OBS data/metadata preparation recommendations

Wayne Crawford

Version: 20230619

Version	Modifications
20230619	Based on OBS data proposal v202203

Preamble

Ocean bottom seismometers (OBS) provide seismological access to the 70% of the earth's surface that is covered by water. For full usage, they should be archived in and distributed from FDSN compatible seismological databases.

The following is a list OBS-specific standards and "best practices" for using the stationXML and miniSEED formats to make OBS data clear, easy to use and intercompatible.

OBS clocks generally have a non-negligible drift because of the lack of GPS signal at the seafloor. The resulting time offsets must be corrected or at least indicated in any data archived at data centers. OBS time bases are generally chosen to have small and first-degree linear drift. Their drift is calculated by synchronizing the instrument clock to GPS before the deployment and then comparing the instrument clock to GPS after the deployment. If the instrument clock cannot be compared to GPS at the end of the experiment, the drift can be calculated a posteriori by calculating the noise correlation between this instrument and another synchronized instrument over the length of the experiment.

Information about the existence of ~~linear~~ clock drift, its value if measured (linear at least, non-linear if possible) and its probable range if not measured, should be provided in the data and metadata.

obsinfo software package

The Pure Python obsinfo software, available at gitlab or using the "pip" command-line interface, creates StationXML files including OBS-specific information. A full StationXML file requires instrument responses in obsinfo-compatible format, but a relatively simple "subnetwork" file is sufficient to generate StationXML-compatible fields, which can then be inserted into StationXML files (a future version might automatically inject them).

Obsinfo can also be used to create standard (RESP or AROL) instrument files from basic component information.

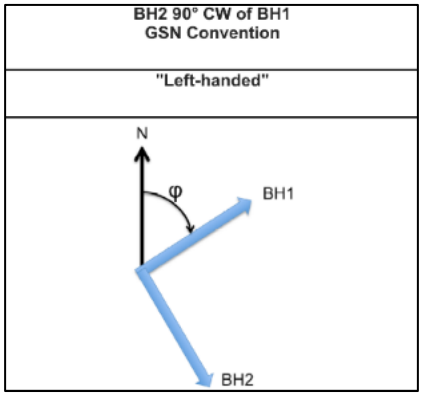
Provenance files

Processing done on data files (from data download to delivery to the data center) should be recorded in text-based, structured files. The JSON process-steps format is an example.

Source Identifiers

The following source-subsource codes (see FDSN Source Identifiers documentation)should be used for the following types of sensor/data:

Unoriented seismometer: “1” for the “N” channel, “2” for the “E” channel (GSN standard, see image below)
Seismometer/geophone with inverted vertical channel (positive voltage is down): “3”
Hydrophone: “DH”
Differential pressure gauge: “DG”
“Absolute” bottom pressure recorder: “DO”



Station names for repeated deployments

If OBSs are deployed repeatedly at one site (to make a long series), use an incrementing alphanumeric character at the end of the station name (i.e., A01A, then A01B then A01C for subsequent deployments at the same approximate location).

Metadata (StationXML)

See StationXML Reference for details of StationXML elements

Clock drift

Indicate in a <Comment>, using JSON syntax, for example (using obsinfo):

```
<Comment subject=”Clock Drift”>
    <Value>“{Linear Clock Correction: {time_base: Seascan MCXO, ~1e-8 nominal
drift, reference: GPS, start_sync_reference: 2015-04-
22T09:21:00Z,start_sync_instrument: 0, end_sync_reference: 2016-05-
28T22:59:00.1843Z,end_sync_instrument: 2016-05-28T22:59:02Z}}”</Value>
</Comment>
```

Use absolute datetimes when possible.

If you must use relative times, use (and name) “instrument _correction”: Seconds to add to instrument time at a given reference time (this value has the same sign as the “Instrument Correction” record header field in miniSEED 2.4 and the FDSN -> Time -> Correction header in miniSEED 3).

Deployments in lakes

Set the <WaterLevel> to the elevation of the lake surface

Positions

Use the “plusError”, “minusError” and “measurementUncertainty” attributes to specify uncertainties in Latitude, Longitude and Elevation and how you measured them.

Leap seconds

Indicate in a <Comment>, using JSON syntax (for example, using obsinfo).

```
<Comment subject=”Leap Second”>
    <Value>“{time: 2016-082T23:59:60Z, “description: Positive leap-second (a
61-second minute)}"</Value>
</Comment>
<Comment subject=”Leap Second”>
    <Value>“{correction_data: msmod --timeshift -1 -ts
2016,182,23:59:59.999999 –s, correction_end_sync_instrument: subtracted one
second from displayed instrument time"</Value>
</Comment>
```

Orientation information

Set the following <Azimuth> and <Dip> values for source/subsource codes 1, 2, 3, DH, DG and DO:

1	<Dip unit="DEGREES">0.0 </Azimuth><Azimuth minusError="180.0" plusError="180.0" unit="DEGREES">0.0</Azimuth>
2	<Dip unit="DEGREES">0.0 </Azimuth><Azimuth minusError="180.0" plusError="180.0" unit="DEGREES">90.0</Azimuth>
3	<Dip unit="DEGREES">90.0 <Azimuth unit="DEGREES">0.0</Azimuth>
DH, DG, DO[1]	<Azimuth unit="DEGREES">0.0</Azimuth> *If value DECREASES for a positive pressure :* <Dip unit="DEGREES">90.0</Dip> *If value INCREASES for a positive pressure :* <Dip unit="DEGREES">-90.0</Dip>

The “3” code is used for vertical channels where an increase in value corresponds to a DOWNWARD movement. If an increase in value corresponds to an UPWARD movement, (the seismological standard), use the standard “Z” code and Dip = -90

Data completeness

[1] Pressure sensor dip corresponds to “Z/3” conventions for UPGOING waves

Use Station <CreationDate> and <TerminationDate> fields to specify when the data was supposed to start and end, and <StartDate> and <EndDate> to specify when it actually starts and ends.

Standard values marine seismologists may not know:

Within each <Channel>, set `<Type>CONTINUOUS</Type>` and `<Type>GEOPHYSICAL</Type>`

Data (miniSEED)

Clock drift correction

Three main possibilities for distributing data are proposed:

1. Indicate the time correction in each record header but do not apply it (RAW).
2. Indicate the time correction in each record header and apply it (SHIFTED).
3. Resample the data at the originally intended rate (RESAMPLED)

The SHIFTED method as it allows the user to work with time-corrected data which has not been modified but for which the time is as close as possible to GPS time. Users of very long-period data often prefer RAW data because it is easier to concatenate daily files. RESAMPLED data offers the best of both worlds, but could distort waveforms/spectra (only if not correctly performed?).

Until consensus is reached, we propose below how to distinguish between these methods.

If the time correction has been calculated:

- RESAMPLED data: Use a non-standard Instrument Code, as the data themselves have been modified.
- SHIFTED data:
    - Indicate time correction applied in record header field 16 ("Time Correction" and set field 12, bit 1 ("Activity flag, time correction applied") to 1. The 'qedit' software does all of these at once to each header ("add_trend corr" and "apply_corr keep" commands).
    - Indicate that the time correction code has been applied by:
        - Setting the data quality flag to "Q"
        - Alternatively, specify a location code between 00 and 49
- RAW data.
    - Indicate time correction applied in record header field 16, without applying it. The 'qedit' software can do this using its "add_trend corr" command.
    - Indicate that there is no time correction by:
        - Setting the data quality flag to "D
        - Alternatively, specify a location code between 50 and 99

If the time correction has not been calculated

Set bit 7 of the data quality flag ("time tag is questionable") to 1.

?If possible, add blockette 500, field 10 ("Clock status") indicating the linear drift (i.e. "Unmeasured linear drift on Seascan MCXO, expected order(1e-8)")?

Leap seconds

Leap seconds should be corrected in the data and the record containing the leap second should be flagged.

If the leap second is positive (the most common case: 61 seconds in the minute):

- Shift all record times AFTER the leap second back one second.
- Set activity flag bit 4 to 1 in the header of the record containing the leap second.
- Change 'end_sync_instrument' to be one second earlier than what the instrument indicated

If the leap second is negative (59 seconds in the minute):

- Shift all record times AFTER the leap second forward one second.
- Set activity flag bit 5 to 1 in the header of the record containing the leap second.
- Change 'end_sync_instrument' to be one second later than what the instrument indicated

Here is how to do this using msmod, assuming a positive leap-second at 23:59:60 on day 182, 2016:

```
msmod --timeshift -1 -ts 2016,182,23:59:59.999999'
msmod –actflags '4,1' –ts 2016,182,23:59:36 –te
  2016,183,00:00:36
```

The times in the second command are hand-chosen to bracket the record containing the leap second.

Proposed modifications

Allow sampling rate to be specified as double precision. (in miniSEED3 draft)

This is the only way to accurately represent OBS clock rates, which are regular but off of the specified sampling rate by a factor of approximately 1e-8 (MCXOs) or 1e-9.5 (CSACs), requiring 27- or 32-bit floating-point mantissas, respectively, to be correctly specified. Single precision floats only have 23-bit mantissas, double precision floats have 52-bit mantissas.

More data quality flags, with clear hierarchy

Data quality flags are the only clear way to distinguish between levels of data processing, but the choices are too limited. Additional data qualities that cannot currently be specified are: Data directly translated from another format , or data for which the header values have been changed, but not the data itself. A possible hierarchy would be (new in italics):

- "D" : The state of quality control of the data is Indeterminate
- *"T": Translated Raw Waveform Data from another initial format*
- "R": Raw Waveform Data with no Quality Control (reserved for SEEDlink)
- *"H": Quality controlled Data, processes have been applied only to the headers*
- "Q": Quality controlled Data, some processes have been applied to the data *(does this mean time-series values)?*
- *"C": Quality controlled Data, No processes applied to time-series or header*
- "M": Data center modified, time-series values have not been changed

In the miniSEED3 header, the Data publication version replaces the data quality flags (can still have flags in “Extra Header Fields” (field 14). This offers a clear hierarchy, but not a way to specify that one wants uncorrected or corrected data (recommended RAW=1 could be used for uncorrected data). Could this be put in field 14: extra header fields? In any case, would have to be searchable using web tools