HDF5 for Archiving and Delivering Active/Passive Full Wavefield Data

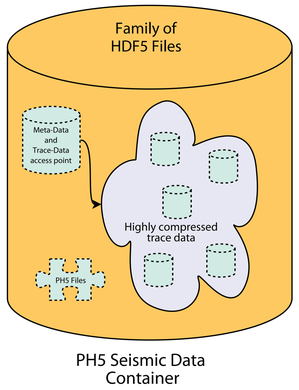
Steve Azevedo, Derick Hess, & Bruce Beaudoin
IRIS PASSCAL Instrument Center, New Mexico Tech
Socorro New Mexico, USA (bruce@passcal.nmt.edu)

ABSTRACT

Archiving and serving data from active/passive full-wavefield data sets presents several challenges. Active data is best suited to be stored and utilized in SEG-Y. Passive, continuously recorded data is most frequently stored in SEED or some similar format conducive to continuous data and capable of handling instrument response meta-data. Neither SEG-Y nor SEED are appropriate formats for storing both active and passive data. Surveys that combine both active and passive sources also frequently need to be served to the end users in multiple formats. To mitigate these and other problems with active/passive data sets, IRIS PASSCAL developed a storage format based on the HDF5 data format.

HDF5 was chosen for the following desirable characteristics: 1) Freely available and open source 2) Well supported by the HDF Group (part of NCSA) 3) Support for numerous data models including time series data 4) Self describing hierarchical data format 5) Large number of data tools already available 6) Elements can be accessed generally more quickly than if they were in an SQL database 7) Large number of programming APIs available 8) Easy to edit elements in-place 9) Ability to accommodate extremely large data sets. The PIC KITCHEN is used to organize all of the experiments data and meta-data into PASSCAL HDF5 format (ph5). The extensibility and portability of HDF5 allows the PH5 format to evolve and operate on a variety of platforms and interfaces. Storing data in HDF5 format also facilitates the interactive production of SEG-Y gathers, as well as other common seismic data formats based on varying and flexible data request parameters.

To make PH5 more flexible the seismic meta-data is separated from the time series data in order to achieve gains in performance as well as ease of use and to simplify user interaction. By separating the meta-data we are able to provide a better web based interface for data requests. Our web interface generates a unique web form and pre-populates much of it based on the meta-data provided to it from the PH5 file. The data requester then can intuitively select the extraction parameters as well as data subsets they wish to receive. The web interface then passes this on to the PH5 processing tools to generate the requested seismic data.



PH5 Seismic Data
Container

Photo credit T. Parker, G. Slad

Sweetwater

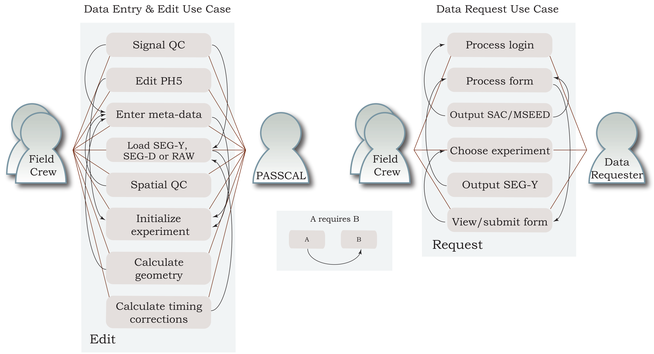
Though not technically a complete-wave field experiment, Sweetwater collected data using instrumentation typically used for both passive and active source experiments. This experiment fielded a large number of channels recording continuously resulting an a much larger data volume than IRIS is accustomed to handling. This data set gives us a good opportunity to test and refine our data handling tools and methods.

Nodal Seismic Array and Data Set

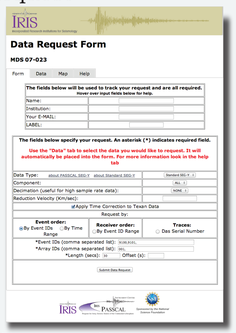
- Nodal Seismic Vibro-Seis survey recorded on Fairfield Nodal ZLand seismic systems. The survey ran from mid-March to April 30.
- The initial data release will consist of approximately 2600 nodes covering 12.5 sq. miles. The data volume is roughly 27 TB of 2ms, continuous data.

PASSCAL Seismic Array and Data Set

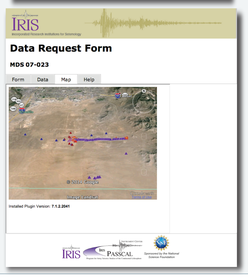
- 25 PASSCAL station array including 20 - Trillium Compact Post Hole sensors (20 sec), 5 - Trillium 120 Post Hole sensors, 200 sps and 20 sps, 92 GB stand-alone data collected, 98.7% data completeness, 5 telemetered stations, 200 sps (via cell modems), and 99.5% data completeness (real-time).
- Data available at IRIS/DMC under Network code XB.

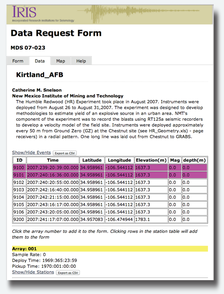


Request

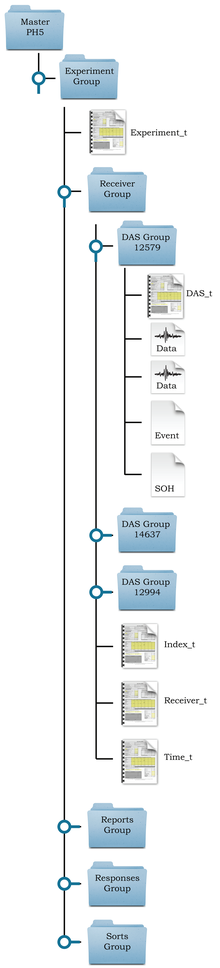


Connecting to
the Web Form





Selecting data for
export

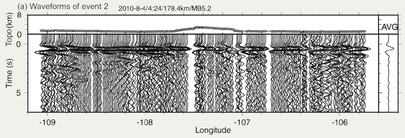


THE STRUCTURE OF PH5

- Top level **Experiment Group**: Shows the groups within the experiment group and the contents of the experiment table. The information in the experiment table gives general information about the experiment such as location, experiment name, and names of PI's and their institutions.
- **Receivers Group**: Shows the groups within the receivers group and the contents of the index table, the receiver table, and the time table. All of the DAS groups in the receivers group are external links. The index table is used to track these links. The receiver table gives the orientation of a sensor. The time table gives the time-drift of the clock for each data logger.
- **DAS Group**: The DAS group holds the trace data. The data arrays hold the time series as 32-bit integers, the event arrays hold data logger generated information about the recording windows as character arrays, the SOH arrays hold data logger generated state-of-health information as character arrays.
- **Reports Group**: Holds any PI generated reports.
- **Responses Group**: Shows the response group and the contents of the response table. Potentially complete response information can be held in character arrays holding evalresp style response information
- **Sorts Group**: This group holds information about events in the events table, information about the stations in the array table, information about shot-to-receiver distances in the offset table, and information that ties events to data traces in the sorts table.

Some Advantages of PH5

- Waveforms are stored separately from meta-data
  Simplifies archive updates
  More timely archiving
- Customizable data requests
  Unlike SEG-Y where the archiver determines data availability and gather configurations.
- Data is compressed
  Lossless data compression and de-compression is done in the HDF5 libraries and is completely transparent to the user. Compression is ~20% better than steim2 on the same data.
- All recorded data can be archived
  Below a teleseismic event recorded on RefTek RT125A instruments deployed after a controlled source experiment. By archiving all recorded data, sources of 'opportunity' can be later extracted from the archive.



Record section from: Zhaohui Yang et al. "Imaging Basin Structure with Teleseismic Virtual Source Reflection Profiles." Geophysical Research Letters 2012.