SECOND EUROPEAN CONFERENCE ON EARTHQUAKE ENGINEERING AND SEISMOLOGY, ISTANBUL AUG. 25-29, 2014

AN ADAPTABLE AND MODERN SEISMIC DATA FORMAT INCLUDING DATA PROVENANCE

Lion KRISCHER[1], James SMITH[2], Wenjie LEI[3], Matthieu LEFEBVRE[4], Ebru BOZDAG[5], Norbert PODHORSZSKI[6], and Jeroen TROMP[7]

The widely acknowledged surge in the amounts of data available in seismology comes with a range of interesting challenges. Increases in problem size and complexity already cause existing file formats to be major bottlenecks when striving for more efficient and scalable workflows and simpler data exchange practices. One problem is the fact that existing file formats have been designed for a particular purpose and are not well suited for parallel file systems available in high performance environments. Furthermore, many problems require the incorporation of various types of metadata and data relations which nowadays is usually handled by ad hoc and thus non-exchangeable solutions. A third problem that everyone is aware of but that is usually not dealt with is reproducibility and data provenance. Finally, there is no standardised way to store and exchange derived data like cross correlations from ambient seismic noise analysis or adjoint sources common in full waveform inversions. We propose and hope to establish a new file format largely reusing existing technologies to tackle these problems.

The file format consists of an hierarchical structure stored within a container format like ADIOS (https://www.olcf.ornl.gov/center-projects/adios/) or HDF5 (http://www.hdfgroup.org/HDF5/). The use of these established tools and associated libraries grants parallel read and write capabilities on a large number of systems and, depending on the format, further features like built-in compression algorithms and data check-summing. The containers enable the storage of large amounts of array-like data with efficient access patterns. Waveform and derived data is stored in this structure alongside the metadata and provenance resulting in a single, self-describing file.

Seismological metadata is stored by using the FDSN (http://www.fdsn.org/) StationXML and QuakeML formats directly enabling a large variety of use cases without reinventing the wheel. Associations between waveforms and various components in a QuakeML file are provided by reusing the QuakeML Resource Identifiers. The StationXML blocks amongst other things store information about station locations and instrument responses.

[1] Ludwig-Maximilians-Universität, Munich, krischer@geophysik.uni-muenchen.de

[2] Princeton University, Princeton, jas11@princeton.edu

[3] Princeton University, Princeton, lei@princeton.edu

[4] Dr., Princeton University, Princeton, ml15@princeton.edu

[5] Dr., Géoazur, Valbonne, ebru.bozdag@geoazur.unice.fr

[6] Dr., Oak Ridge National Laboratory, Oak Ridge, pnorbert@ornl.gov

[7] Prof. Dr., Princeton University, Princeton, jtromp@princeton.edu

Provenance information is stored in a domain specific extension for PROV-XML (http://www.w3.org/TR/prov-xml/), the XML representation of W3C PROV, in the context of seismological data processing and generation. We describe the history of data using a process-centered provenance perspective. For example, waveform data in different stages of a processing chain is described as entities whereas the actual processing steps are described as activities using existing entities to generate new ones. The domain specific extension aims to incorporate processing operations common in seismology while being flexible enough to describe workflows that lead to the previously mentioned derived data. We will present examples of processing chains, synthetic waveforms, and derived data described with it and stored in the file format.

We will furthermore present prototype implementations of the format in Python and Fortran/C alongside some benchmarks and real world use cases that demonstrate the possibilities and advantages of the proposed data exchange format.