Geophysical Research Abstracts
Vol. 16, EGU2014-10972, 2014
EGU General Assembly 2014

EUDAT strategies for handling dynamic data in the solid Earth Sciences

Alberto Michelini (1), Peter Evans (2), Mark Kemps-Snijder (3), Jani Heikkinen (4), Justin Buck (5), Jozef Misutka (6), Sebastian Drude (7), Massimo Fares (1), Claudio Cacciari (8), and Giuseppe Fiameni (8)

(1) Istituto Nazionale Geofisica e Vulcanologia, Centro Nazionale Terremoti, Roma, Italy (alberto.michelini@ingv.it, 0039 06 5041303), (2) German Research Center for Geo-sciences, (3) Royal Netherlands Academy of Arts and Sciences, (4) CSC - Finnish IT Center for Science, (5) British Oceanographic Data Center, (6) Charles University in Prague, (7) Max Planck Institute for Psycho-linguistics, (8) CINECA

Some dynamic data is generated by sensors which produce data streams that may be temporarily incomplete (owing to latencies or temporary interruptions of the transmission lines between the field sensors and the data acquisition centres) and that may consequently fill up over time (automatically or after manual intervention). Dynamic data can also be generated by massive crowd sourcing where, for example, experimental collections of data can be filled up at random moments. The nature of dynamic data makes it difficult to handle for various reasons: a) establishing valid policies that guide early replication for data preservation and access optimization is not trivial, b) identifying versions of such data – thus making it possible to check their integrity – and referencing the versions is also a challenging task, and c) performance issues are extremely important since all these activities must be performed fast enough to keep up with the incoming data stream. There is no doubt that both applications areas (namely data from sensors and crowdsourcing) are growing in their relevance for science, and that appropriate infrastructure support (by initiatives such as EUDAT) is vital to handle these challenges. In addition, data must be citeable to encourage transparent, reproducible science, and to provide clear metrics for assessing the impact of research, which also drives funding choices.

Data stream in real time often undergo changes/revisions while they are still growing, as new data arrives, and they are revised as missing data is recovered, or as new calibration values are applied. We call these “dynamic” data sets, DDS.

A common form of DDS is time series data in which measurements are obtained on a regular schedule, with a well-defined sample rate. Examples include the hourly temperature in Barcelona, and the displacement (a 3-D vector quantity) of a seismograph from its rest position, which may record at a rate of 100 or more samples per second. These form streams of data, and a complete data set may contain many streams. It is important to distinguish between "observation time" (or measurement time, MT), at which a measurement is made and "access time" (or state time, ST, or ingestion time), at which data arrived in this database system. Granularity in time is the third ingredient and it mainly refers to reproducibility: is there a requirement to distinguish between states a few seconds, minutes, days apart? This time may be as small as the time between successive samples. Granularity relates to accountability.

We present a possible solution which stems on representing each observation in a dynamic data stream by a point in the bi-temporal space defined by (MT, ST) corresponding to the time at which the data arrived and the time to which it refers. Features of dynamic data such as latency and gaps can then be visualised: delays in arrival of the data (latency) move observations down from the line labelled “real time”, while missing values (gaps) cause the sequence of observations to jump to the right. Revisions of values at one or more time appear below the original values.