Proposal for FSDN standardization of waveform quality metrics

V1.4 25 Feb 2016

Knowledge of quality of seismic waveform data and related metadata is essential for any scientific analysis and interpretation of the data. Automated processes to calculate data quality parameters are required (a) to handle huge amounts of data, (b) to enable data centres to automatically monitor changes or variations in data quality over different time scales and (c) to provide services to the research community to search for and harvest the best quality data based on these parameters.

Currently, parallel developments are on-going (e.g. IRIS DMC, ORFEUS EIDA) to calculate data quality parameters and use these in different services. This document proposes the standardization of a number of basic metrics of which most are common in both systems. The metrics are written for mini-SEED data.

Quality parameters are calculated for a time windowed time series. In the following a time series is considered to belong to a data stream uniquely identified by a SEED network code, stations code, channel code, location code and data header/quality indicator (D|R|Q|M). Default time window length is 24 hours, starting at 00:00:00.0 UTC.

Definitions

Data (dis)continuity, gap and overlap

A sample value at time t represents a continuous signal within time window [t, t+Δt), in which Δt is the sample interval between two samples as defined by the sample rate factor and the sample rate multiplier in the data record header (fields 10 and 11).

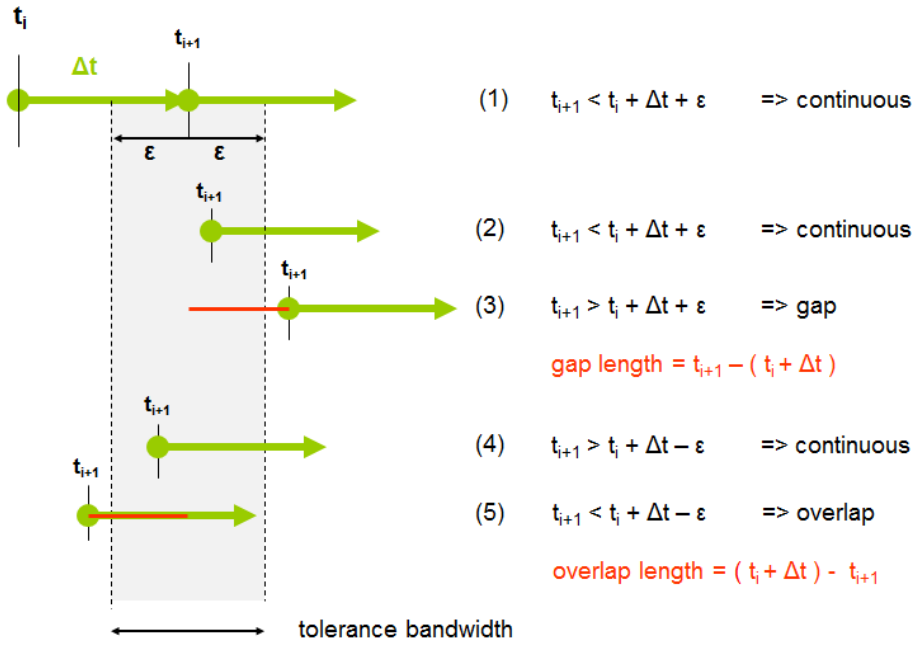
A continuous (discrete) time series is defined as a time series in which the time interval between two adjacent samples (a) is constant (Δt) or (b) does not differ from this constant by more than a certain time tolerance (**ε**). The tolerance value is proposed here as 50% of the sampling rate.

Continuity condition between two samples: $\Delta t - \varepsilon <= t_{i+1} - t_i <= \Delta t + \varepsilon$.

A *discontinuity* thus occurs when the time interval between two adjacent samples (with corresponding times t_i and t_{i+1}) exceeds the sample rate interval (Δt) by more than the (sample rate dependent) time tolerance ε: $|t_{i+1} - (t_i + \Delta t)| > \varepsilon$

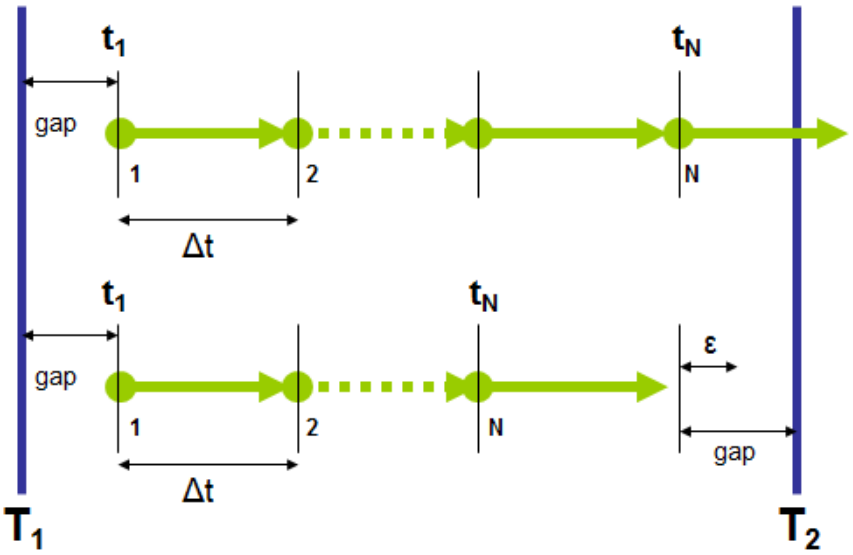
A *gap* is a positive valued discontinuity, an *overlap* is a negative valued discontinuity.

Gap condition:	$t_{i+1} - t_i > \Delta t + \varepsilon$	with gap length:	$t_{i+1} - t_i - \Delta t$
Overlap condition:	$t_{i+1} - t_i < \Delta t - \varepsilon$	with overlap length:	$t_i + \Delta t - t_{i+1}$



Within a time window [T_1, T_2) a time series may have a start time t_1 (defined by the first sample in [T_1, T_2)) later than T_1 or an end time t_N (defined by the last sample) before T_2. These start and end gaps are defined as:

start gap:	$t_1 - T_1$	when	$t_1 - T_1 > 0$
end gap:	$T_2 - (t_N + \Delta t)$	when	$T_2 - t_N > \Delta t + \varepsilon$



Record start time, end time and length

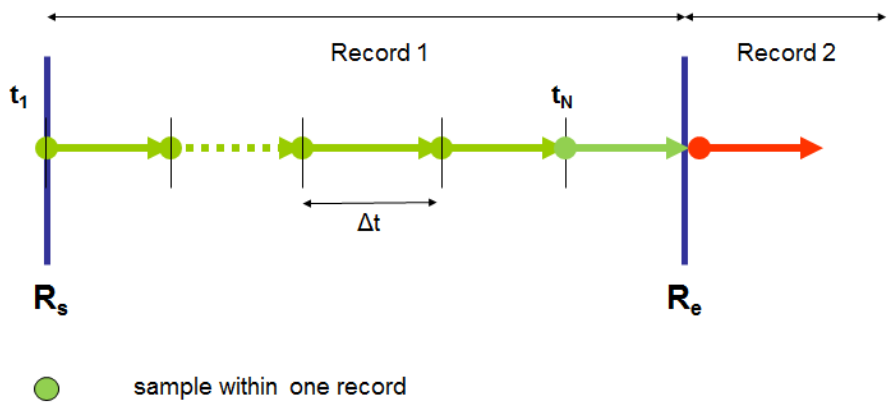
Data in SEED format is stored in records. Each SEED record contains a continuous time series. Gaps or overlaps in data may occur between records.

The **start time** $\mathbf{R_s}$ of a mini-SEED record is the time t_1 of the first sample in the record (including time correction if applicable).

The **end time** $\mathbf{R_e}$ of a record is defined by the end of the time interval represented by the last sample in the record (including time correction if applicable).

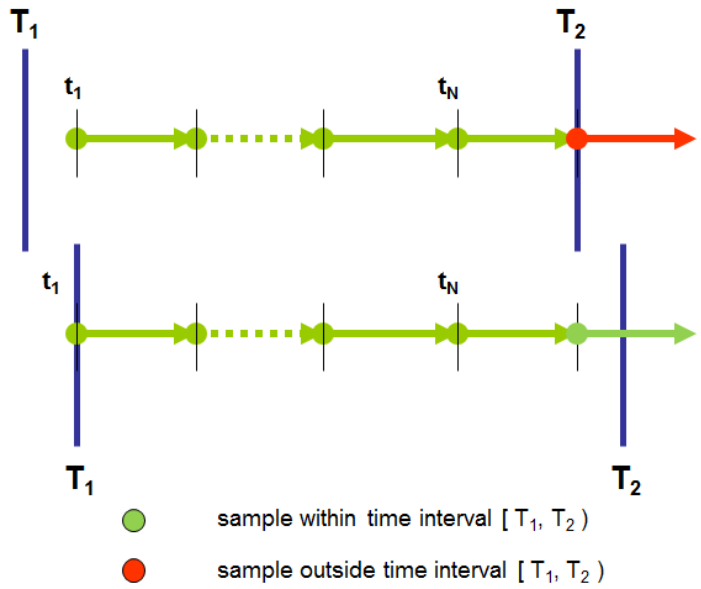
The time interval of a record containing N samples is thus defined as $[t_1, t_N + \Delta t)$.

The **length** of a record is $\mathbf{R_e} - \mathbf{R_s}$.



Metrics calculation in time window

Metrics are calculated within a time window [T_1, T_2). Time T_1 is included, time T_2 is excluded.



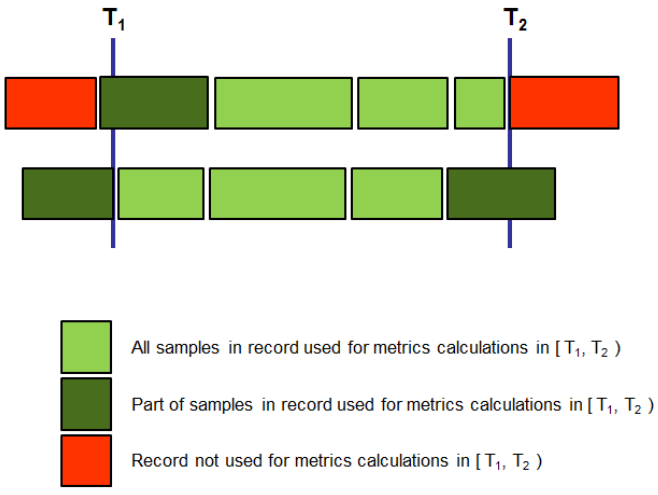
Samples at t=T_1 are included in the metrics calculations, samples at t=T_2 are excluded.

Data records for which $\mathbf{R_e < T_1}$ are not used in the metrics calculations.

Data records for which $\mathbf{R_s \geq T_2}$ are not used in the metrics calculations.

Data records for which $\mathbf{T_1 \leq R_e < T_2}$ are (partially) used in the metrics calculations. In these records the samples are used with time **t** for which $\mathbf{t \geq T_1}$.

Data records for which $\mathbf{T_1 \leq R_s < T_2}$ are (partially) used in the metrics calculations. In these records the samples are used with time **t** for which $\mathbf{t < T_2}$.



Data availability

Data availability is the percentage of data in a time window [T_1, T_2). It is the length of the time window minus the sum of all gaps in this time window (including start gap and end gap), relative to the length of the time window [T_1, T_2).

$$availability = 100 \times \frac{(T_2 - T_1) - sum\_gaps}{T_2 - T_1}$$

Metrics based on sample values

The following metrics are calculated from all sample values within time window $[T_1, T_2)$. Samples at $t=T_1$ are included, samples at $t=T_2$ are excluded.

- **sample_mean**

Average value of all samples $(x_1, .., x_N)$.

$$mean = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- **sample_max**

Maximum value of all samples $(x_1, .., x_N)$.

- **sample_min**

Minimum value of all samples $(x_1, .., x_N)$.

- **sample_median**

Median value of all samples $(x_1, .., x_N)$. The middle value of the sorted samples.
The 50-th percentile of all samples $(x_1, .., x_N)$.

- **sample_upper_quartile**

The 75-th percentile of all samples $(x_1, .., x_N)$.

- **sample_lower quartile**

The 25-th percentile of all samples $(x_1, .., x_N)$.

- **sample_rms**

The root mean square of all samples $(x_1, .., x_N)$.

$$rms = \sqrt{\sum_{i=1}^{N} \frac{x_i^{\,2}}{N}}$$

- **sample_stdev**

The standard deviation of all samples $(x_1, .., x_N)$.

$$stdev = \sqrt{\sum_{i=1}^{N} \frac{(x_i - mean)^2}{N}}$$

- **percent_availability**

Data availability is the percentage of data available in a time window [T_1, T_2). It is the length of the time window minus the sum of all gaps in this time window, relative to the length of the time window [T_1, T_2).

$$availability = 100 \times \frac{(T_2 - T_1) - sum\_gaps}{T_2 - T_1}$$

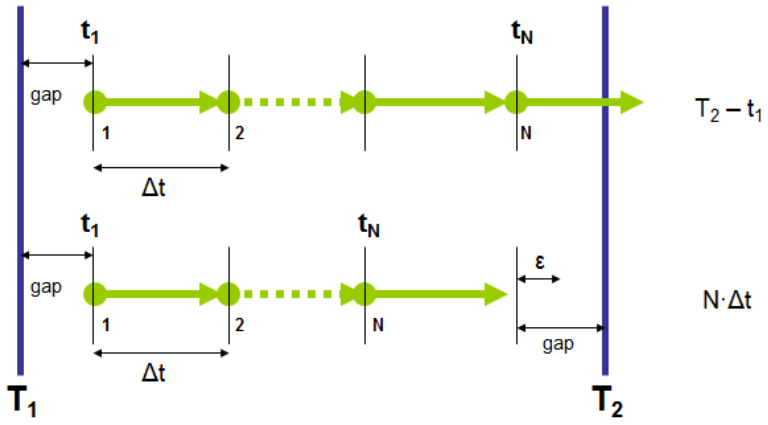
availability in seconds: $(T_2\text{-}T_1) - sum\_gaps$

Metrics based on mini-SEED data record header

The following metrics are extracted from header flags in data records fitting time window **[T_1, T_2).**

The metrics are based on flags that are specifically defined in SEED. To distinguish these from metrics that are calculated from the sample values (time series) the names are pre-pended by 'ms_'. As the time window [T_1, T_2) usually does not correspond exactly with record start time R_s and/or record end time R_e only the samples fitting the time window [T_1, T_2) – as defined above – are considered.

A flag in the record header of a SEED record applies to all samples in the data record.



The length of continuous data in a time window [T_1, T_2) is defined as:

$T_2 - t_1$	when $t_N + \Delta t > T_2$
$N \cdot \Delta t$	when $t_N + \Delta t + \varepsilon < T_2$

t_1 is time of first sample in [T_1, T_2), t_N is time of last sampe in [T_1, T_2)

- **ms__data_quality_flags__bit_0__amplifier_saturation**

Percentage of data in time window [T_1, T_2) for which bit 0 in the Data Quality Flag byte is set to '1'.

Calculation: sum the data availability (in seconds) of all continuous time segments that fit time window [T_1, T_2) in records for which bit 0 in the Data Quality Flag byte is set to '1' and (b), divided by the time window length $T_2 - T_1$, times 100.

- **ms__data_quality_flags__bit_1__digitizer_clipping**

Percentage of data in time window [T_1, T_2) for which bit 1 in the Data Quality Flag byte is set to '1'.

- **ms__data_quality_flags__bit_2__spikes**

Percentage of data in time window [T_1, T_2) for which bit 2 in the Data Quality Flag byte is set to '1'.

- **ms__data_quality_flags__bit_3__glitches**

Percentage of data in time window [T_1, T_2) for which bit 3 in the Data Quality Flag byte is set to '1'.

- **ms__data_quality_flags__bit_4__missing_padded_data**

Percentage of data in time window [T_1, T_2) for which bit 4 in the Data Quality Flag byte is set to '1'.

- **ms__data_quality_flags__bit_5__telemetry_sync_error**

Percentage of data in time window [T_1, T_2) for which bit 5 in the Data Quality Flag byte is set to '1'.

- **ms__data_quality_flags__bit_6__digital_filter_charging**

Percentage of data in time window [T_1, T_2) for which bit 6 in the Data Quality Flag byte is set to '1'.

- **ms__data_quality_flags__bit_7__suspect_time_tag**

Percentage of data in time window [T_1, T_2) for which bit 7 in the Data Quality Flag byte is set to '1'.

- **ms__activity_flags__bit_0__calibration_signal**

Percentage of data in time window [T_1, T_2) for which bit 0 in the Activity Flag byte is set to '1'.

- **ms__activity_flags__bit_2__event_begin**

Percentage of data in time window [T_1, T_2) for which bit 2 in the Activity Flag byte is set to '1'.

- **ms__activity_flags__bit_3__event_end**

Percentage of data in time window [T_1, T_2) for which bit 3 in the Activity Flag byte is set to '1'.

- **ms__activity_flags__bit_6__event_in_progress**

Percentage of data in time window [T_1, T_2) for which bit 6 in the Activity Flag byte is set to '1'.

- **ms__io_and_clock_flags__bit_5__clock_locked**

Percentage of data in time window [T_1, T_2) for which bit 5 in the I/O & Clock Flag byte is set to '1'.

- **ms_timing_correction_perc**

Percentage of data in time window [T_1, T_2) for which field 16 (“Time correction”) in the record header is non-zero.

- **ms_timing_quality**

Average of the timing quality percentage value stored in miniSEED blockettes 1001. Value is NULL if not present in the data records.

- **ms_timing_quality_median**

The 50-th percentile of all timing quality percentage value stored in miniSEED blockettes 1001 in time window [T_1, T_2). Value is NULL if not present in the data records.

- **ms_timing_quality_lower_quartile**

The 25-th percentile of all timing quality percentage value stored in miniSEED blockettes 1001 in time window [T_1, T_2). Value is NULL if not present in the data records.

- **ms_timing_quality_upper_quartile**

The 75-th percentile of all timing quality percentage value stored in miniSEED blockettes 1001 in time window [T_1, T_2). Value is NULL if not present in the data records.

- **ms_timing_quality_max**

Maximum of the timing quality percentage value stored in miniSEED blockettes 1001. Value is NULL if not present in the data records.

- **ms_timing_quality_min**

Minimum of the timing quality percentage value stored in miniSEED blockettes 1001. Value is NULL if not present in the data records.

- **num_gaps**

The number of gaps in time interval [T_1, T_2).

- **sum_gaps**

The sum of all gaps in time interval $[T_1, T_2)$.

- **num_overlaps**

The number of overlaps in time interval $[T_1, T_2)$.

- **sum_overlaps**

The sum of all overlaps in time interval $[T_1, T_2)$.

- **max_gap**

Largest gap in seconds in time interval $[T_1, T_2)$.

- **max_overlap**

Largest overlap in seconds in time interval $[T_1, T_2)$.
