

Transfer Formats

Electronic data transferring, whether downloading from our archives to a client or uploading from a client to our archives is an important part of NRI's ongoing service. Most data transfers are from a disk file of some type or another from one node to another node on a network. There are few problems in file transferring as long as sender and receiver know the transfer protocol and know about the file being transferred.

Providing on-line electronic transfer of seismic data that has historically been accessed from magnetic tape media, presents the problem of preserving the tape's blocked format on magnetic disk. This is a concern since the electronic transfer will seldom be directly from or directly to tape media, but rather from and to magnetic disk. Therefore, the sender and receiver must understand each other with respect to the way this problem is resolved.

There have been several formats proposed to handle this problem. NetDriven Solutions has looked at most of them and has initially decided to include the following options in our client download software. A description of the format and its suitability with respect to various formats are indicated.

Additional download options will be considered and may be added as time and demand warrants. The following table lists the code and descriptions used to populate the edm_data_transfer table. A more detailed description of the formats follows.

Transfer Format	Description
1	Single File Byte Stream (.tape, .tif, or .txt)
2	Index Data File Pair (.idx & .tape)
3	Eight Byte Preamble Header (aka Lacey)
4	Four Byte Preamble Header and Postamble Trailer (Geo-x)
5	Dual Files Per Shot (Vanguard-like)
6	Xtended InDeX File Pair (.xidx & .tape)

Note: Currently, only formats 1,2,3,4,6 are implemented in the NDS Download Server program. Option 5 is available by interactive download through the NDS Query Archives program.

SINGLE FILE BYTE STREAM

This format supplies the data set as a single disk file that consists of all the 'blocks' in the data set concatenated together one after the other. No added pre or post block encapsulation information is present.

This format is suitable to data set formats that were originally in this format, that is, were disk files to begin with. It would also be suitable to data sets of fixed record/block size provided that the receiver/sender knows the record/block size, or where the data set itself defines the record block sizes.

Examples of data sets suitable for transfer in this format would be:

- tiff file, usually already a byte stream disk file

The NDS Download Program

- text file, again usually a byte stream disk file
- SEG Y, known 3200 byte EBCDIC header, 400 byte binary header and known 240 byte trace header, number of data samples in a trace is indicated in the trace headers and the data sample type is indicated in the binary header so data sample size of trace can be determined.

Software accessing these files must know the file format to correctly handle it. The user would be responsible for combining the downloaded file with the desired software.

EIGHT BYTE PREAMBLE HEADER (aka LACEY format):

This format is used to place any type of data set from tape media to disk media. On tape, the data set may consist of multiple files of variable or fixed sized blocks or a single file of variable or fixed size blocks. The disk file created is a byte stream file where all the blocks are concatenated together prefixing each with an eight byte preamble header.

The eight byte preamble header consists of two four byte unsigned integers. The first is the sequence number of the block within the data set, starting from one. The second is the byte size of the block directly following the preamble header. An end of file mark, for those data sets that include them, would be indicated with a preamble header by itself with a byte size of zero.

Support software is required to extract the blocks correctly by reading the preamble header to determine the size of the next block.

This format is already used by some existing copying and pre-processing software when copying data from tape onto disk or from disk to tape. It is also used for storing datasets onto CD-ROM.

If demand exists, or resources become available, NetDriven Solutions may develop, in the form of a linkable object library, software support for accessing datasets in this format.

FOUR BYTE PREAMBLE HEADER AND POSTAMBLE TRAILER:

This format is used to place any type of data set from tape media to disk media. It is similar to the eight byte preamble header described previously. It uses a four byte unsigned integer preamble header that indicates the number of bytes in the following block. After each data block a four byte postamble trailer is present indicating the length of the data block it followed. This format allows for easier backward movement within the data file.

Support software is required to extract the blocks by reading the preamble header for the data size before reading the data block and then to read over the post-amble trailer. This format is already in use by some processors in-house.

The NDS Download Program

Support software has to be created by the client at this time. If demand exists, or resources become available, NetDriven Solutions may develop, in the form of a linkable object library, software support for accessing data sets in this format.

INDEX DATA FILE PAIR (.idx - .tape format):

This format supplies the data set as two disk files: a data file and an index file. The data file is just like the byte stream file described previously, it is identified with a '.tape' file extension. Since this '.tape' file contains all the data blocks concatenated together there is no way to know where a block starts and how big it is, thus the need for the index file. This file is identified with a '.idx' file extension and contains records that describe the location and size of blocks in the '.tape' data file.

Each block in the original data set, including end of file marks, has an index record to describe it. Every index record is 16 bytes in size, the very first index record being slightly different than the rest. The first index record is used to indicate how many more index records there are in the '.idx' file. Bytes five through eight in this index record are used to represent an unsigned four byte integer count. Bytes one to four and bytes nine to sixteen are set to zero.

For the remaining index records the following fields are defined:

- a four byte integer field used for indicating this block's relative block number within the data set, starting from one
- a four byte integer field optionally used to contain the SEG file/record id of the block
- a four byte unsigned field used to indicate the byte offset of the data block from the beginning of the '.tape' file
- a four byte unsigned field used to indicate the byte size of the data block

NOTE: To indicate an end of file mark, typically used after a header data block pair on original tape media, an index record is used where the byte size will be zero and the byte offset will be equal to the following block's byte offset.

NOTE: To process the data set in this format, a program must access both the '.idx' and the '.tape' file. The '.idx' file is used to extract the blocks from the '.tape' file. If used correctly the '.idx' file can be used to skip blocks and end of file marks in the '.tape' file.

This format is useful for any format that may have varying block sizes, and uses end of file marks for separating files within a data set.

Support software is required to extract the blocks correctly. This format was initially used extensively in house at NRI ON-LINE INC. The need for additional fields in the index led us to develop an extended version.

Support software has to be created by the client at this time. If demand exists, or resources become available, NetDriven Solutions may develop, in the

The NDS Download Program

form of a linkable object library, software support for accessing data sets in this format.

DUAL FILES PER SHOT (aka VANGUARD):

This format is used to place SEG field format data sets from tape media onto disk media. On tape, the data set usually consists of multiple files each consisting of a small header block, a large data block and an end of file mark. In an attempt to maintain this header-data block separation on disk, each block is written to separate files creating a pair of header/data disk files for each 'file' on tape.

To identify the header/data file pairs, they share the same file name made up from the file number extracted from the header block appended to a common prefix. The common prefix is usually 'FILE' or just 'F' and the file number is usually appended with leading zeros so that the file names are the same length and can be listed/sorted alphabetically. For example file 1 on a tape could have a disk file name of FILE0001 or F0000001 (NDS uses the latter).

A file extension is used to indicate if the file is from a header or a data block. Optionally a part of the file extension may indicate the original SEG field format. For example an extension of '.hb' and '.sb' could be used to indicate a SEG B header and a SEG B seismic block respectively. Selecting the extensions so that the header block will appear alphabetically before the data block is desired, thus the use of 's' for seismic data rather than a 'd' for data in the example.

A problem arises when file ids are duplicated within a data set to be transferred. A duplicate file may overwrite or be appended to a previous header-data file pair with the same file number. To avoid this an offset can be added to the duplicate file id and to subsequent file ids to maintain the alphabetic ordering of the files. Since most SEG field format ids are 4 Binary Coded Decimal digits, an offset of 10,000 is suitable. This is the scheme NDS uses and explains why the single letter 'F' file prefix is used as mentioned above.

The header/data file pairs are usually placed in a directory that is named to identify the data set, likely by line identification or reel identification. Included in this directory with the header-data file pairs would be a .log file created to list the actual ids with the created header/data file pair names and the header and data block sizes.

SEG A, SEG B, SEG C and SEG D are examples of data sets suitable for transfer in this format.

This format is already used for storing data on CD-ROM.

EXTENDED INDEX DATA FILE PAIR (.idx - .tape format):

The NDS Download Program

This format supplies the data set as two disk files: a data file and an index file. The data file is just like the byte stream file described previously, it is identified with a '.tape' file extension. Since this '.tape' file contains all the data blocks concatenated together there is no way to know where a block starts and how big it is, thus the need for the index file. This file is identified with a '.xidx' file extension and contains records that describe the location and size of blocks in the '.tape' data file.

Each block in the original data set, including end of file marks, has an index record to describe it. The first index record is used to indicate how many more index records there are in the '.xidx' file. To indicate an end of file mark, typically used after a header data block pair on original tape media, an index record is used where the byte size will be zero and the byte offset will be equal to the following block's byte offset.

To process the data set in this format, a program must access both the '.xidx' and the '.tape' file. The '.xidx' file is used to extract the blocks from the '.tape' file. If used correctly the '.xidx' file can be used to skip blocks and end of file marks in the '.tape' file.

This format is useful for any format that may have varying block sizes, and uses end of file marks for separating files within a data set.

Support software is required to extract the blocks correctly. This format is used extensively in house at NetDriven Solutions.

NDS ARCHIVED FILE (aka .d2 file format):

This format is used by NDS to archive data sets from various tape media to D2 tape. The format was created with the following considerations in mind:

- to be able to handle any data set recorded on various tape media
- to include identification information about the data set and
- to make efficient use of the D2 transport.

The format created consists of a byte stream file that encapsulates the identification information and the data sets in 'd2 blocks'. Each 'd2 block' consists of a four byte d2 block header and one or more 'segments'. Each segment consists of a segment header and an optional data segment. The 'segments' allow for a number of small 'blocks' to be encapsulated into one 'd2_block' as well as to allow for very large 'blocks' to span more than one 'd2_block'. Details of the encapsulated format are available for the asking from NetDriven Solutions Inc.

This format is useful for any format that may have been recorded on any tape media.

Support software is required to extract the blocks correctly.

Support software to access data blocks from this format by means of a linkable object library is available for SUN Solaris/Sparc 2.4 and SGI IRIX 5.3 and later.

The NDS Download Program

While some of our earlier archiving was done using this format, NDS no longer uses this as the primary archiving format.
