Geoscience Australia Data Repository – Information about Client Data Delivery

12 July 2010

This document provides more detailed information about the data you have received from Geoscience Australia's Data Repository.

Delivery Media

All data are delivered to you on electronic media. You will have received data on a USB memory stick or disk. If you requested data on tape you will have received data on one or more 3592 tape media in addition to the USB disk.

Tape Data Source

All data you have received have come from Geoscience Australia's Repository RDS Data Storage system. These data have been captured on the RDS as images of media on which the data were received. With very few exceptions, we have provided to you data as we have received it. We have no tried to correct errors in the data, nor to convert non-standard data to a standard form. Hence the data you receive on disk is identical to the source medium that contained it. If you received data on tape it may contain data from more that one input tape concatenated to a single output tape if the data is from a single survey. Data from separate surveys is always written to separate tapes.

The only case where there may be a difference from the original is where there were extra or missing tape file marks on the source tape. The data itself has been recovered to the greatest extent we have been able but we have not reproduced errors in occurrence of tape file marks.

Source Media Identifiers

Each physical medium received by Geoscience Australia's Data Repository is given a bar code label with a unique identifier comprising a single letter followed by exactly 8 digits. This is the Source Medium Identifier (SMI). The letter component of the SMI may be one of:

Prefix	Meaning
L	Produced by Geoscience Australia's onshore (Land) program
M	Produced by Geoscience Australia's offshore (Marine) program
P	Received from a Petroleum Industry exploration company under requirements of Commonwealth legislation

Data on tapes captured in the RDS are stored as an image of the data on tape with the SMI as the Directory name. When Geoscience Australia delivers data to clients on disk we duplicate the SMI directories. If we have delivered the data on tape, the disk we also send will contain all related files about the data except the data we have written to tape.

Source media

When we load a source medium, we record three additional data files:

1. A PDF™ image of the tape. This contains whatever information was printed on the tape label when the tape was produced. This file has the SMI as the file name with a .pdf suffix
2. A log file which records the transfer of the data from tape to disk. This provides a summary of what happened as the data was loaded and may assist in understanding errors or problems in the data. It has the SMI as the file name with a .log suffix.

3. An index file which is a record of the record structure of the data as it was read from tape. The format of this file is described below. The index file has the SMI as the file name and a .idx suffix.

The source medium we have used may not be the original medium on which the data were recorded because we may have remastered the original media to new, higher density media at least once and possibly more times.

Data files on Disk

In most cases the data from the tape is written to disk so that one set of data records followed by a tape file mark is written to a single disk file. The data file names are structured as follows:

`FILE_nnnnnn.suffix`

Where:

`FILE_` is the file name prefix

`Nnnnnn` is a sequential number starting at 1 and incrementing by 1

`suffix` is a three character suffix determined by the data type.

The following suffixes are used:

Suffix	Meaning
sgy	SEG-Y data
tar	Unix tar format data
nav	Navigation data
dat	Unknown or non-standard data format

Quality Control Files

Geoscience Australia applies a Quality Control (QC) process to data loaded to its RDS. The nature of the QC process depends on the type of data being loaded and this is explained below under the relevant data type headings. The QC information is included with the data we deliver to you and can be identified by the suffix. The file name is the same as the file name of the data, the suffix indicates the content and follows:

Suffix	Meaning
csv	Comma Separated Values data. This is the output of a QC process in CSV format. These files can be read using a spreadsheet program.
err	This is an error file where any problem we found with the data is reported. Typically this is where the data does not conform to the relevant standard. If no .err file is present, then no errors were uncovered.
asc	This is an ASCII text file. These are produced when we encounter text files such as navigation data in EBDCIC and reproduce the file in ASCII as part of the QC process
lis	This is a list file produced from a tar file by the Unix tar command. It is produced as a validation of the tar file and contains a list of the file names in the tar archive.

Index Files

Index files are the record of how the data was recorded on tape and are used to:

- reproduce data to tape; and
- aid reading the data for our QC programs, especially when there are errors in the data as read from tape.

Data are read from tape in one of two ways:

1. Each set of data records is written to a separate sequentially named file (Sequential form). This is used for SEG-Y data and for most kinds of non-seismic data
2. All the data on the tape is written to a single file (Single form). This is used for SEG-D data. The file name is the SMI of the source medium and the suffix depends on the data type, usually .sgd.

The index file always has the same name as the SMI and a *.idx* suffix. The structure of the index file varies slightly between Sequential and Single forms.

The first line of the Index File is a Path record which shows the destination directory. This can be ignored.

Each subsequent line reports either:

- A File record noting the commencement of a file[1] of data on tape
- A sequence of record lengths in bytes;
- An End of Tape File indicator;
- A repetition of a previous file;
- A total file count for the tape marking the end of data on the tape.

The File record comprises the text “File”, a file sequence counter and a file name which is the name by which the data described in the following records is stored. An example is:

```
File 1  FILE_000001.sgy
```

which says that the first file on the tape will be written to the named file.

If the data is being written to a single file rather than sequential files, the file name element of the File record is present only on the first File record.

The list of record lengths follows the File record. It is a sequence of comma separated values indicating the lengths of each tape record until a tape file mark is found. Spaces are added for readability. If more than one record has the same length then the repetition is indicated by a number in angle brackets. For example:

```
3200, 400, 6240 <1818>
```

indicates a sequence of record lengths 3200 byte, 400 bytes and a 6250 byte record repeated 1,818 times. Thus the total tape file is 1,821 records.

When a Tape File Mark is encountered, an End record is written. An example is:

```
End 1  11354160
```

This marks the end of the first file on tape and reports the total number of bytes read form the file of data on tape.

If two or more sequential files on tape have the same structure, this is recorded with a Repetition line. The format of the repetition line is slightly different for the sequential and single file forms.

Sequential form:

[1] In tape structured data the term file refers to a sequence of records terminated by a tape file mark. This is different from a file of data on disk.

```
File 12  FILE_000012.sgy
3200, 400, 4240 <138>
End 12  592960

Repeatfile 13 FILE_000013.sgy
```

This says that sequential file 13 has a record structure identical to sequential file 12.

Single form:

```
File 1  P00537719.sgd
4064, 3860 <315>
End 1  1223824
Repeat 91
File 93
4064, 3860 <301>
End 93  1169784
File 94
4064, 3860 <315>
End 94  1223824
Repeat 1125
Total 1219 files
```

This shows that the first file comprised a 4064 byte record followed by a 3860 byte record repeated 315 times. This file structure was repeated 91 times, then a File 93 had the same first record but the 3860 byte record was repeated only 301 times. Then File 94 returned to the originals structure and this was repeated 1125 times.

The final record in the Single form example above shows a Total record which marks the end of data on the source tape and notes the total number of files recorded from the tape. Below are examples of Index files for SEG-D, SEG-Y and tar format tapes:

When the data are well formed the Index file can be concise and short. If the data are poorly formed, containing missing records or irregular records, then the Index file may be long. It is in any case an accurate description of the data as read from the source tape.

Example 1: SEG-D Index file

```
Path /gpfs_cache/open/segdtapes/P00601115
File 1  P00601115.sgd
544, 5140 <240>
End 1  1239284
Repeat 6577
Total 6578 files
```

Example 2: SEG-Y Index file:

```
Path /gpfs_cache/open/segytapes/P00435538
File 1  FILE_000001.sgy
3200, 400, 10480 <37773>
End 1  395875120
File 2  FILE_000002.sgy
3200, 400, 10480 <29996>
End 2  314372160
File 3  FILE_000003.sgy
3200, 400, 10480 <51408>
End 3  538769920
File 4  FILE_000004.sgy
3200, 400, 10480 <24643>
End 4  258272720
File 5  FILE_000005.sgy
3200, 400, 10480 <24340>
End 5  255097280
File 6  FILE_000006.sgy
3200, 400, 10480 <18482>
End 6  193705440
```

```
File 7  FILE_000007.sgy
3200, 400, 10480 <25653>
End 7  268857520
File 8  FILE_000008.sgy
3200, 400, 10480 <30299>
End 8  317547600
File 9  FILE_000009.sgy
3200, 400, 10480 <31107>
End 9  326015440
File 10  FILE_000010.sgy
3200, 400, 10480 <33228>
End 10  348243520
File 11  FILE_000011.sgy
3200, 400, 10480 <66154>
End 11  693308000
File 12  FILE_000012.sgy
3200, 400, 10480 <45954>
End 12  481612000
Total 12 files
```

Example 3: Tar tape index file:

```
Path /gpfs_cache/open/othertapes/P00486136
File 1  FILE_000001.tar
10240 <60003>
End 1  614440960
Total 1 files
```

Geoscience Australia Data Repository – Information about Client Data Delivery

Seismic and Related Data

Seismic data may be in any of the following forms:

Format	Field Data	Processed Data	Disk File	3592 Tape
SEG-D	✓			✓
SEG-Y	✓	✓	✓	✓

Seismic data is always associated with the survey that produced the data and is organised on the disk you received in separate directories for each Survey.

Separately on the disk in a Directory called **`DataSummary`** is a file which lists each survey, the directory name used for the Survey, and the SMI for each source medium produced for you in the request. This file is in plain text format and should contain a complete list of the seismic data you have received and how to find the related disk files.

Ancillary data is included in the disk you have received. Ancillary data includes:

- Navigation data
- Observer's Reports
- Survey Reports
- Velocity data

You have been provided with whatever information Geoscience Australia has available for each kind of information. The data is located in directories under the Survey directory named for the kind of data they contain.

Seismic data tapes are either in SEG-D or SEG-Y format. SEG-D data is always field data. SEG-Y data may be field or processed data. Normally SEG-D data is delivered on tape and SEG-Y data is delivered on disk, although SEG-Y data may be delivered on tape as a special request.

Navigation data may be delivered on tape or on disk. Where it is delivered on tape we deliver the data on disk unless there is a special request to deliver it on tape. Navigation data may be field or processed data, and may be in standard form such as P1/90 or may be in an internal proprietary format.

We have received some source tapes containing a mixture of seismic and other data. While most source tapes are delivered in the format applicable to the data, some contain archives of files of the data, usually in Unix *tar* format. If we received data delivered in this way it is delivered to you in the format in which it was received.

Attributes of data in each of these formats are described in more detail below.

SEG-D Data

SEG-D data can be delivered only on 3592 tape. You receive one or more 3592 tapes containing the SEG-D data from a given survey. As best we can we write the data to tape in the order it was recorded. The SMI codes for the source media should be printed on the tape label and are included in the information we have provided to you on disk. SEG-D data frequently contains errors and we hope the QC information we provide will help you to use the data. The sources of useful information are:

- The Index file for each source medium.
- The PDF file for each source medium
- The CSV file for each source medium

The index file shows the record structure on each source medium. If a SEG-D tape is well formed then this is very short, showing the structure of the first tape file and the number of times it occurs on the tape. If the SEG-D tape is poorly formed, with bad or missing headers, missing traces or missing tape file marks then the index file may be much longer.

The PDF file shows what information is written on the tape label of the source tape.

The CSV file list in CSV form each General Header record as defined in the SEG-D standard. The complexity of this listing depends on how extensively the options in the standard have been used, but the file reports the values for each defined general header field as encountered in the data. For large data sets these files may be very long and may overflow some versions of standard spreadsheet programs.

Tape Label Records

The SEG-D standard (at least its more recent versions) specifies a 128 byte label record to occur at the start of the tape. These may or may not be present on tapes we have received. When they are present we have reproduced them on the output tape. If a tape label is present its content is reported in a .err file and in the CSV file.

In some cases of concatenated tapes we have received, tape labels from tapes input to the concatenation have been reproduced along the output concatenated tape. These are also reported in the .err file and the CSV file.

SEG-Y Data

SEG-Y data is normally delivered as disk files but may on special request be delivered on 3592 tape. If it is delivered on disk then it is contained as a set of sequential files in directories formed from the SMI of the source tape. If the data is delivered on tape then the disk will contain all related files but not the SEG-Y data files.

Each file of SEG-Y data is named sequentially as described above. For each SEG-Y data file there is a .csv file produced by the QC process which has the same name as the SEG-Y file and has a .csv suffix. This file contains:

1. The 3200 byte EBCDIC reel header converted to 40 lines of 80 column ASCII text
2. The data from the 400 byte binary reel header as defined in the SEG-Y standard. The Binary reel header values are indented so it starts at the second column from the left side. They are presented in three columns. The first column is the defined name of the value. The second column contains the value detected and the third column contains the meaning of the value or the unit of the value as defined in the standard.
3. A list of all non-zero trace header values in the file printed as one line per trace with the names of the trace header values at the top of the list. This list is indented so it starts at the third column from the left side.

This file can be long, but it is easy to see the values supplied in the reel and trace headers and hence to obtain a quick view of the range and quality of information in the data.

If the QC process detects data that does not conform to the SEG-Y standard, it reports this in a file that has the same name as the file contain the SEG-Y data with .err as the suffix. If there are systematic faults in the data, such as the trace sequence numbers being incorrectly formed then these files may become large.

The SEG-Y directory also contains:

- The Index file for each source medium.
- The PDF file for each source medium

The index file contains the record structure as read from the source tape and may be useful if there are errors in reading the data.

The PDF file contains an image of the label of the source tape. Both these files use the SMI as the file name with the relevant suffix.

Data received in *tar* Format

Some data is received by Geoscience Australia on tape in Unix *tar* format. This is a file archive format produced from a disk file system. If we receive data in this format then we provide the data to you in the same format. Data received in a *tar* archive may include:

- SEG-Y data files
- Navigation data files
- Images of seismic sections in CGM format.

All kinds of data may be contained in a single tar archive. More than one tar archive may be present on a single source medium.

Each tar archive is captured as a sequentially named file in the form described above with a suffix .tar. We have attempted to validate the tar archive by producing a listing using the *tar* program. This listing is a list of the file names in the tar archive and has the same name as the tar archive file and a .lis suffix.

There are some cases where a tar archive is appended to the data on a tape containing data in another format, such as SEG-Y. In this case the tar archive is included in the SMI directory of the tape but with a .tar suffix and with a .lis file as well.

Navigation Data

The navigation data we provide depends on how the data was received by Geoscience Australia. If we received navigation data on tape in a standard format we have captured in on the RDS in the sequential file name structure described above. Navigation data files loaded from tape have a .nav suffix. If the data files are in EBCDIC we have attempted to validate them by converting to ASCII and we have included the files in the SMI directory with a .asc suffix, written as 80 column records.

If the data was not formed this way we have not include ASCII files.

If the data was in a non-standard format such as in “internal proprietary” format we have loaded the data as sequential files with a .dat suffix but we are unable to apply and QC processes to such data.

Navigation data SMI directories also contain:

- The Index file for each source medium.
- The PDF file for each source medium

The index file contains the record structure as read from the source tape and may be useful if there are errors in reading the data.

The PDF file contains an image of the label of the source tape. Both these files use the SMI as the file name with the relevant suffix.