M1.3 Report: common data formats and analysis tools

Matteo Di Giovanni[1,2], Shahar Shani-Kadmiel[3], Carlo Giunchi[4], Rosario De Rosa[5,6]

[1] Gran Sasso Science Institute I-67100 L'Aquila, Italy
[2] INFN, Laboratori Nazionali del Gran Sasso I-67100 Assergi (AQ), Italy
[3] KNMI NL-3731 De Bilt, Netherlands
[4] INGV, sezione di Pisa, I-56123 Pisa, Italy
[5] Universita degli Studi di Napoli Federico II, I-80126 Napoli, Italy
[6] INFN, sezione di Napoli, I-80126 Napoli, Italy

Keywords: *(min 3, max5) keyword1, keyword2, keyword3, keyword4, keyword5.*

Abstract. This document is aimed at defining the standards for data format and analysis tools that are and will be used for Einstein Telescope site characterization studies. The goal is to provide the appropriate tools and methods to make the analysis easily replicable to facilitate the comparison between the candidate sites.

1 INTRODUCTION

The identification of two candidate sites to host the future 3rd generation gravitational wave (GW) detector Einstein Telescope (ET) prompted, at the end of last decade, extensive long-term site characterization campaigns aimed at the assesment of the suitability of the aforementioned sites to host ET. This meant that a pletora of research groups, from different research institutes all over Europe, each with its own data analysis methods and tools, had to find a way to define common data formats and analysis tools in order to compare the two sites as uniformly as possible and to make the analyses easily replicable by any scientist. Today, the amount of data collected during several years of observations and the number of observables involved (seismic, acoustic, magnetic, etc...) require uniformity and ease of access. Therefore, the goal of this document is to define the standard data formats and tools for site characterization studies.

2 DATA FORMATS

2.1 MiniSeed

The IRIS (Incorporated Research Institutions for Seismology) consortium provides several formats for the exchange of seismic data. These formats can be summarized in:

- SAC;
- SEED;
- MiniSEED;

The SEED (Standard for the Exchange of Earthquake Data) data format is intended primarily for the archival and exchange of seismological data. A so-called "full SEED" volume is the combination of time series values along with comprehensive metadata. Usually, metadata comprise geographic coordinates, response/scaling information and other information needed to interpret the data values are not included. The SEED data format was originally designed in the late 1980s and still remains in widespread use. Nevertheless, this format is designed mainly for archival rather than processing, resulting in large files more difficult to store and handle. Therefore, more convenient data formats were developed.

Among these new formats in MiniSEED, a stripped down version of SEED containing only waveform data and very limited metadata like time series identification and simple state-of-health flags. Time series are stored as generally independent, fixed length data records which each contain a small segment of contiguous series values. A reader of miniSEED is required to reconstruct longer, contiguous time series from the data record segments. Common record lengths are 512-byte (for real time streams) and 4096-byte (for archiving), other record lengths are used for special scenarios.

The lack of metadata information in MiniSEED means that this format must always be accompanied by a separate station metadata file, containing information about geographical coordinates, type of sensors installed, response and other information needed to interpret the data values.

In conclusion, for ET site characterization studies, the collaboration reached an agreement for which seismic data must be stored and shared using the MiniSEED format in conjunction with metadata files.

2.2 Station XML metadata files

The advent of MiniSEED, brought to the development of two different formats to provide a standardized format for geophysical metadata:

- Dataless SEED (developed by IRIS);
- StationXML (developed by FDSN);

The latter being chosen by the ET collaboration as the standard for sharing station metadata. In principle, StationXML is an XML representation of metadata that describes the data collected by geophysical instrumentation. StationXML is an improvement and an extension of the SEED format with which it is fully compatible. Given the information about each seismic station, StationXML files can be easily built and manipulated by any library devoted at handling seismic data (e.g. ObsPy).

2.3 SDS data structure

The huge amount of data produced by seismic sensors pushes for an efficient way to organize MiniSEED files. The most common is the SeisComP Data Structure (SDS). In conjunction with the adoption of the MiniSEED and StationXML formats, the SDS has been chosen as the default structure to store and share seismic data within the ET collaboration.

Given that each seismic station is usually represented by some identifiers such as the code of network to which it belongs (e.g. NET), the code of its name (e.g. STA), the location (e.g. LOC) and the data channel (e.g. CHA), the basic directory and file layout of the SDS is defined as SDSdir/Year/NET/STA/CHAN.TYPE/NET.STA.LOC.CHAN.TYPE.YEAR.DAY. Where NET.STA.LOC.CHAN.TYPE.YEAR.DAY is the file name and:

- SDSdir : arbitrary base directory;
- YEAR : 4 digit year;
- NET : Network code/identifier, up to 8 characters, no spaces;
- STA : Station code/identifier, up to 8 characters, no spaces;

- CHAN : Channel code/identifier, up to 8 characters, no spaces;
- TYPE : 1 characters indicating the data type, recommended types are:
  - ’D’ - Waveform data;
  - ’E’ - Detection data;
  - ’L’ - Log data;
  - ’T’ - Timing data;
  - ’C’ - Calibration data;
  - ’R’ - Response data;
  - ’O’ - Opaque data;
- LOC : Location identifier, up to 8 characters, no spaces;
- DAY : 3 digit day of year, padded with zeros.

The dots in the file names must always be present regardless if neighboring fields are empty. Therefore, data from each day of the year are stored in a single MiniSEED file.

The use of the SDS structure facilitates the uploading and handling of seismic data. Dedicated libraries, such as ObsPy, don’t need to upload data file by file but require only few identifiers, together with the GPS start and end times of the segments of interest, to read hundreds of MiniSEED at a time and merge all the data in a continuous time series ready to be processed by the user.

The SDS data structure proves also convenient for data not only from seismometers. For example, microphone data can be stored in the same data structure and are easily readable by ObsPy.

2.4 NetCDF

We reached an agreement for which spectra should be saved, stored and shared using the NetCDF format. The same format is also the standard for the environmental data download from external providers such as Copernicus.

2.5 CSV

Environmental data from weather stations installed at the sites should be shared in CSV files, one per each month of the year.

3 DATA ANALYSIS TOOLS

To facilitate the sharing of routines and procedures, we decided that the codes should be written in Python for it being open source, easier to share and rich of modules that fit our purposes. It is also easy to port Python codes to MatLab for whoever prefers it. Seismic data can be handled easily using the ObsPy module. Data analysis tools are based on functions and methods from the most common Python modules, such ad NumPy and SciPy.

4 DATA AVAILABILITY

There is still an open debate about the platform through which data are shared. Some propose to create a client on FDSN to download data from the internet. A FDSN client would be open access and whoever wants can access the data without requiring any account nor permission from third party. A client and the use of ObsPy would also make redundant the need to download data on a local machine before handling them, unless a mass download is needed.

On the other hand, there is also a repository at the university of Pisa on which data from the Sardinia site are stored. To access it, an account is required. Access would be granted to any ET collaboration member.

A STANDARD FUNCTIONS

- numpy.fft (beware of normalization)
- scipy.spectrogram
- insert others...