Seismic Data Lifecycle and the Importance of Data Quality.

Keith Capelin, Troika International Limited - 2014.

Seismic data is a very costly product that is expected to conform to strict data standards, and is subjected to rigorous quality control (QC) measures during its creation. These QC measures will involve the quality assurance (QA) regime of the seismic contractor, which is usually re-enforced by the contractual specifications of the client, and overseen by an independent consultant.

In light of the extensive standards seismic data is subjected to, raw data received from a seismic survey is expected to be of a high quality and well documented. However, any Data Set will consist of a variety of elements which are typically separate from the data items: the seismic data, positioning data, logs, and reports.

Data processing is carried out on the raw seismic data which produces more data products. This results in a Data Set that is comprised of a large number of data items on various media and in a range of standardised and non-standardised data formats.

There is an ongoing need to manage the Data Set during its formation and onwards through its lifecycle. Managing the Data Set includes maintaining a database to facilitate and organise the various items, and from an archival point of view a large database to manage a large number of Data Sets.

Seismic Data Sets are huge (often running into terabytes) and tend to grow larger through their lifecycle as more intermediate processed products are added. One notable contributing factor is historical data being archived along with transcribed data because of fears over possible lost data in transcription (QC issues).

The geophysical industry has defined a number of standard data exchange formats for both raw and processed seismic data, as well as for related positioning data. For seismic data these are SEG-D and SEG-Y in a number of revisions as defined by the Society of Exploration Geophysicists (SEG). For positioning data the standard is SPS as defined by the SEG and The International Association of Oil and Gas Producers (OGP - was UKOOA), P2 and P1 in a number of revisions.

A couple of fundamental requirements to achieving a Data Set of good quality are adherence to the industry accepted formats and computer readability. Data errors may degrade data, but if the errors are properly documented and easily accessible, then the data

may be tolerable. Identifying and documenting errors and issues in data items are fundamental requirements for good data quality.

A seismic Data Set is usually stored offline once the processing sequence is completed, but may be recalled and re-used for a number of reasons; it may be resold in whole or in part to a third party, data may be re-processed to investigate other geophysical potential within the survey area, data may be re-processed to take advantage of new processing techniques, a Data Set may be added to another overlapping or adjacent Data Set to image a larger area, data may be used as part of a time-lapse seismic sequence to model how a reservoir is being depleted during production, or data may be required to be transcribed or copied for statutory archive requirements.

Current common practice is to store data on tape (magnetic cartridge) for long term archiving and to reload it to a disk drive when it is needed. Long term storage is often physically managed by boxing cartridges and reloading handled by robotics. Some assets are moved towards long term storage on magnetic disk rather than on cartridges (mainly by disk manufacturers), but the problem is still the sheer volume of seismic data when weighed against intermittent use.

Data may be used several times (in whole or in part) during its lifecycle, which creates an urgent need for the durability and quality to be maintained. Physically the data is stored on magnetic media, often tape, which may have become outmoded as computer technology advances. It is necessary to review the data media from time to time to ensure there is no physical degradation of the media which could cause loss of data, and to ensure that the media reading devices are still available, connectable, and supported. There is usually an ongoing requirement to transcribe data from potentially outmoded media types to new media types. Transcribing to a new media type (usually smaller in physical size but larger in data content) means re-organising the Data Set and possibly losing information on labels. Therefore the data is impacted by computer system changes, operating systems, device drivers, physical interfaces, as well as the tape handling devices and media types.

Seismic data quality can be adversely affected during these lifecycle changes. Numerous component elements of data can easily become lost without proper management, especially supporting reports. For example logs and other metadata which are possibly stored separately, maybe even as paper or PDF/Excel files on DVDs or worse still floppy disks, can be mishandled & misplaced by various users throughout the data's lifecycle. Data can be lost during transcription due to process or manual error or perhaps due to poor handling, quality control and reporting. Each manual intervention step also allows further opportunity for error to occur. Data can also be lost due to media degradation or become inaccessible due to hardware obsolescence. What is important is that data remains accessible, that quality is maintained and any degradation or error is observed, documented

and the information accessible. Generally seismic data is robust, but unforeseen errors can cause major interruptions to computer processes and introduce a need for further manual intervention and potentially costly problems.

Current business process trends such as Business Process Management (BPM) seek to automate the management of Big Data, and large seismic data sets seem like ideal candidates. However, seismic data is "different", it is complex and usually made of many different files; it has errors, data issues, format variants, incorrect metadata, and missing data all of which make these new trends in data processes realistically impractical to implement. When data errors are flagged and recognized, automated processes can possibly be configured to accommodate them. When there are data issues/errors that are unknown within a workflow, they will cause problems in computer processes which require time consuming manual investigation and often result in failure. It might be the case that 80% of the data is clean but the remaining 20% with issues is probably distributed throughout the entire Data Set and will cause repeated process interruptions.

In summary, the future of working with seismic data will become more automated by moving towards the BPM model, but this will rely on good quality clean data to process. From a corporate standpoint, seismic data has represented a large investment and is an irreplaceably valuable and important company asset; it has to be maintained through standardized QC measures to monitor the quality of its value. Investing in good quality practices rather than looking for the lowest cost solution to transcription will simultaneously make data management easier, less costly, and extend seismic data's lifecycle. Therefore by investing a relatively small amount of time and money into the early stages of seismic data's lifecycle, the stability and profitability of the lifecycle increases naturally.

Troika specialises in developing software tools to manage seismic data during its lifecycle, while also provides consultancy, advice, and third party quality control services on seismic data. Troika's management are passionate about data standards, and have been actively involved at various levels with the development and introduction of data standards. Troika has a vested interest in helping to improve seismic data quality, and can highlight the lucrative opportunities presented to the industry by having good quality seismic data.