Efficient Access and Manipulation of Big Seismic Data from Disparate Sources

Seismological data recordings have been growing exponentially in the last three decades. For instance, the data archive at the IRIS Data Management Center (DMC) grew from less than 10 Tebibytes in 1992 to greater than 750 Tebibytes today (in 2022). In addition to such big data archives, retrieving and merging data from various disparate seismic sources also creates big data which will enable obtaining higher-resolution seismic images and understanding phenomena such as earthquake cycles. Moreover, recent progress in geosciences with the application of AI/ML using big data has shown the potential of discovering patterns that were not previously recognized. However, aggregating large seismic datasets introduces its own challenges. Some of these challenges arise from the fact that many data centers have their own way of distributing data, and the format of the data and metadata are different in many cases.

The objective of this investigation is the development of data access and manipulation tools for retrieval, merging, processing, and the management of big seismic data from disparate seismic data sources. We develop a free, open-source, direct data accessing, gathering, and processing software toolbox for disparate sources using Python. Aggregating data from different data centers will enable us to investigate the seismic structure beneath a region of interest at a higher resolution by merging the seismic databases. Such a merged dataset can be applied on studies around the boundaries between countries if those countries have different networks. One boundary region of geologic interest to exemplify the benefits of aggregated seismic datasets from different networks in two countries is the southern side of the Rio Grande Rift including the bordering areas between the US and Mexico. Previous more detailed seismic studies on Rio Grande were conducted mostly on the US side of the Rift Valley.