Data loss at the Swiss seismological service in 2004

Roman Racine, racine@sed.ethz.ch

January 6, 2017

Abstract

The swiss seismological service has started to operate a broadband digital network in 1999. Starting form August 1999, this data is archived at its original sampling rate as well as downsampled to 1sps. The archived data set is continuous from August 1999 up to this day, except for a large gap in 2004. This report focuses on the nature and the reasons for this gap and what measures have been undertaken to prevent further gaps.

Overview

The SED (Schweizerischer Erdbebendienst, Swiss seismological service) operates a broadband digital network since 1999. Archival of continuous data has started in August 1999. Data has been stored at 120sps and at 1sps. Since then, various data mediums and data formats have been used to store the data. As well, data sampled at 40sps has been added in retrospective in 2016, downsampled from 120sps.

1999-2003

Data was stored in SED KP binary format. This format is propriatary to the SED. One data file contained five minutes of waveform data for all available channels for the data sampled at 120sps and one day of data for all channels for the data sampled at 1sps. Only these two sampling rates were stored permanently. Chunking data into pieces of five minutes was used as the standard practice up to 2012. The main servers held a few weeks of data locally. For long term archival, data was stored on DDS2 tapes. These tapes are still stored in the SED safe in the B floor of the NO building up to this day.

2003-2007

The data format was now changed from the SED internal KP format to the standardised and internationally used GSE2 format, originally developed for the International Monitoring System (IMS) of the Comprehensive Nuclear Test Ban Treaty. [1] Still, data was chunked into blocks of five minutes. Each five minute file contained all available channels.

Data storage was moved from tapes to external disks.

In addition to the sampling rates 120sps and 1sps which had been created and archived since 1999, the sampling rate 40sps was created in real time and sent to ORFEUS. However, this sampling rate was not archived at the SED.

2007-2012

Data format remained GSE2. Several improvements were made to the data storage. A RAID system was purchased to replace the lose external disks in 2007. As well, data backup on LTO3 tapes was introduced and the tapes were stored in the SED safe. In 2008, old data in binary KP format on DDS2 was converted to GSE2 and stored on the RAID system.

In 2008, permanent archival of the data was moved away from SED-run hardware to the centralised IT services of the ETH (Informatikdienste). The SED still made backups on LTO3 to store them in the safe.

2012-2016

Data format changed to miniseed, the SDS data structure was adopted. [2] All existing data was converted from GSE2 to miniseed in 2012. The data archive

was made available via arclink [3] in 2012. Backups were shifted from LTO3 to hard disks in 2015. Since then, regular full copies of the archive are made, whereof one is stored outside of the ETH. The archive was made accessible via fdsnws [4] in 2014.

Data loss in 2004

In 2004 the typical size of an affordable external hard disk (250GB) was suitable to hold about five months of data. Apparently, one of the disks, holding data from April 2004 to August 2004 failed partially, resulting in approximately 75% data loss. As a strategy to have at least two separate and independent copies of the waveform data was only adopted in 2007, there is no way to recover the lost data with the means of the SED.

However, in 2004, data was already converted to miniseed, decimated to 40sps on the fly and streamed to ORFEUS [5] in real time. Note that this 40sps stream was not stored locally at the SED but only streamed to ORFEUS. After contacting ORFEUS (Reinoud Sleeman), it turned out that data of only four stations was stored permanently. The data stored at ORFEUS was transferred to the SED again upon request. As 28 broadband stations were streaming in real time during the affected time window, data of 24 stations of all sampling rates is lost permanently and 40sps data of four stations has been recovered. (CH.BNALP, CH.BOURR, CH.FUORN, CH.SALAN). 1sps data can of course be recreated for these four stations.

Resulting situation

The resulting situation for the time period starting from April 2004 to August 2004 is as follows:

- 120sps data: 75% data loss, recovered files scattered amongst the time period
- 40sps: four stations fully recovered (CH.BNALP, CH.BOURR, CH.FUORN, CH.SALAN) from ORFEUS, the rest of the data is decimated from the 120sps data set in retrospective, resulting in the same amount of data loss as for the 120sps channels.
- 1sps: 2004 apparently was completely copied onto the faulty disk and is therefore completely lost, spare for a few days recovered from the faulty disk. Data can be recreated from 40sps streams where available.

Measures taken since the data loss

The most important measure taken after the data loss was the strategy to have a second copy of the data on LTO3 tapes. It's most interesting to notice that ever since a complete copy of all the data has been made, not a single backup medium has ever failed again.

Starting from 2015, two copies of the data on hard disks are made regularly, where one is stored at the ETH and one is stored outside of the ETH, further enhancing data data safety.

To further enhance data completeness, all new dataloggers in the Swiss broadband network phased in starting from 2014 are equipped with a large local storage which can hold the data for months. In case of a communication failure, the data can be retrieved and pushed to the archive at a lager stage.

Plots

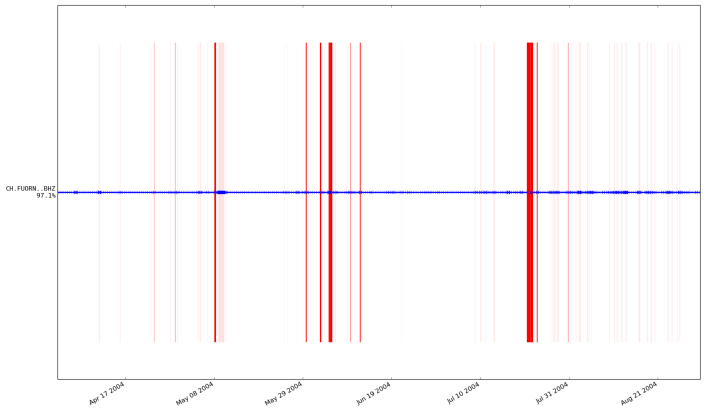

Figure 1: CH.FUORN..BHZ, data retrieved from ORFEUS in 2016 and pushed back to the SED archive

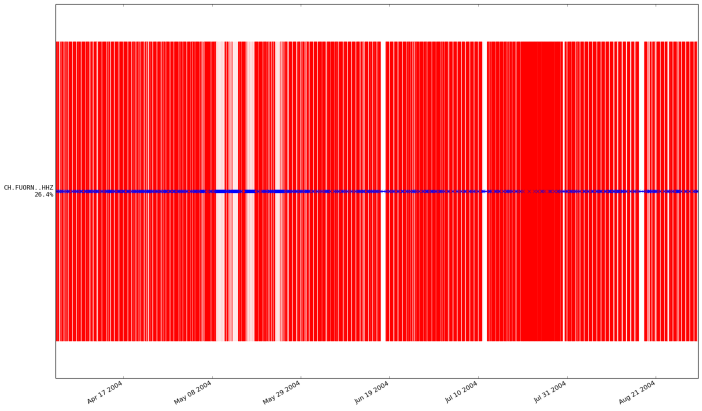

Figure 2: CH.FUORN..HHZ, 73.6% of the data irrecoverably lost

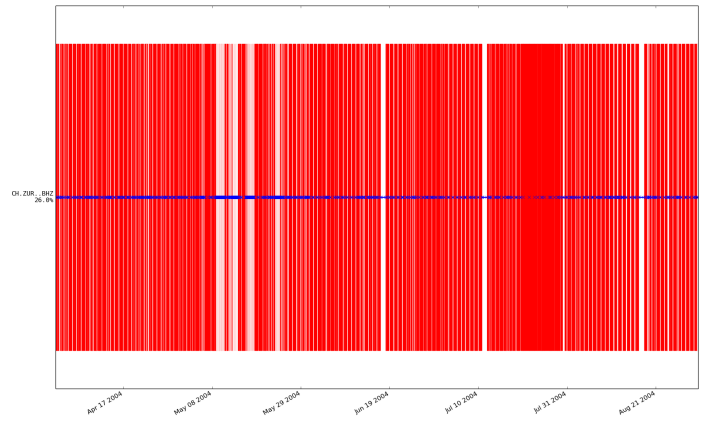

Figure 3: CH.ZUR..BHZ, data not available from ORFEUS and recreated in 2016 from the incomplete HHZ set

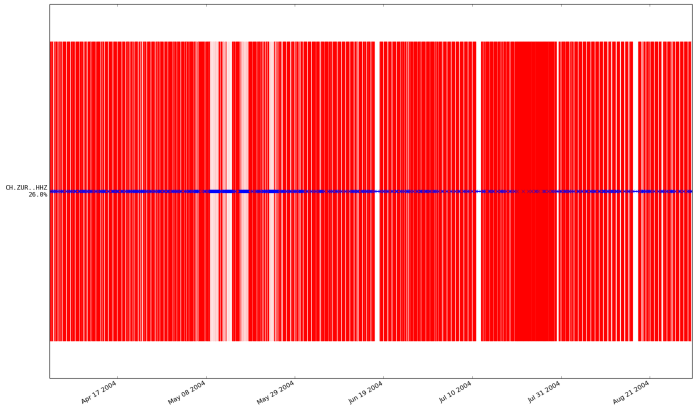

Figure 4: CH.ZUR..HHZ, 74.0% of the data irrecoverably lost

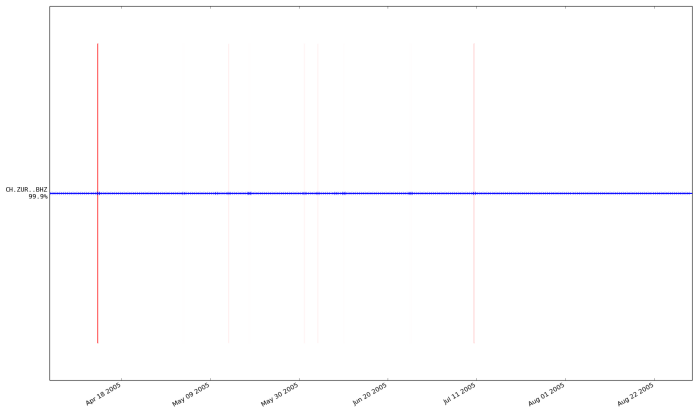

Figure 5: CH.ZUR..BHZ, Same time of the year in 2005 for comparison

Appendix

A more detailed overview over the gaps by station is provided here. It's easy to see that most of the channels share the same pattern of gaps. This is no surprise, as the data was stored in five minute GSE2 files containing all channels at that time. Usually, files were completely lost due to the disk failure or were completely recoverable. The same period for 2005 is provided as well for comparison.

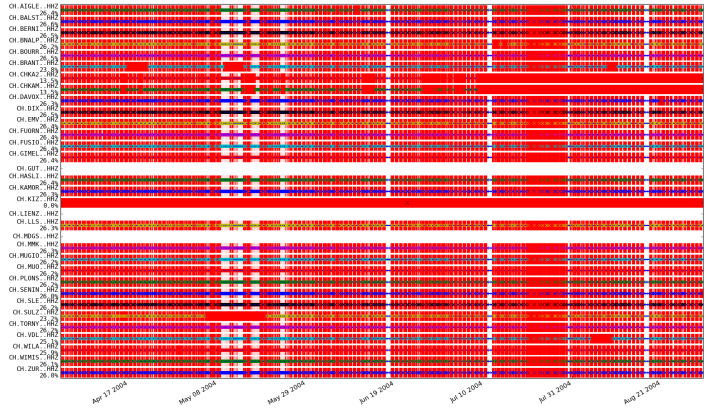

Figure 6: Overview over all channels from April to August 2004

Station	Available data
CH.AIGLE	26.91%
CH.BALST	27.04%
CH.BERNI	27.01%
CH.BNALP	26.69%
CH.BOURR	26.99%
CH.BRANT	24.27%
CH.CHKA2	14.07%
CH.CHKAM	14.07%
CH.DAVOX	26.75%
CH.DIX	26.94%
CH.EMV	26.92%
CH.FUORN	26.87%
CH.FUSIO	26.90%
CH.GIMEL	26.89%
CH.HASLI	26.86%
CH.KAMOR	26.82%
CH.KIZ	0.66%
CH.LLS	26.79%
CH.MMK	26.75%
CH.MUGIO	26.66%
CH.MUO	26.72%
CH.PLONS	26.71%
CH.SENIN	26.53%
CH.SLE	26.65%
CH.SULZ	23.67%
CH.TORNY	26.64%
CH.VDL	25.62%
CH.WILA	26.41%
CH.WIMIS	26.55%
CH.ZUR	26.52%

Table 1: Data completeness on the HH channels April to August 2004

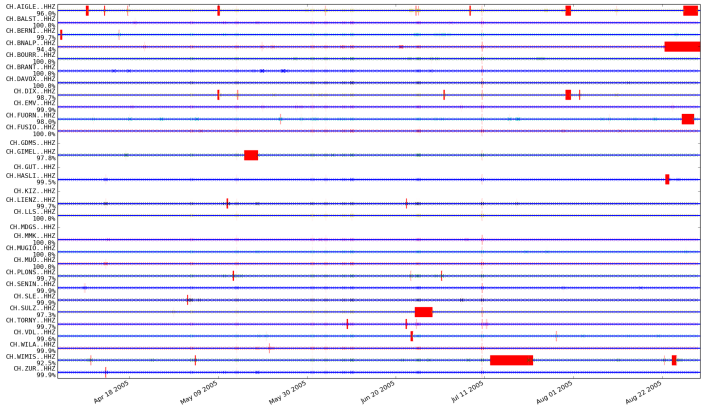

Figure 7: Overview over all channels from April to August 2005

Station	Available data
CH.AIGLE	96.05%
CH.BALST	99.96%
CH.BERNI	99.67%
CH.BNALP	94.45%
CH.BOURR	99.95%
CH.BRANT	99.95%
CH.DAVOX	99.96%
CH.DIX	98.67%
CH.EMV	99.95%
CH.FUORN	98.00%
CH.FUSIO	99.96%
CH.GIMEL	97.86%
CH.HASLI	99.46%
CH.LIENZ	99.71%
CH.LLS	99.96%
CH.MMK	99.96%
CH.MUGIO	99.96%
CH.MUO	99.95%
CH.PLONS	99.72%
CH.SENIN	99.94%
CH.SLE	99.89%
CH.SULZ	97.28%
CH.TORNY	99.72%
CH.VDL	99.57%
CH.WILA	99.92%
CH.WIMIS	92.56%
CH.ZUR	99.87%

Table 2: Data completeness on the HHZ channels April to August 2005 for comparison

Bibliography

[1] GSE2.1 provisional standard
http://www.seismo.ethz.ch/export/sites/sedsite
/research-and-teaching/.galleries/pdf_products_software/provisional_GSE2.1.pdf

[2] SeisComP Data Structure (SDS)1.0, GFZ Potsdam
https://www.seiscomp3.org/wiki/doc/applications/slarchive/SDS

[3] http://arclink.ethz.ch/webinterface/

[4] http://eida.ethz.ch/fdsnws/ For a more comprehensive description, see
http://www.fdsn.org/webservices/

[5] Observatories & Research Facilities for European Seismology
http://www.orfeus-eu.org/