the data by allowing faster and better quality control.

With a more complex example such as lithologies and their major descriptors such as texture and composition the storage of this data in a single field causes many future problems. For example if a Quartz rich sandstone with crossbedding is coded as a single coded – qzSDSTxbe then the data is neatly stored in a single entity and extractable and plottable as a single entity. However, if you have 100 minerals, 300 lithologies and 400 textures then the number of possible codes you can build with these becomes astronomical – 100 * 300 * 400 = 12 000 000. Due to this variety standardisation in a single project using concatenated codes is difficult and on wider scales nearly impossible.

CENTRALISATION

Centralisation of data is the collation of multiple datasets into the one place. It is an issue that must be examined in terms of scale. On the small scale the integration of data from two surveys over the same area is an obvious requirement. On larger scales, however, the advantages are diminished or at least become less obvious.

Within an office or project the centralisation of digital data has obvious and immediate advantages. Data that is fragmented can be updated and used independently of other data or other copies of the same data that can easily result in the loss of information. Fragmented data in multiple files and locations also relies on specific knowledge of the location of the data. This specific knowledge usually resides with an individual within an organisation. If the individual is unavailable, it is common that the digital data is difficult or impossible to retrieve and decipher. The collation of all data to a single database in an office removes the need for the specific knowledge and the data should then be retrievable on basic criteria such as project, prospect or location.

On a wider scale the collation of a company's data into a single database will have similar benefits. The centralisation of an entire company's data is also a useful archive guarding against data loss from a particular office and will make the data available to all relevant groups within the company. Having a centralised version of the data as well as a project based version can cause disparity between the two versions. To avoid this users should always access the centralised copy of the data. There are however disadvantages in accessibility with current WAN speeds. The day to day requirements for access to data usually forces a local copy of the data to be maintained with a method of replicating changes made to the two copies of the information.. This maintenance will cost time and money and, unless the data is being regularly used in more than one office, it is probably unnecessary. If each office or project use different data formats, codes or methodologies then the centralisation becomes a difficult, time consuming and costly process.

Centralisation of data into a single database may not be as effective as networking of a number of key databases. In this scenario each office would have a copy of the data within their own jurisdiction and through WAN links would also have access to other offices data. Requests for other offices data could then either be processed on or offline depending on the size of the request the urgency for the data and the speed of the connection. Why spend the time and money transferring, maintaining and storing data in two places when it is mostly used in the one place. More sensible is to retrieve the data from the one place when needed.

Without a standard structure in each office either method of centralisation is made more difficult. If the structures are standardised then the information can be seamlessly combined or fragmented as required with no reassignation of fields and data.

CONCLUSION

A standard data structure will solve many of the long-term data problems while also streamlining current data gathering. Centralisation on a project level is critical. Broader centralisation of a company's data may be desirable but is only cost effective if standardisation of data management has occurred on the same scale.

Received: October 1999
Published: April 2000