Straw man OBS data/metadata proposal

Wayne Crawford

Version: 20161117

Preamble

Ocean bottom seismometers (OBS) greatly improve the range and coverage of possible seismological studies by providing access to the 70% of the earth's surface that is covered by water. Studies using OBS data have been cited many times and OBS data is starting to appear in standard seismological databases

Two major blocks in getting OBS data to the public remain: 1) the relative lack of data and metadata in data centers; and 2) differences or inconsistencies in data and/or metadata formats used for OBS data. The distribution of data by OBS facilities is hampered by a lack of clear standards and procedures for archiving data and metadata.

I propose here standards for OBS data and metadata formats, as well as post-processing tools that should allow OBS data to be more available and useful to the general public. This is a "strawman" proposal, in which I present my ideas in an initial draft. Anyone reading this is invited to supplement or suggest corrections or deletions to the proposed standards. The near-term goal is to establish a reference document that will provide clear and simple instructions for preparing OBS data, serve as a reference for those using OBS data, and establish guidelines for future development in OBS instruments and OBS data processing.

The document is divided into four sections: Data, Metadata, Software and User Manual. In each of the first three sections, I outline problems, propose solutions, and list alternative solutions. Seismological standards are used wherever possible, with additions made only where necessary for OBS data. In the last section, I outline a first-draft user-manual for OBS data so that new users are made aware of potential differences with respect to land data.

Many standards used for storing seismological data have incomplete documentation. It would be useful to outline them at the beginning of a document for OBS facilities. They include, but are not limited to, the following:

- Standard tools used to prepare/process miniSEED, dataless SEED, and StationXML files
- The specification of data types "Continuous" and "Geophysical" in the metadata
- Preferred data formats for miniSEED (I have received numerous recommendations to use "Steim1" rather than "Steim2", for example).
- The fact that most software that compresses existing miniSEED data does not optimize the length of record sections, but instead stuffs as many compressed data record sections as possible into an existing record section, keeping that sections' header. This is not necessarily bad because it avoids inventing or extrapolating a new section header, which could be problematic for time-corrected data. But it should be mentioned and the reason explained.

Data

OBS data should be delivered to the data centers in the standard format for the data center. I assume here that that format is miniSEED, which is generally used by data centers. Many other formats (including SAC) do not allow the fine-scale time correction specified below.

Timing corrections

OBS clocks generally have a non-negligible drift because of the lack of GPS signal at the seafloor. This may change in the future with atomic clocks, but for current and past data this must be corrected before it is archived at the data centers. Some parks provide daily files with a corrected start time but no further drift correction applied to the data itself. This can cause problems for people wanting to "stitch" time series together, as the drift over one day may be high compared to the sampling interval.

Recommendation

Apply the time correction to each miniSEED record header (generally every 4096 bytes, so every 40 seconds for efficiently compressed 24-bit data at 100 sps), specify the correction applied and the fact that it was applied in each header. Also, set the "data quality" field to "**Q**". If there is no information about clock drift but there is a risk of clock drift, set the "data quality field to "**D**" and set data quality flag bit 7 to "1".

The "fixed section of data header" fields concerned are:

Field	Bytes	Name	Comments
2	7	Data header/quality	"Q" for clock drift corrected, "M" for possible clock drift but not corrected
8	21-30	Record start time	Clock drift corrected time
12	37	Activity flags	Set bit 1 to 1 if corrected
14		Data quality flags	Set bit 7 to 1 if NOT corrected
16	41-44	Time correction	.00001 seconds that were ADDED to the initial time to obtain the corrected time

The 'qedit' software can apply a linear clock correction to miniSEED data, taking care of Fields 8, 12 and 16. I have heard some concerns that 'qedit' is not a standard miniSEED tool but I have not had any problems and I am not aware of standard software that can do a linear clock correction. There may be others that I am not aware of. The 'msmod' software can be used to set the data header/quality field and the data quality flag.

Alternatives

- Put the time correction in the header but do NOT apply it. This would avoid all problems with stitching data sets together, but could lead to errors in pick times made by users unaware of this convention or lacking tools to apply the corrections.
- ALSO specify the exact sampling rate in a "high precision sampling rate" blockette using a 64-bit floating-point representation[1].

[1] The standard sampling rate blockette (B100) is inadequate because it uses 32-bit floating point, which only is accurate to 7 significant digits, whereas common MCXO drifts are on the order of 1e-8, whose effect

- Resample the data at the specified sample rate. This seems too risky in terms of waveform distortion.
- Define a new field or flag that specifies that there is an uncorrected linear clock drift in the data. This would allow users to still use the data and then calculate the clock drift based on changes over time in, for example: time residuals in earthquake locations, or instrument noise cross-correlations. This field could be a new "io and clock" flag, for example (only 6 of 8 are currently defined)

Pressure instrument names

Currently it appears that all OBS pressure channels are named ?DH, which is specified in the SEED manual as meaning "hydrophone". Should differential pressure gauge channels be named "?DF" (infrasound) and absolute pressure gauge channels be named "?TZ" (tide instrument)?

The instrument code "D" is used for pressure data, however the "H" orientation code should be used for hydrophones and "F" for infrasound according to the SEED Manual version 2.4.

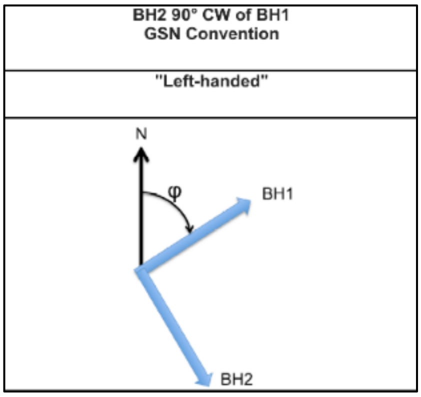
Channel naming

Short period seismic instrument channels at OBSIP are named EL? for high-rate and LL? for low-rate data. "E" for band type is Extremely Short Period, or sampling rates between 80 to 250 samples per second. "L" for band type is Long Period, or sampling rate of 1 sample per second. The "L" for the sensor type indicates a Low Gain Seismometer. These are both defined in the SEED Manual version 2.4.

Channel Orientation Codes

If the instrument was not oriented along the tradition axes (N-S and E-W), the orientation codes for the horizontal channels shall be "1" and "2" according to the geometry specified by the GSN standard shown below. This corresponds to 'N'->1 and 'E'->2 if the seismometer provides N and E outputs, but if the seismometer provides 'X' and 'Y' outputs there is no sure mapping: if the instruments' 'X-Y-Z' mapping is right handed than Y->1 and X->2, but if the mapping is left-handed than Y->2 and X->1.

The preferred standard would be "Left-Handed" as utilized by the GSN:



is seen on the eighth and higher digits. IEEE 64-bit floating point provides 14 significant digits, adequate for drifts down to ~3e-14 (1 μsec/year).

Channel Polarity

All data should be delivered with GSN-standard polarity (including positive Z = -90º dip from horizontal). It will not be acceptable to change the orientation specification in the metadata to a non-standard value in order to accommodate reversed-polarity data.

Site names for repeated deployments

If OBSs are deployed repeatedly at one site (to make a long series), should we give them the same name but different locations? Up to what point can we keep the same name? Does it depend on the array spacing?
According to FDSN(?), any sites within 1 km can be given the same station name, with the different deployments given different location codes, each one with it's precisely-specified position. IRIS, however, prefers to use an incrementing alphanumeric character at the end of the station name for OBS deployments, to indicate subsequent deployments.

Processing information

OBS data may go through a number of steps before being ready for archival at data centers. These processing steps should be well documented, so that any mistakes can be traced and corrected. The most obvious example is for the timing corrections, but other steps may also be useful. One possibility is to create 'opaque' miniSEED files with this information. Another would be to provide a text file (perhaps structured, such as JSON) with this information. The text or structured file would be more readable, whereas the opaque miniSEED file fits in some data structures (such as SeisComp3 data structure). I think a text file would be easiest for the user to read, but this depends on the data centers having some standard place for these text files (with State of Health data?)

Metadata

Metadata should be in dataless SEED or StationXML format, which are the most complete formats used in seismology and from which it is easy to extract simpler formats (such as RESP and SACPZ). In addition to the standard information for all seismological data, the following should be added:

- Water level: this is useful for removing/exploiting water surface reflections. In general, this is 0 (sea level), but not if deployments are made in lakes. It seems "water level" would be simpler to implement than "water depth" because the default value would be "0" rather than "-elevation".
- A comment on the method used to determine position as well as estimated position errors (the latter using the existing "plusError" and "minusError" attributes in the Latitude, Longitude and Elevation elements). Possibilities include:
  - Surface release position
  - Near-bottom release position using short acoustic baseline
  - Airgun survey
  - Acoustic survey
- If the horizontal channels are not geographically oriented, their "Azimuth" field should be set to "0"

A way to specify data recovery/loss

It would be very useful to have some way to make metadata (or something) for stations that did NOT return data. This metadata would include a comment (or field) specifying why there is no data (such as "not recovered", "logger failure", etc. This would be useful for

- Users expecting data from a location to see/understand why that site did not return data
- Automatically creating maps of station distribution, with non-working stations indicated by a different symbol or color
- Generating statistics on data recovery rates and common failure modes

For the same reasons, it would also be useful if stations that worked for some part of the experiment specified their initially planned recording period. I don't see a way of specifying this in StationXML

Software

Standardizing OBS data and metadata storage should also allow efficient validation and correction of OBS-specific problems. Software for at least the following examples should be written

Clock drift confirmation

Software to calculate clock drift based on drift in hypocenter travel time residuals
Software to calculate clock drift based on noise correlation between instruments

Noise removal

Removal of noise caused by infragravity waves (based on pressure measurements) and dynamic tilt (based on correlation with horizontal channel noise).

Sensor orientation calculation/correction

Determining the horizontal orientation of the instrument using teleseisms and/or local earthquakes. A common method used is the Rayleigh wave polarization method, outlined in the Stachnik et al., 2012 paper with code made available on the OBSIP website (http://www.obsip.org/data/obs-horizontal-orientation/).

Stachnik, J.C., A.F. Sheehan, D.W. Zietlow, Z. Yang, J. Collins, and A. Ferris (2012), Determination of New Zealand Ocean Bottom Seismometer Orientation via Rayleigh-Wave Polarization: Seismol. Res. Lett., 83, 704–712, doi:10.1785/0220110128.

Common pipeline for data preparation

It would be efficient and may avoid incompatibilities if a standard data preparation software was written which input uncorrected miniSEED data and integrated the clock drift correction and any other "standard" processes. The OBS facilities would still be responsible for converting from their proprietary format to miniSEED (with appropriate SEED-based channel names) and to provide clock drift measurements. This would provide the additional advantage that any clock drift corrections found afterwards could be applied almost automatically at the data center level

Active seismic data extractor

A program to input continuous data-center-level data and a shot file (in some standardized format) and output SEG-Y files.

Advantages: Reduce work for OBS facilities (both in writing extraction software and in re-extracting data once clock drift or shot corrections are found). Facilitate data sharing (access on-line at a secured site). Put more OBS data on data-centers (some non-shot parts may be useful for earthquake seismology). Securely archive this data.

Disadvantages: Puts a lot of "non-useful" data on data centers.

Note: I know of no current standard for storing/distributing active seismic data/metadata, although OBSIP now archives all active source data in continuous miniSEED and cut SEGY files. The current data standard is universal (SEGY), but not the metadata (neither for instrument responses, nor for shots), although a shift to PH5 (an implementation of HD5) could provide this.

User manual

This section presents the information that should be provided to the end user, so that OBS data usage is clear.

OBS-specific information

OBS data have some differences from conventional land data. The main differences are:

- The OBS clock may drift relative to absolute time, because there is no GPS signal at the seafloor. OBSs are synchronized at the beginning of each experiment and use precise clocks to minimize this problem. The drift is generally corrected for, but an incorrect correction may occasionally be applied. The correction is usually a linear correction, though drift has been shown to be non-linear.
- Most OBSs are deployed in free-fall and their horizontal axes are not aligned to geographic cardinal directions. Generally, their orientation is unknown, as magnetic compasses and gyrocompasses generally give unsatisfactory results.
- At low frequencies (<~0.1 Hz) there may be significant noise on the seismometer channels from ocean surface waves and instrument tilting under currents. There are algorithms to remove the ocean wave and current noise from the vertical seismometer channel using recorded pressure data.
- If an experiment is deployed close to naval activities, sometimes the OBS data is redacted or filtered to remove sensitive information.

There are algorithms available to evaluate and reduce the problems caused by these differences, but they are not trivial to most new users.

OBS clock drift

If the instrument is corrected for the clock drift, the miniSEED header Data Quality Code will be ‘Q’, otherwise it will be ‘M’. The correction is calculated for and applied to every miniSEED record header and indicated in record header fields 8, 12 and 16.

Channel names

Channel Name	Instrument	Sampling Rate	Gain	Notes
BH?	Broad/wideband seismometer	10-80 sps	High	
HH?	Broad/wideband seismometer	80-250 sps	High	
EL?	Geophone	80-250 sps	Low	
LL?	Geophone	1 sps	Low	
BX? HX? EX?	Varies	Varies	Varies	Filtered or otherwise altered data
BDH	Hydrophone	10-80 sps		
HDH	Hydrophone	80-250 sps		
LDH (or LDF?)	Differential pressure gauge	1 sps		
BDH (or BDF?)	Differential pressure gauge	10-80 sps		
HDH (or HDF?)	Differential pressure gauge	80-250 sps		

HDH (or HTZ?)	Absolute pressure gauge	80-250 sps		
BN? LN? BY?	Strong Motion	Varies		
VKI, LKI, HKO	Temperature	Varies		

OBS-specific processing codes

Should be able to work on data-center supplied data:

- Instrument orientation
- Timing validation/correction (using noise and/or EQ residuals)
- Noise removal
- SEG-Y file creation (from standardized shot files, can also transform from zero-phase to minimum phase if necessary)