Network of European Research Infrastructures for Earthquake Risk Assessment and Mitigation

Report

Automatic data QC and distribution statistics for data providers

Activity:	*Expanding access to seismic waveforms in the Euro-Med region*
Activity number:	*NA2*
Deliverable:	*Automatic data QC and distribution statistics routines for data providers*
Deliverable number:	*D2.3*
Responsible activity leader:	*Torild van Eck*
Responsible participant:	*KNMI*
Author:	*Reinoud Sleeman*

Seventh Framework Programme
EC project number: 262330

1. Introduction

Knowledge of quality of seismic waveform data and related metadata is essential for any scientific analysis and interpretation of seismic data. Measuring and monitoring the quality of such data is crucial for reliably estimate the location, depth and magnitude of an earthquake and is probably best done by experienced seismologists during daily routine analysis. Automated processes to calculate data quality parameters are required, however, to handle huge amounts of data and to enable observatories and data centers to automatically monitor changes or variations in data quality over different time scales. Automated processes can rapidly detect malfunctioning or degenerating of equipment, installation problems, recording failures, timing errors, or incomplete or wrong metadata. Calculating and keeping track of quality parameters serve thus the purpose to notify station operators about problems in a timely matter as well as the research community to search for and harvest the best quality data.

In a heterogeneous distributed network of data repositories, like EIDA and beyond, a coordinated, common approach in defining and calculating quality parameters is crucial for effectively exchanging quality information to be used by any service to provide the highest quality data (according to the criteria of the user). Also novel quality parameters, like sensor orientation or polarization, as derived by new and innovative techniques developed in the scientific community, should be incorporated in this framework as they can be extremely useful to further enhance the quality of the data by notifying the station operator. However, for a number of these new techniques (e.g. synthetic seismograms, cross correlations) we need to rely much on work done outside this consortium in projects like VERCE, EUDAT and COMMIT.

Although reliable and excellent software tools exists for real-time analysis of data (e.g. Antelope, Seiscomp3, EarthWorm) there is no common, accepted standard framework for seismological quality parameters. Our current practices are to use a set of parameters to describe specific features in the data (consisting of time series and metadata), like the mean value in a time series, the number of gaps in a time interval or the PSD of the ground acceleration at a certain frequency in a certain time interval, and call these parameters 'Quality Control (QC)' parameters.

Although the definition of most parameters may be unambiguous, the algorithms, their implementations to derive these parameters from the data, the archiving and the reporting are certainly not. This document aims at making an overview of current practices in QC data analysis, to provide basic QC parameter definitions and describe the workflow towards a common QC framework. However, due to the complexity of this challenge the present document is not static and will evolve in time based on the evaluation of different aspects described in the workflow, international developments (e.g. FDSN, IRIS) and the developments in other projects.

Expanding the access to seismic waveforms in the Euro-Med region, which is the main task in this work package, most likely will further enhance when data can be qualified in a uniform matter. Monitoring the use of data, in terms of distribution statistics, is therefore important both for data providers and for data centers. The current technology used in EIDA, called ArcLink (http://www.seiscomp3.org/wiki/doc/applications/arclink), enables monitoring of each data request including the user's email address, the amount of data shipped, the type of request (waveform data, inventory, routing), time window and stations from which the distribution statistics can be extracted.

2. Overview of current practices

2.1 European Integrated Data Archive (EIDA)

EIDA (http://www.orfeus-eu.org/eida/eida.html) is a distributed Data Center initiative within ORFEUS (a) to securely archive seismic waveform data gathered by European research infrastructures, and (b) to provide transparent access to the archives for the geosciences research communities. Current EIDA nodes are: ODC/ORFEUS, GEOFON/GFZ/Germany, SED/Switzerland, RESIF/CNRS/ INSU/France, INGV/Italy and BGR/Germany. These nodes signed an MoU with ORFEUS in which the commitments are established. Steering of EIDA is done by the EMB (EIDA Management Board) and the ETC (EIDA Technical Committee).

Currently, ArcLink is the protocol which technically connects the distributed archives and provides uniform access to the data archives. ArcLink uses the SeisComP3 database structure and provides access to the Quality Control database tables providing basic QC information (latency, offset, RMS, spike, gap, timing).

As 5 EIDA nodes are also participants in NA3 the EMB and ETC are taking the lead in coordinated quality control developments. New challenges in efficient data management, data content metadata, quality maintenance, provenance and access services in a distributed network of archives, are part of this discussion and related to developments in NERA NA9, NERA SA2, VERCE, COMMIT and EUDAT.

Current interactive data availability tools:

- http://www.orfeus-eu.org/cgi-bin/inventory_new.cgi
- http://geofon.gfz-potsdam.de/waveform/archive/data.php (Fig. 1)
- http://www.orfeus-eu.org/eida/marker_manager.htm (Google Maps with available stations within EIDA)

Basic QC parameters (currently for a subset of EIDA stations):

http://www.orfeus-eu.org/data/dailyqc-browse.php (Fig. 2)

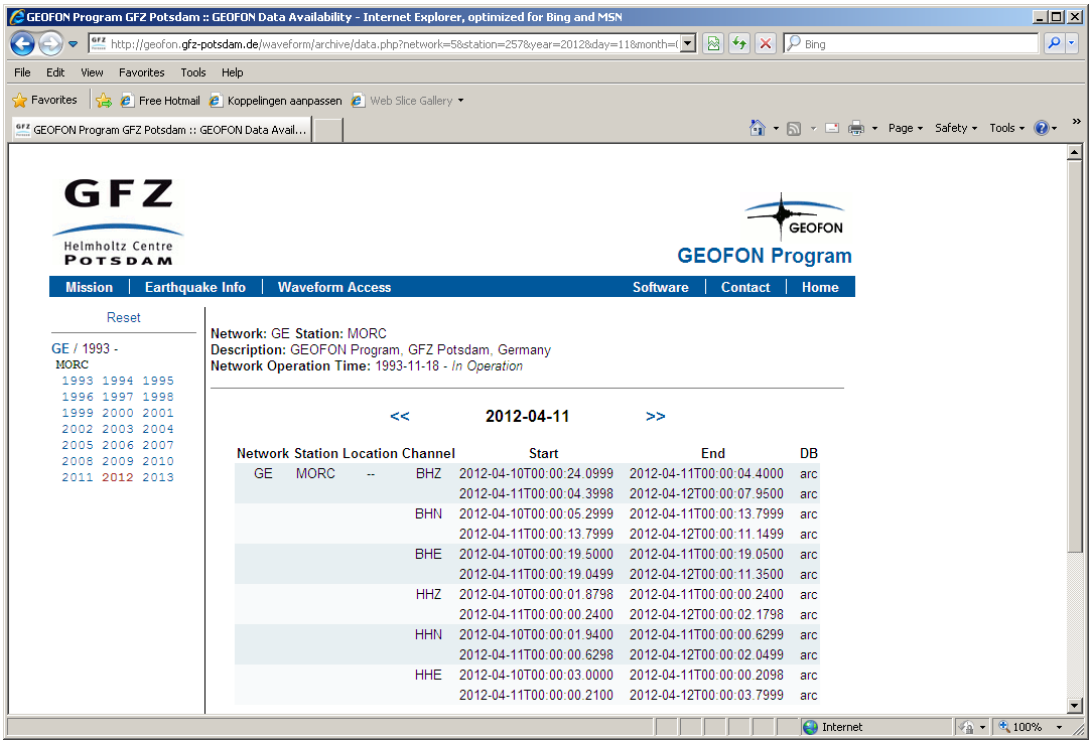


Fig 1. Example of data availability page (GFZ)

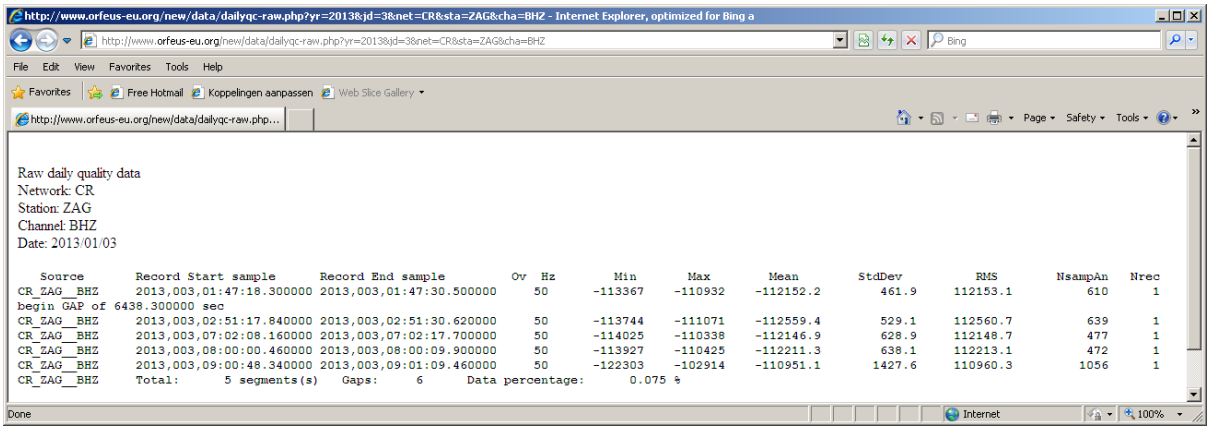


Fig 2. Example of on-line available basic QC information (ODC)

2.2 PSD / PQLX

Monitoring the Power Spectral Density (PSD) of the background noise and the representation of these values as (1) function of time ("power timelines") and (2) as Probability Density Functions (PDF) is a common (and very useful) tool to monitor the quality of the station (and the data) in time. An extremely important advantage of these techniques is the overall monitoring of site conditions, instrument stability over time and quality of metadata. A PSD versus time monitor enables rapid detection of significant changes in data quality, while the PDF representation provides insight in the overall performance of the station.

PSD/PQLX monitors:

- http://www.seismo.ethz.ch/research/groups/alrt/products/pqlx/index (Fig. 3)
- http://www.orfeus-eu.org/data/psd-versus-time-monitor.html ("power timelines")
- http://www.orfeus-eu.org/cgi-bin/pqlx_new.cgi

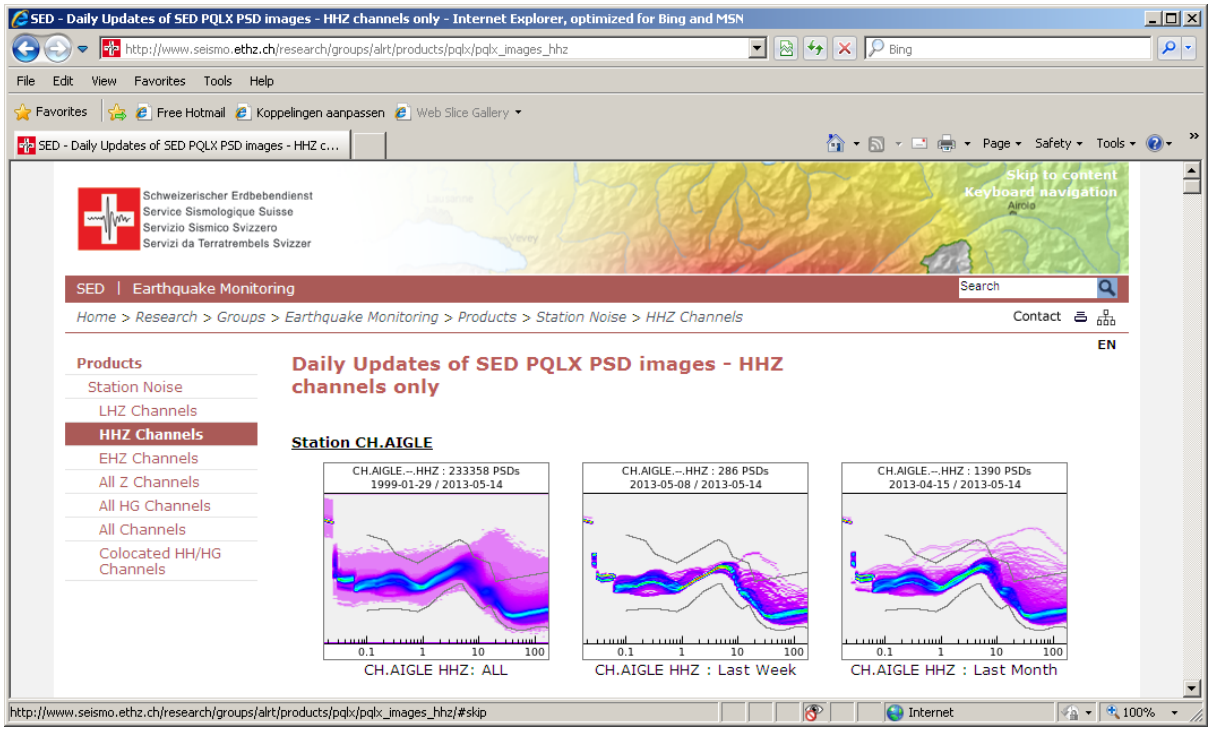


Fig 3. Daily update of PQLX images at ETHZ

3. QC parameters classification

A general classification of QC parameters is:

1. status variables determined at a defined point in time; this is mainly state-of-health (SOH) information read from digitizer logs like digitizer temperature, voltage, memory usage.
2. parameters assigned to a time window, derived from the data (e.g. gaps, mean value) or added by the network operator (information on station maintenance, defective instrument, etc.)
3. more or less continuous observation variables, e.g. PSD values, data latency or sensor orientation derived from cross correlation procedures

Based on current practices at data centers and seismic observatories as well as common needs for quality parameters expressed by users we propose the following list of basic quality parameters:

Essential:

(a) reliability of timing / known timing errors; type 2
(b) amplitude information (SNR, clipping, statistics (max, min, mean, standard deviation, RMS, mode, median)); type 2 and 3
(c) sign/polarization inversion; type 2
(d) severely disturbed signals (e.g. maintenance activities, construction activities, mass recentering operations, external noise sources, spikes); type 2
(e) gaps / gap density; type 2

Important:

(f) overlaps; type 2
(g) orientation of horizontal sensors; type 3
(h) PSD; both tracking PSD values over time and stacked PSD (PDF plot; PQLX) type 3
(i) data availability in a certain time window; type 2

Additional:

(j) data latency; type 3
(k) status variables read from digitizer log, the available parameter set depends on the digitizer manufacturer, i.e. not all parameters can be retrieved at all stations; all these parameters are of type 1:
timing information/quality, temperature, voltage, current, clock lock/unlock messages, time jumps, reboots/resets/restarts. memory usage, firmware version, software version, latitude, longitude, elevation, system notifications (e.g. write errors to storage media), memory errors of other faults
(l) manually inserted information about environmental influences on data like reasons for data outages, qualitative comments to data status or known problems type 2
(m) data integrity; although this is an important topic for data availability and reliability it is considered as a second order topic in this scope.

Parameters of type 1 are in general provided by the datalogger and no common algorithm or reporting is used. These parameters are usually stored as SOH files and/or in a database. Type 2 parameters usually require processing of the waveform data in a time window to automatically produce (b), (c), (d), (e), (f) and (i). The appendix provides a definition of the basic QC parameters (b), (e), (f) and (i). Additional information on (a), (d) and (l) may be provided by the network operator and added to the database. Parameters of type 3 requires processing of the continuous waveform data (j) and the corresponding metadata (g) and (h).

4. On-going work

Selection and definition of QC parameters. The above outline serves the on-going discussions for the QC parameters to be considered. Focus is on the essential and important classified parameters. On-going discussion on definitions within NA3 and which parameters are realistic to provide within NA3. Definition of common quality parameters is given below (Appendix).

Processing of time-series and SOH: For this task it is required to process time series within a specified time window and store the extracted values. Software to calculate the parameters (Fig. 2) is based on the 'libmseed' software package (Chad Trabant; IRIS DMC) and modified by ODC following the definitions.

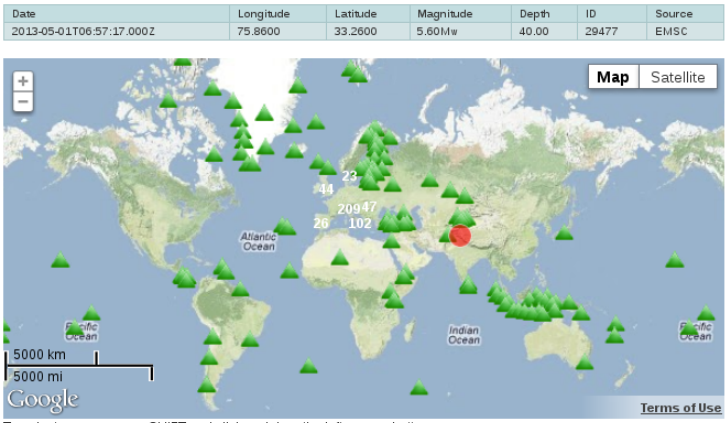
In addition, available SOH channels and/or logfiles need to be parsed on relevant information. Evaluation of existing SOH parsing tools (e.g. BGR).

Identifying parameter storage/format: Evaluation of existing (e.g. SeisComP3 (www.seiscomp3.org), SeisHub (www.seishub.org), BGR) QC databases and new developments (MonetDB, www.monetdb.org (ODC), Mustang (IRIS)) to store (and access) the different quality parameters. Using databases for QC parameters is efficient for data selection purposes (queries) based on such parameters. Within the current EIDA technology (in a distributed data archive system), however, it seems not realistic to

synchronize huge quality parameter databases between different EIDA nodes. Besides, pre-defined time window lengths ("granularities") in which quality parameters are calculated may not serve all users or clients. An alternative approach within the current technology to enable on-the-fly calculations of parameters in user defined time windows will be evaluated within MonetDB. MonetDB is a column store database technology with a CPU-tuned query execution architecture enabling fast queries on both static (PID based) and dynamic (on-the-fly calculations) of quality parameters. New QC algorithms may be added easily by extending the query language. PID implementation: the PID is a reference to an object (piece of data) which can be hooked with metadata like QC data. Currently investigated in relation with MonetDB (COMMIT).

For the noise characterization (PSD/PDF) the PQLX software and database is commonly used. An output format schema (XML, JSON) will be defined and implemented to be accessed by a webservice (ODC). Enhancing the webservice interface to the PQLX database to produce "power timelines".

Resolving time errors: For this parameter it is needed to evaluate different approaches. Possible approaches are 1) the extraction of timing information from SOH files, 2) using station time residuals based on manual picks (requires specific coordination with EMSC and ISC), 3) using station time residuals based on automatic pickers (preliminary developments on EIDA event data using AR-AIC picker (Fig. 4); JSON format), 4) array processing for teleseismic events and 5) noise cross-correlation within a network (costly in computational power).



To select an area press SHIFT and click and drag the left mouse button.

Date	Longitude	Latitude	Magnitude	Depth	ID	Source
2013-05-01T06:57:17.000Z	75.8600	33.2600	5.60Mw	40.00	29477	EMSC

Network	Station	Longitude	Latitude	EpiDist	BackAz	SNR	Tres
II	BRVK	70.283	53.058	20.19	166.35	3515.81	-0.60
II	KURK	78.620	50.715	17.56	187.69	806.26	1.96
II	AAK	74.494	42.639	9.42	172.99	434.21	-2.37
CN	FRB	-68.545	63.747	79.29	29.76	26.88	0.83
GE	KMBO	37.252	-1.127	49.91	43.11	24.75	-0.44
CN	RES	-94.896	74.689	72.16	8.13	20.48	0.07
II	ESK	-3.205	55.317	57.41	77.55	19.85	1.98
TU	ILGA	33.717	41.052	34.21	89.35	15.80	-1.25
KO	RSDY	37.327	40.397	31.47	90.68	6.21	2.82
BN	EKB	-3.180	55.340	57.39	77.58	6.01	2.08
NL	OPLO	5.812	51.589	52.78	81.54	5.26	2.82
IU	PAB	-4.350	39.545	62.86	68.11	5.20	0.53
NO	NC405	11.715	61.113	49.11	94.12	5.15	0.42

Fig 4. Example of preliminary results providing automatic signal-to-noise (SNR) information and P-onset time residual (based on AR-AIC picker) for EIDA station (in development).

IT challenges for the long term are: 1) an efficient algorithm to locate the best quality of data in a distributed archive system based on a set of parameters 2) keeping track of data provenance 3) using QC parameters in services (queries). These developments are strongly related to NERA NA9, VERCE, COMMIT and EUDAT.

Appendix: Definition of common QC parameters

Parameters are calculated in a single, continuous time series in a defined time window $[t_0, t_1]$. A single time series is identified by a network code NN, station code SSS, channel code CCC, location code LL and a data quality code Q ('D', 'R', 'Q', 'M').

Software to calculate the parameters is based on the libmseed software package (Chad Trabant; IRIS DMC) and uses the default definitions for gap and overlaps (see below).

Mean (Average, Offset) [counts]

Average value of all samples in a specified time window $[t_0, t_1]$.

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

N and M represent sample indices for the earliest/first and latest/last sample within time window $[t_0, t_1]$ respectively.

RMS [counts]

The root mean square (RMS) is a statistical measure of the amplitude of all samples in a time window $[t_0, t_1]$.

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

Indices N and M as above. (No check on gaps or overlaps is done here).

Reporting parameters: $[t_0, t_1]$, rms

Standard deviation [counts]

The standard deviation is the root mean square (RMS) deviation of values from their arithmetic mean.

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

Indices N and M as above. (No check on gaps or overlaps is done here).

Reporting parameters: $[t_0, t_1]$, stdev

Mode

The most frequent value in the time series (not included yet)

Median

Middle value separating the greater and lesser halves of sample value dataset (not included yet)

Max

Maximum value in the time series.

Min

Minimum value in the time series.

Continuous time series - definition

A continuous (discrete) time series is defined as a time series in which the time interval between two adjacent samples (a) is constant or (b) does not differ from this constant by more than a certain time tolerance. In other words, a continuous time series is defined as a time series in which the samples are equally spaced in time within a certain time tolerance.

$\Delta t - \varepsilon <= T_{m+1} - T_m <= \Delta t + \varepsilon$.

m is the sample index number in the time series, $N <= m < M$, T_m is the time corresponding to sample m, Δt the (constant) sample rate (in s), ε is the time tolerance is s. Default time tolerance is 1e-4 s (following libmseed.h)

Therefore, changes in the sample rate which are larger than the time tolerance define a discontinuity in the time series.

Gap [s]

A discontinuity in the time series is a gap when the time interval between two consecutive, continuous time segments exceeds the sample rate interval (Δt) by more than the time tolerance ε. The start time ($T_{2,start}$) of the second segment is later than the end time ($T_{1,end}$) of the first segment by more than the sample rate and the time tolerance ε .

Overlap [s]

An overlap is the time interval between two consecutive, continuous time segments for which the start time ($T_{2,start}$) of the second segment is earlier than the end time ($T_{1,end}$) of the first segment. According to definition above this difference exceeds the sample rate interval (Δt) by more than the time tolerance ε.

Data availability:

Data availability is the length of the time window (t_0, t_1) minus the sum of all gaps in this time window, relative to the length of the time window t_1-t_0.

$$availability = \frac{(t_1 - t_0) - \sum gaplength_i}{t_1 - t_0} \text{ x } 100\%$$

Mean time reception quality [percentage]

In case the mini-SEED data records have blockette 500 the mean time reception quality in time window (t_0, t_1) is the mean of field 6 values in blockette 500 for all data records that (partially) fit in time window (t_0, t_1).

Spike (time, max. amplitude [counts], interval [s], interval [s])

The occurrence of one or more spikes in a record can be described by the number N of spikes, the time corresponding to the maximum value of each spike, the maximum value of the spike, and the mean time interval between adjacent spikes (for N>1). A spike finder algorithm is still preliminary. (Not included).