



Moving towards persistent identification in the seismological community

Javier Quinteros, Peter Evans, Angelo Strollo, Damian Ulbricht, Kirsten Elger, and Roland Bertelmann
Helmholtz-Zentrum Potsdam - Deutsches GeoForschungsZentrum (GFZ), Potsdam, Germany (javier@gfz-potsdam.de)

The GEOFON data centre and others in the seismological community have been archiving seismic waveforms for many years. The amount of seismic data available continuously increases due to the use of higher sampling rates and the growing number of stations. In recent years, there is a trend towards standardization of the protocols and formats to improve and homogenise access to these data [FDSN, 2013]. The seismological community has begun assigning a particular persistent identifier (PID), the Digital Object Identifier (DOI), to seismic networks as a first step for properly and consistently attributing the use of data from seismic networks in scientific articles [Evans et al., 2015]. This was codified in a recommendation by the international Federation of Digital Seismic Networks [FDSN, 2014]; DOIs for networks now appear in community web pages.

However, our community, in common with other fields of science, still struggles with issues such as: supporting reproducibility of results; providing proper attribution (data citation) for data sets; and measuring the impact (by tracking their use) of, those data sets.

Seismological data sets used for research are frequently created “on-the-fly” based on particular user requirements such as location or time period; users prepare requests to select subsets of the data held in seismic networks; the data actually provided may even be held at many different data centres [EIDA, 2016]. These subsets also require careful citation. For persistency, a request must receive exactly the same data when repeated at a later time. However, if data are curated between requests, the data set delivered may differ, severely complicating the ability to reproduce a result. Transmission problems or configuration problems may also inadvertently modify the response to a request.

With this in mind, our next step is the assignment of additional EPIC-PIDs to daily data files (currently over 28 million in the GEOFON archive) for use within the data centre. These will be used for replication and versioning of the data. This will support reproducible, fine-grained citation of seismic waveform data in a consistent fashion. Moreover, we plan to create also PIDs for collections of PIDs, in order to support the citation of a set of many data files with a single identifier.

The technical information describing the instruments used to acquire the data and their location will most probably be also identified with a PID (to a StationXML record) and pointed to from the metadata of the waveform PID. StationXML will also include the DOI of the network for citation purposes.

With all these elements, progress towards reproducibility and better attribution are gained.

References

- EIDA (2016): European Integrated Data Archive (EIDA) . <http://www.orfeus-eu.org/eida/eida.html>
- Evans, P., Strollo, A., Clark, A., Ahern, T., Newman, R., Clinton, J. F., Pedersen, H., Pequegnat, C. (2015 online): Why Seismic Networks Need Digital Object Identifiers. - Eos, Transactions American Geophysical Union, 96. <http://doi.org/10.1029/2015EO036971>
- International Federation of Digital Seismograph Networks (FDSN) (2013): FDSN Web Service Specifications, Version 1.1b, 2013/10/25. <http://www.fdsn.org/webservices/FDSN-WS-Specifications-1.1.pdf>
- International Federation of Digital Seismograph Networks (FDSN) (2014), FDSN recommendations for seismic network DOIs and related FDSN services [WG3 recommendation], <http://doi.org/10.7914/D11596>.