The CREWES data library

Rolf Maier, Kevin W. Hall, and Han-Xing Lu

ABSTRACT

The data most frequently used by researchers and students are being put into a unified format described below.

INTRODUCTION

Data on disk have always been easier and quicker to access than data on tapes. Furthermore, disk storage has been cheaper than magnetic tapes for several years now. Hence, the data sets most frequently requested by researchers and students have been collected on disk over the last few years. This allows not only for quick access and backup, but also simplifies the task of converting the data into a common format following a common set of specifications.

DATA FORMAT

The aim is to have data stored in one basic format, although there are two variants depending on the data's age and whether they have been prepared. Prepared data sets are thus far all in a version of SEGY standard rev. 0 format described in the following subsection, while unprocessed, newly acquired field data, data from the physical modelling system, and data from numerical simulations, are in SEGY rev. 1 format (Norris and Faichney, 2002). Of course, there are many older data sets, but they are not collected on any disk.

Not all data are, nor can be, stored in SEGY format. While the number of samples in the trace header is a four-byte integer and can hence get quite large, the file header takes only a two-byte integer for the number of samples per trace which is insufficient for many data from continuous monitoring experiments. Hence, some data are in SEGD format.

Prepared data sets

In order to simplify data access, a small on-disk seismic data library has been assembled over the years, driven by the necessity to keep popular data sets readily available. Most data are in SEGY standard rev. 0 format (Barry et al., 1975) as IBM floating point numbers. All known information about the data for which there is a predefined location in the reel header is put into the SEGY text (ebcdic) header.

Trace header word locations are only partly standardized, and the standards are generally adhered to only partly. Since CREWES has enjoyed the availability of ProMAX for many years, the header word locations ProMAX assignes by default are used. Hence, for information such as first breaks, station number, inline number, and cross-line number, we use the same (ProMAX) format for every dataset when written as a SEGY file. The location of any non-standard header word is defined in the text header. The following little table lists those trace header words which are not part of, or differ from, the rev. 0 standard:

ProMAX mnemonic	header bytes	type	description
sou_sloc	181-184	4I	source (station) number
srf_sloc	185-188	4I	receiver station number
iline_no	189-192	4I	inline number, in 3D
xline_no	193-196	4I	cross-line number, in 3D
cdp_x	197-200	4R	x coordinate of CDP
cdp_y	201-204	4R	y coordinate of CDP
R_LINE	205-208	4I	receiver line number, in 3D
S_LINE	209-212	4I	Swath or sail line number, in 3D
FB_PICK	213-216	4R	first break pick time
geo_comp	217-220	4I	the component number in 3C
tr_type	221-224	4I	1: live trace, 2: zeroed out bad trace

Most of the items in this list are not defined in the rev. 0 standard and are assigned to undefined header bytes. The exceptions in this list are the first and last item.

The first item is defined simply as a unique number pertaining to a shot. The standard calls this the energy source point number and uses bytes 17-20. Sometimes the field file ID (FFID) is used; we use the shot point. Using the shot point as a unique identifier implies that shots are not repeated, or that repeated shots are all bad but one, and only the good shot has been kept after preprocessing.

The last item, the trace type, called the trace identification code in both versions of the standard, would be in bytes 29-30 if the standard were followed.

Not all data have the trace kill flags set for bad traces. If it is set then the data are preserved. This allows for the testing of noise removal algorithms and the easy preservation of trace distances which many algorithms depend on.

RELATED FILES

If geometry is in the headers then there likely exists geometry information in the form of a SEGP1 file, and observers' notes. If a SEGP1 file does not already exist then one is made during preprocessing. If first breaks and trace kills are in the headers then the geometry has been put into the headers in ProMAX/CREWES style by Han-Xing Lu.

Observers' notes are often in a data base format; upon preprocessing a text file is created. If the observers' notes are on paper then a PDF version is added to the data directory.

AVAILABILITY OF DATA

Data from each project are kept in a separate directory, and all data directories are collected in one directory. Data which are not open to everyone have restricted access.

The following data are currently ready for immediate use: 2008 and 2009 field shool data, separated by acquisition system and format (SEGY and SEGD); recent Priddis data from the land surrounding the Rothney Astrophysical Observatory (RAO), again separated by acquisition system and format and including land streamer and hydrophone data from a

well; and 2008 Spring Coulee data of vertical and three-component geophones.

Both accoustic and elastic Marmousi data are also available. A locally generated set of accoustic Marmousi data is also available in MATLAB format.

REFERENCES

Barry, K. M., Cavers, D. A., and Kneale, C. W., 1975, Report on recommended standards for digital tape formats: Geophysics, **40**, 344–352.

Norris, M. W., and Faichney, A. K., Eds., 2002, SEG Y Data Exchange Format: Society of Exploration Geophysicists.