NERA

Network of European Research Infrastructures for Earthquake Risk Assessment and Mitigation

Report

Data quality improvement statistics

Activity:	*Expanding access to seismic waveforms in the Euro-Med region*
Activity number:	*NA2, 2.5*
Deliverable:	*Data quality improvement statistics*
Deliverable number:	*D2.4*
Responsible activity leader:	*Torild van Eck*
Responsible participant:	*KNMI*
Author:	*Reinoud Sleeman (KNMI)*

Seventh Framework Programme
EC project number: 262330

Summary

A new, flexible approach to describe seismic waveform data in terms of quality parameters has been designed and implemented with the aim to serve both network operators, data centres and researchers in using a variety of quality parameters to serve different purposes, like detecting a decrease in data quality, to monitor overall network performance or to search for (specific) high quality data.

Description

The research community requires increased data quality and data completeness as well as services to search for data, given a range of quality parameters. Common quality parameters are: gaps, timing quality, latency, signal-to-noise ratio, statistical values, power spectral parameters, etc. Preserving of the highest quality of data available from a data centre requires advanced monitoring tools and mechanisms to detect changes in such quality parameters which may indicate a performance decrease.

During the NERA project the acquisition and archiving policy within ORFEUS changed from a centralized archive at ORFEUS Data Centre (ODC) into a distributed archive system called EIDA (European Integrated Data Archive) and hence influenced the development of procedures and services with respect to data quality of waveform archives in Europe. A centralized service could no longer facilitate the need to monitor or provide quality parameters and had to evolve into a distributed service. Therefore the focus and outcome of this task shifted accordingly, into a solid core component to serve a variety of needs in the next generation software to build in EIDA.

The description of work for task 2.4 also includes the installation of secondary data retrieval routines to improve data completeness. This task, however, has become less important with the implementation of EIDA in ORFEUS. Within EIDA each archive (node) is responsible for a one or more seismic networks. Nodes that are responsible for a national network (e.g., ETH for CH network, INGV for IV network) do have the highest quality data by default, which is often organized and anchored on a national level. Often procedures are in place to (for example) fill gaps through regular synchronization with the station data buffer. Only ODC and GFZ are archiving data from a multitude of national networks, and here the issue of data completeness between archives may be investigated. However, the risk of data loss during transmission in real-time has been reduced significantly due to the use of native data exchange protocols. Mainly SeedLink is used as robust and stable protocol in Europe. For a few networks running the SCREAM protocol (Guralp) or Antelope, ODC uses these protocols to minimize data loss which mainly occurs on transmission protocol interfaces. Furthermore, ODC improved the SCREAM plugin for SeedLink by using a ring buffer and TCP requests for missing data records. But common practice has shown that the most efficient way to ensure availability of the highest data quality at a centralized centre like ODC is to copy the national archives at regular times rather than to maintain an (often) complex system of procedures to automatically detect gaps, request data for those gaps and merge them in the existing archive.

A few approaches in the past to monitor and provide quality parameters to both users and network operators were initiated but turned out not to be adequate or sustainable. A system like SeisComP3 has a build-in module for monitoring and storing quality parameters but this was never used extensively by data centres. More common practices in monitoring services are directed to websites showing, for example, latency, Power Spectrum Probability Density distributions (PDF's), PSD variations over time, station time residuals (using e.g., EMSC locations), station magnitude residuals (compared to e.g., EMSC magnitude) or interfaces showing gaps and/or statistical values per day. Most of these approaches were (and still are) related to a service connected to a local archive, without a coordinated, common approach in Europe, and often needed considerable resources for long term maintenance.

Although the services were (and still are) useful for providing a quality snapshot in time (e.g., latency), or providing a more general view over a longer time frame (e.g., PSD versus time, or histograms), none of these services could be of help to users to search for and select data that would fit one or more quality criteria set by the user. Besides, no common or standardized monitoring technique was in place at different data centres when EIDA was initiated. A similar drawback is seen towards the post-acquisition processing of raw data in order to, for example, repair timing problems, fill gaps or detect metadata errors.

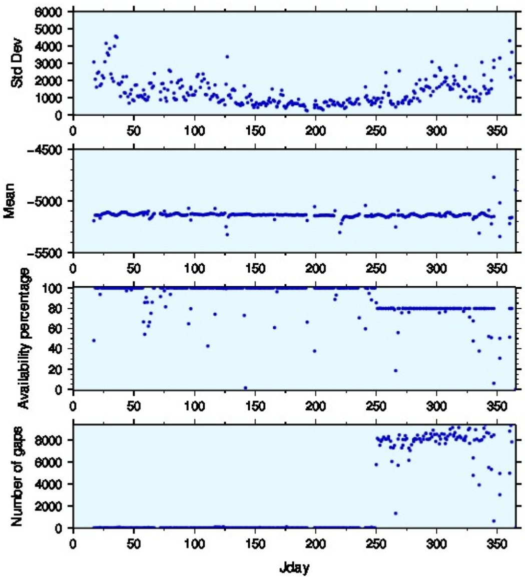
Task 2.4 therefore focussed on the work done in Task 2.3 (definition of QC parameters) and the implementation of software towards a service that could benefit both the research community and the data provider community. The philosophy is to design a flexible core system, which would host a suite of parameters describing the waveforms (waveform metadata), while new parameters could be added easily, including for example the analysis of State-Of-Health (SOH) channels. A variety of clients can connect to this system in a uniform way to retrieve waveform metadata (e.g., quality parameters) and make decisions (based on the metadata) whether the data fits the criteria. This approach enables and allows different selection criteria by the users, as well as by network operators, as quality may be defined in various ways. As an example, the existence of data gaps beyond the earthquake signal are not of interest for a user who only needs the earthquake signal, while they may make the data less valuable for users analysing long time series (e.g., for ambient noise studies or cross correlations). The design and implementation of this service is therefore an important step forward in the development of a Mediator/Broker within the next generation EIDA software, to make decisions whether data fits the criteria, by using webservices.

The ODC developed a first webservice to collect waveform metadata for this purpose which may potentially serve a new, broad range of clients for data selection clients based on any combination of quality parameters. This waveform metadata webservice will be used in the EIDA Broker/Mediator to distinct “good” data from “bad” data. The service essentially exists of the following components:

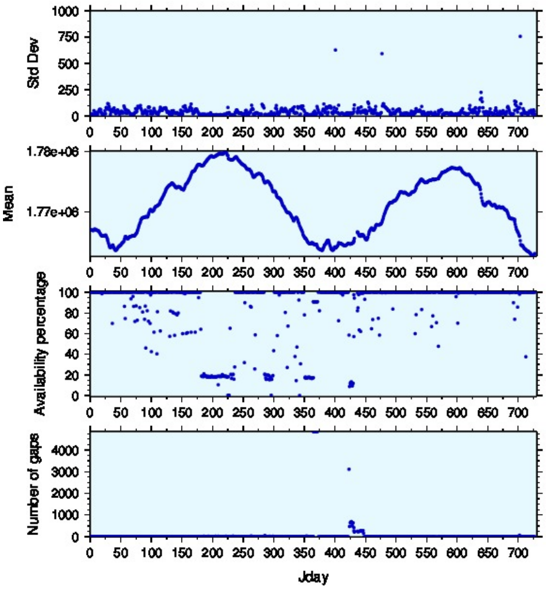
- mseedMetadatCollector (python)
  this tool calls the core of the QC processing, called ‘msstatqc_json’ which processes mini-SEED files in an archive, according to the QC document provided in task 2.4 (http://www.orfeus-eu.org/man/msstatqc.html), and parses the JSON output into a database (MongoDB). The choice of using MongoDB has made the database flexible to extend and add new parameters to the existing ones. Current granularity of the data being processed is 1 day, with the aim to demonstrate the feasibility of the system in a distributed system. The flexibility of the system allows for smaller time-windows (e.g., 1 hour) to be deployed in the next release.
- odcws_wfmetadataselect (java)
  this webservice actually translates the user’s input into querying the database to provide the corresponding metadata
  http://www.orfeus-eu.org/man/odcws_wfmetadataselect.html

Examples

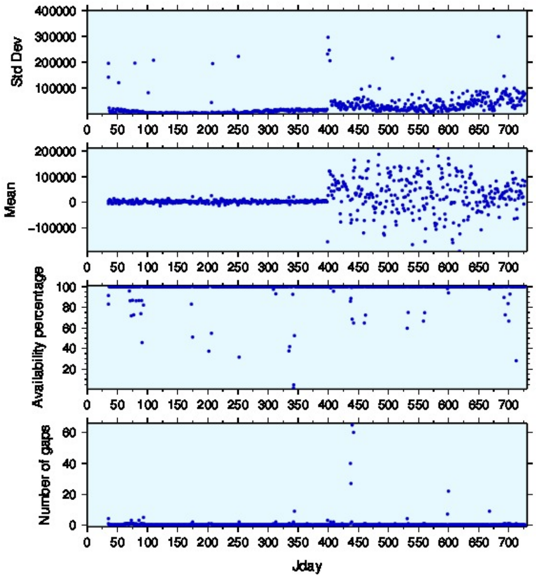
This section shows some examples of the usefulness of this system by applying queries to the system which easily can be transformed into a graphical representation. The examples show 4 parameters (from bottom to top) as function of Julian day: number of gaps per day, the availability percentage per day, the mean values of the raw data and the standard deviation (both per day). Note that all plots are made for raw digital counts. 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 (default here is 1e-4 *s*). The total of gaps is represented by the data availability.



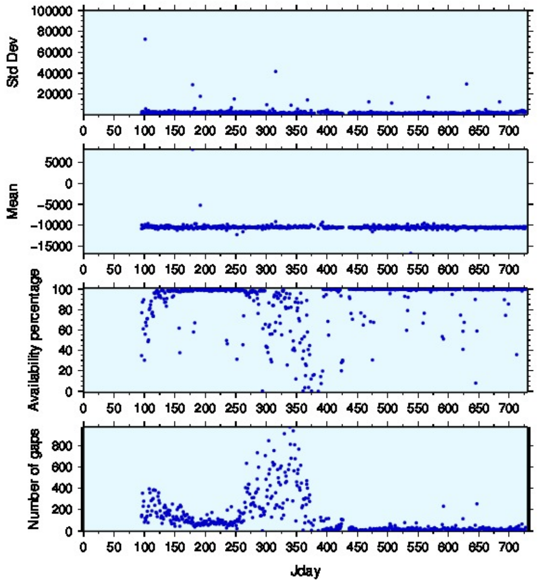
Example 1: Data from station BN_WOL_BHZ in 2012. This example shows a distinct increase in number of gaps towards the end of the year, and a corresponding drop in data availability. Besides, the standard deviation in the raw data follows the well-known seasonal variation in background noise (relatively high noise in winter and low noise in summer on the northern hemisphere).



Example 2: This example shows the same 4 parameters for station BS_JMB (2012, 2013). Most interesting observation here is the mean trace which inversely follows the annual variation in background noise and may therefore reflect local temperature variations. In terms of quality improvement the data completeness has increased significantly in the second year, mainly caused by an improved communication configuration at the server side.



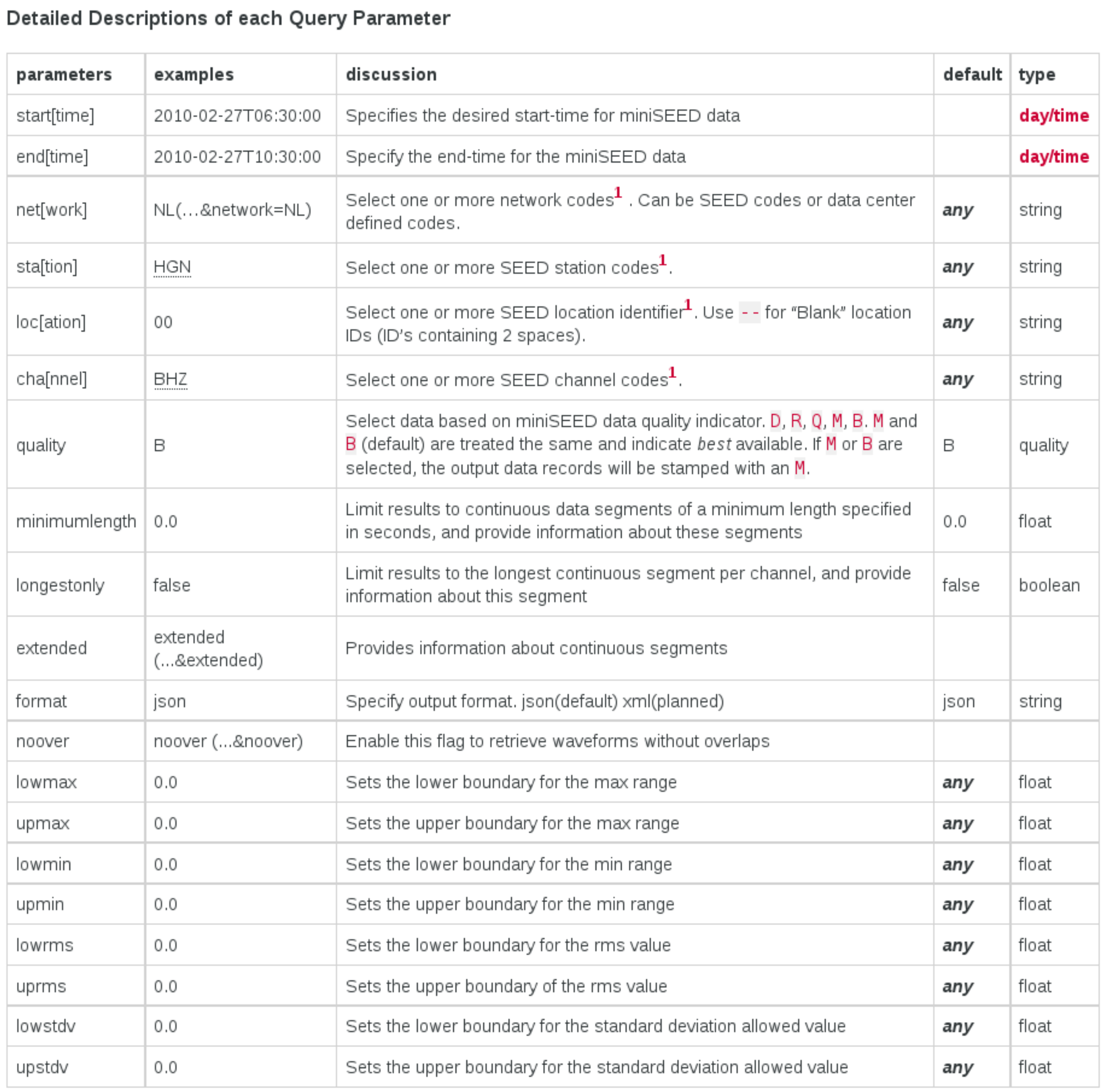
Example 3: This example shows the 4 parameters for station NO_TROLL_BHZ (2012, 2013). Clearly both mean values and standard deviations increase significantly around day 400. This example demonstrates that the interpretation of the quality parameters cannot be done without using the station metadata, showing a change in sensitivity at the same day.



Example 4: Significant improvement in data availability due to a reduction in gaps for station/stream GB_ELSH_HHZ.

Links to concrete results

odcws_wfmetadataselect: http://www.orfeus-eu.org/man/odcws_wfmetadataselect.html
core QC software: http://www.orfeus-eu.org/man/msstatqc.html



Detailed Descriptions of each Query Parameter

parameters	examples	discussion	default	type
start[time]	2010-02-27T06:30:00	Specifies the desired start-time for miniSEED data		**day/time**
end[time]	2010-02-27T10:30:00	Specify the end-time for the miniSEED data		**day/time**
net[work]	NL(…&network=NL)	Select one or more network codes[1] . Can be SEED codes or data center defined codes.	***any***	string
sta[tion]	HGN	Select one or more SEED station codes[1].	***any***	string
loc[ation]	00	Select one or more SEED location identifier[1]. Use -- for "Blank" location IDs (ID's containing 2 spaces).	***any***	string
cha[nnel]	BHZ	Select one or more SEED channel codes[1].	***any***	string
quality	B	Select data based on miniSEED data quality indicator. D, R, Q, M, B. M and B (default) are treated the same and indicate *best* available. If M or B are selected, the output data records will be stamped with an M.	B	quality
minimumlength	0.0	Limit results to continuous data segments of a minimum length specified in seconds, and provide information about these segments	0.0	float
longestonly	false	Limit results to the longest continuous segment per channel, and provide information about this segment	false	boolean
extended	extended (...&extended)	Provides information about continuous segments		
format	json	Specify output format. json(default) xml(planned)	json	string
noover	noover (...&noover)	Enable this flag to retrieve waveforms without overlaps		
lowmax	0.0	Sets the lower boundary for the max range	***any***	float
upmax	0.0	Sets the upper boundary for the max range	***any***	float
lowmin	0.0	Sets the lower boundary for the min range	***any***	float
upmin	0.0	Sets the upper boundary for the min range	***any***	float
lowrms	0.0	Sets the lower boundary for the rms value	***any***	float
uprms	0.0	Sets the upper boundary of the rms value	***any***	float
lowstdv	0.0	Sets the lower boundary for the standard deviation allowed value	***any***	float
upstdv	0.0	Sets the upper boundary for the standard deviation allowed value	***any***	float

Current query parameters supported in the waveform metadata query system.