PH5 Data Processing In a Nutshell

PH5 is the recommended archiving format for controlled-source experiments and mixed-mode experiments. This abbreviated guide will walk you through creating the directories and adding data to the PH5 archive (steps 1-3), formatting and loading metadata (4-6), calculating derived tables and loading responses (7-11) and validating, viewing, and extracting data from the PH5 archive (12-14). **UNIX commands (bold print)** and any command line arguments are highlighted yellow. Input/output files are denoted by <*filename*>. To activate the PH5 environment, run **source activate** ph5 before starting to work on your PH5. To deactivate the PH5 environment and return to your normal python environment, run **source deactivate** ph5 or **conda deactivate** for newer versions of conda.
Additional documentation can be found on our website: https://www.passcal.nmt.edu/content/data-archiving/active-source

1. Create an organized directory structure for the data. Start by creating a main directory for the project. Once this is made, create subdirectories within it for the raw data (RAW), the metadata (METADATA), and the final PH5 formatted files (PH5). For example: **mkdir** RAW (to put raw files in).

2. Create a list of the files in the raw directory with their full path. From within the main project directory:

ls -d $PWD/RAW/*.TRD > trd-list.txt

This command will create a list of Texan raw files with their complete path—modify as necessary for RT130 (*.ZIP) and nodes (*.fcnt). If you have multiple types of instrumentation, combine your raw file lists into a single master list that contains all raw data including final node file names after running **unsimpleton** (if applicable; see 2a).

2a. If you are using nodes, after creating the raw node list fcnt-list.txt), you will have to use the **unsimpleton** command from the main project directory to create hard links to the raw node files and output them into a NODE_DATA directory:

unsimpleton -f fcnt_list.txt -d NODE_DATA

Once this is done, the final list of the renamed data can be made from within the main project directory:

ls -d $PWD/NODE_DATA/*.rg16 > node-list.txt

3. Open **pforma,** a program that will combine the raw files into a PH5 archive. At the bottom of the GUI window, select the file you created in step 2 for the raw file list (or your node list in 2a) and choose the PH5 directory you made in step 1 for the processing directory. If you are using nodes, set the UTM Zone (zone number plus hemisphere, N or S), Number of SEG-D Traces to Combine (~30), and Timeout in **pforma** (200000 sec) from File menu (see the long document for more details). Click 'Run,' and allow **pforma** to create the required files. Finally, once all processes are finished, go to the File menu and click "Merge." This will create the Sigma folder in the PH5 processing directory where the final PH5 will be located.

4. Kitchen exchange format files, or kefs, are used to import metadata into the PH5 file. Go to the METADATA directory. Build the experiment summary kef with the GUI: **experiment_t_gen**
You will need information such as the project name, PIs, and general location to complete this step. Afterwards, return to the Sigma directory and load the resulting kef (in this example it was saved as experiment_t.kef) into the PH5 archive:

keftoph5 -n master.ph5 -k ../../METADATA/experiment_t.kef

*note: Every time you load a kef using **keftoph5**, an entry is made in a keftoph5.log file. This can be useful for checking the progress of your steps!

5. Return to the METADATA directory to prepare the receive/station metadata in a spreadsheet. **If you are using Magseis Fairfield nodes, read the "For Magseis Fairfield nodes" section on the last page.**
Required columns are: Array, Station (id_s), Seed station name, DAS serial number, Sensor serial number, Channel number, Sample rate, Sample rate multiplier, Seed channel, Latitude, Longitude, Elevation, Deploy time, Pickup time, DAS manufacturer, DAS model,. Optional fields include Sensor manufacturer and Sensor model.
Receivers/stations in PH5 are organized into arrays. The station metadata is also referred to as the array metadata or the array tables. Please fully read the 'Metadata Tips' section on the last page of this document before proceeding further and see the receiver/station and event template files on the PASSCAL website for further information.

Once the spreadsheet for the receiver/station metadata is completed, save it as a .csv file.

Also, if applicable, create an event table as well. Required columns are: Array, Shot_ID (id_s), Latitude, Longitude, Elevation, Shot Time. While the Size and Depth of the event are not required, they are recommended if available. Save the event table as a .csv file.

Follow the steps below to build the array (and event if applicable) kef(s).

a. Open **noven**; under the File menu, select and open the receiver or shot csv file.
b. Under File > Configure, set as appropriate: "Input Type", "Column Separator" and "View Lines" (lines of the csv file visible in the GUI; set to 1 to view the csv headers).
c. In the main GUI window, assign the appropriate field name from each column drop down menu (check the Help menu for meanings of the different field names); repeat as necessary to ensure all required columns are defined with a field name. Under the File > Configure menu set "Skip Lines" to the number of header line(s) in the file.
d. Select "Check input" under the File menu; act on errors reported in pop-up.
e. Save the output, placing the suffix ".kef" at the end of the file name (e.g. array_1.kef).

6. Build a Google Earth KML file to view the experiment geometry via File > Map locations menu in **noven**. This file only covers the currently loaded receiver array or event geometry. Open the resulting KML file with Google Earth to review your receiver geometry. Check each receiver array and any events. If necessary, correct any errors in the csv file and rebuild the kef before proceeding.

7. Return to the Sigma directory and load all of your receiver/station and event (if applicable) kef(s):

```
keftoph5 -n master.ph5 -k ../../METADATA/array_1.kef
```

8. Update the response table references for multiple datalogger types. **Skip this step if you only have one type of datalogger.** In your PH5 directory containing the sub families and Sigma directory run:

```
set_n_i_response
```

This will create a new directory called RESPONSE_T_N_I. Move into this directory.

From the RESPONSE_T_N_I directory run:

```
load_das_t --path=../Sigma --onlysave
load_das_t --path=../Sigma --onlyload
```

Delete the old response table and load the new Response_t_cor.kef found in the RESPONSE_T_N_I directory.

```
delete_table -n ../Sigma/master.ph5 –R
keftoph5 -n ../Sigma/master.ph5 –k Response_t_cor.kef
```

9. To load responses into PH5, you must run **resp_load** twice. First, from the Sigma directory, run the command with a comma-separated list of arrays you want to load the responses for:

```
resp_load -n master.ph5 -a 1,2
```

This will generate a template csv file of the das and sensor combinations and save it as input.csv. Edit this template file with the path to the RESP files that you wish to use for the experiment. Most conventional responses exist on the Nominal Response Library page: http://ds.iris.edu/NRL/.

For example:
Template csv: *rt130,cmg3t,100,1,1,,*
Modified csv: *rt130,cmg3t,100,1,1,/path/to/RESP/RT130_100SPS.RESP ,/path/to/RESP/CMG3T.RESP*

Then, you must run **resp_load** again to create new response and array tables:

resp_load -n master.ph5 -a 1,2 -i *<input.csv>*

10. If you have a shot array table, in the Sigma directory calculate the source-to-receiver offset for each station and shot and load the resulting kef:

geo_kef_gen -n master.ph5 > offset_t.kef then run **keftoph5** -n master.ph5 -k offset_t.kef

11. Texan data may need timing-drift corrections; calculate as shown below in the Sigma directory. **Skip this step if you do not have Texans**. (RT130s and nodes are GPS-timed during data acquisition.)

time_kef_gen -n master.ph5 > time_t.kef then run **keftoph5** -n master.ph5 -k time_t.kef

12. Run **sort_kef_gen** in the Sigma directory, to produce a kef file containing information that optimizes data searches, and load the resulting kef:

sort_kef_gen -n master.ph5 -a > sort_t.kef then run **keftoph5** -n master.ph5 -k sort_t.kef

13. In the Sigma directory, examine your PH5 for errors using **ph5_validate** and **ph5tostationxml**. **ph5_validate** will output a text file of any warnings about your archive and can automatically fix some errors:

ph5_validate -n master.ph5

To look at the stationXML of your ph5 archive, you can use the **ph5tostationxml** command. ph5tostationxml generates stationXML format metadata at the station or response level. To generate stationXML run:

ph5tostationxml -n master.ph5 -o exp-sta.xml --level=station
ph5tostationxml -n master.ph5 -o exp-resp.xml --level=response

ph5tostationxml will also generate a Google Earth KML file for all arrays and shot lines:

ph5tostationxml -n master.ph5 -o exp-geo.kml -f KML

14. You can build SEG-Y shot gathers using the PH5 command **ph5toevt**:

ph5toevt -n master.ph5 --use_deploy_pickup -o Gathers -N -l 10 -A 1 -x U -e 5012 --shot_line 1

The SEG-Y shot gather in this example is written to the Gathers directory and contains 10 second long un-time-corrected traces from array 1 receivers (all components) starting at the time of event 5012. A log of **ph5toevt** activity is also saved in the specified output directory. If you are ready to produce gathers for all events, you can run **ph5toevt** with the '-E' flag to create gathers for all of the events. Run **ph5toevt** -h for more options. You can also build receiver gathers using **ph5torec**. See **ph5torec** -h for the available options.

15. Output SAC or miniSEED data. PH5 also supports writing data out as SAC or miniSEED. You can write out all of the data from the PH5 archive or just specified time windows using the **ph5toms** command. **ph5toms** defaults to miniSEED for the output format unless the -F flag is specified. Try **ph5toms** -h for all options.

To output miniSEED for two days of data for array 1, run the following:

ph5toms -n master.ph5 -p . -a 001 -s 2017:052:00:00:00 -t 2017:053:23:59:59 -o mseed -F MSEED

To output SAC files for all data contained in the PH5 archive, run the following:

ph5toms -n master.ph5 -F SAC -o sac_out/

Essential Tips on Preparing Metadata

1. Texan serial numbers in the metadata must have 10,000 added to their value for the csv file. So, Texan 2345 should be listed as 12345 and Texan 643 would be 10643. This reflects the internal serial number for the unit, which is represented as I2345 and I0643 in the file name.
2. SEG-Y station IDs must be numbers <= 65536; no letters, punctuation or special characters.
3. Consider using 4 or 5 digit station and shot IDs where the left most digit is the number of the line (array).
4. SEED station names are required. SEED station names must be 3 to 5 characters, alphanumeric, all capital letters.
5. Lines (Arrays) must be numbers. Consider starting the line values at 1 for receiver arrays. An array or line is a logical grouping of stations.
6. RT125s (Texans) and RT130s use the channel nomenclature of Z=1, N=2, and E=3. Nodes, however, use the nomenclature N=1, E=2, and Z=3.
7. When station sites are re-occupied (e.g. installed Texans are swapped with fresh Texans to allow continuous recording at a site) build a separate csv and receiver kef file for each deployment. In other words, no station and component combination should have more than one deployment and pickup cycle in any given csv or receiver kef. This makes it easier to replace metadata for an array if errors are discovered.
8. The columns of the csv file may be in different orders, but the format used in each field is required—be sure to use proper decimals. Deploy, pickup and shot times must be in the format YYYY:DOY:HH:MM:SS.ss.

How to Remove Tables

If any metadata in the csv files is found to be incorrect after loading the subsequent receiver/station or event kefs into the PH5 file, delete the associated tables containing the metadata or tables with calculations based upon the metadata with **delete_table**. Correct the errors in the csv file, rebuild appropriate station/event kef and reload the corrected kef into the PH5 file. Repopulated any other affected tables.

For example, if a receiver/station location is in error after loading the array_1.kef into the PH5 file:

A. Delete affected array tables and any related tables: **delete_table** -n master.ph5 -A 1 followed by **delete_table** -n master.ph5 -O a_e (removes the source-receiver offset information, where 'a' is the array number and 'e' is the shot line number). If all arrays are contained in one csv, you must remove all array tables.
B. Correct the station location in the csv file; rebuild the array_1.kef with **noven** (steps 4-5).
C. Load array kefs: (step 7).
D. Re-calculate the offsets by building a new offset kef and then load it (step 10).

For Magseis Fairfield nodes

If you are using Magseis Fairfield nodes, the array metadata is automatically generated by pforma using the header information in the raw Fairfield continuous data files. If all metadata/location information was correct when the data was written out from the node server, then you do not need to generate new arrays tables. You will still need to create any shot metadata following the instructions outlined in Step 5 of this document.

If the metadata/location information was *not* correct in the node server when you wrote out the data, then you will need to remove the automatically created array tables using the **delete_table** command and replace them with the correct array kef files created by using noven. See "How to Remove Tables" above and Step 5 of this document to remove the original array tables and create new array kef files with the correct information.

Loading Individual RAW files into PH5

If you have already built your PH5 archive using **pforma**, but need to add additional raw data you can do this through a command line performed in the Sigma folder:

125atoph5 -f *<file_list>* -n master.ph5 >& 125atoph5.out

The file list should contain the full path for the additional raw data files. Use **125atoph5** to load Texan raw data, **130toph5** for RT130 data, and **segdtoph5** for nodes.

NOTE: If the orientation of the receivers/stations does not align along true north or east, there is a section in the PH5 Long Document regarding how to account for this.

Last revised: April 16, 2020