Working Groups
Dynamic Data

Dynamic Data at the Second EUDAT Conference

Overview

During the service building process and roll out of the Safe Replication (B2SAFE) service dynamic data has been a challenging subject. It is difficult to keep consistency between data objects, which are eligible to change and are replicated in a distributed environment. This use case is prominent within the seismology community (EPOS[1]) dealing with sensor-generated data in earthquake sensitive areas across Europe and data streams that are generated by mobile devices at unpredictable times and in unpredictable order (CLARIN[2]). Dynamic data is a broad subject, not only from sensor-generated data, but is seen within communities who have to deal with many unstructured and independent non-scientists (e.g. citizen scientists or crowdsourcing). Dynamic data is one of the working group interest fields. During the 2nd EUDAT Conference, 28-30 October 2013 – Rome, Italy[3], the New Services track included a specific session on Dynamic Data presenting both the outcome of the Dynamic Data working group meeting in Barcelona (September 2013) and the EPOS and CLARIN communities use cases on dynamic data.

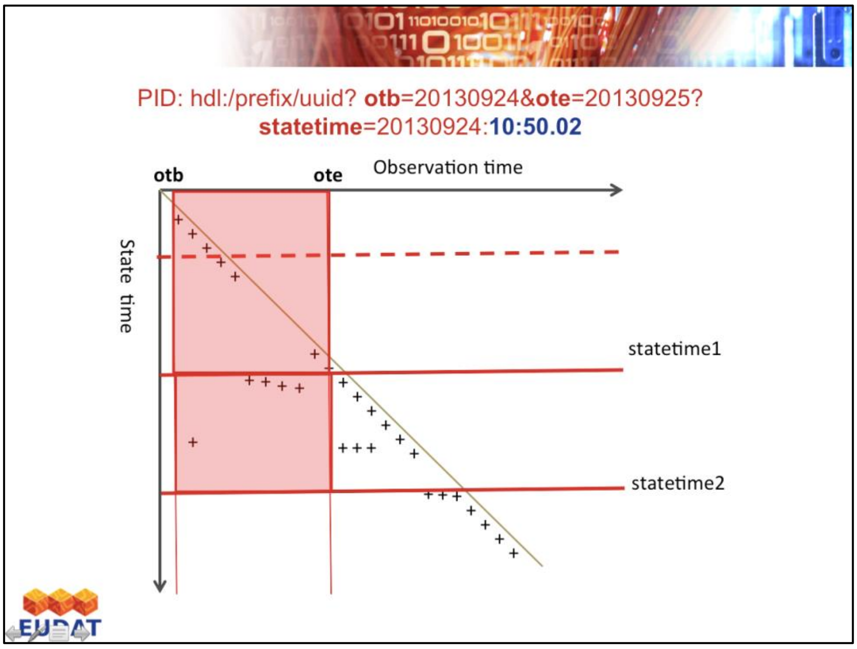
EPOS (European Plate Observing System) use case

The European Plate Observing System (EPOS) collaboration brings together the European seismology community to come to a common vision and approach to enable innovative multidisciplinary research to better understand the physical processes controlling earthquakes, volcanic eruptions, unrest episodes, tsunamis, tectonics and earth surface dynamics. The goal is to establish a long-term plan to facilitate the integrated use of data, models and facilities from existing, and new distributed research infrastructures. The EPOS community is dealing with data generated by sensors situated at earth sensitive locations across Europe. Data is transferred via phone lines or via satellite or radio links. These transmission methods have a level of uncertainty in which data fragments are not received in the right order, sensor frames can be delayed between minutes, hours or days or are never received. The challenge is to keep consistency between these changing data objects in a distributed environment and methods to enable reproducible science. An important aspect to enable reproducible science is to be able to track which version of a data object has been used to generate scientific results. With data objects, which are eligible, to change in time you need a method to identify a version of an object at time X and a method to identify a time frame within an object. This subject was discussed extensively within the Dynamic Data working group at the Barcelona meeting. The result from the working group meeting was to use a bi-temporal scheme to identify a version of a data object with two separate timelines: observation time and state time. The observation time indicates the time frame of the event/measurement, described with a begin and an end time. The state time describes the state or version of an object at a time. The relationship between observation and state time and the difference between versions of the selected observation time between “otb-ote” at state time 1 and 2 is explained in the Figure below.

[1] http://www.epos-eu.org/

[2] http://www.clarin.eu

[3] http://www.eudat.eu/parallel-track-iv-new-services-overview



PID: hdl:/prefix/uuid? **otb**=20130924&**ote**=20130925?
statetime=20130924:**10:50.02**

Figure 1 - Diagram describing the difference between observation and state time

CLARIN MPI-PL The Language Archive use case

The CLARIN MPI-PL *The Language Archive*[4] use case looks at the Dynamic Data challenge from a different view-point. The Language Archive is a unit of the Max Planck Institute for Psycholinguistics concerned with digital language resources and tools. It provides a large data archive holding resources on languages worldwide. The service is open to everyone to store good quality data. In the current age with mobile devices, data is easily generated. The challenge is to predict, store, manage and curate this vast growing data volume, to track intellectual property, to ensure data privacy and to engage a diverse growing user population (e.g. crowd sourcing). Engaging thousands of subjects in tests by using mobile devices means that data from participants will come in at unpredictable moments and in an unpredictable order, nevertheless researchers want to start using the results for calculating evidence and obviously using them for publications. Similar to the case in EPOS, the concern is thus how to cite to a data matrix that is being filled at random uncontrolled moments. Also here observation and state time are different, since mobile devices could be off-line while an experiment is carried out etc.

Dynamic Data Discussions & Conclusions

The main discussions focused on the terminology used to identify versions of dynamic data objects and about the feasibility of supporting this within persistent identifier and repository systems, about how to handle the vast growing data volumes and about intellectual property rights.

The proposed bi-temporal scheme (e.g. observation and state time) for data objects appears to be a proper solution for tackling scientific reproducibility issues, and for data analysis carried out on real-time data. During the discussions some similarities were drawn with the spatial science domain to identify location areas.

Consensus on the terminology used to define the states is important to enable referencing and accessing data objects on basis of a bi-temporal scheme. It is recommended to interact with the RDA data citation working group.

[4] http://tla.mpi.nl/

In the discussion about the vast growing data volumes, the challenge is to manage and to store the data volumes and to be reservedly in destroying data. The current approach is to store all data and not to delete data objects, because the value of the data in the future is hard to predict.

In the crowdsourcing use case there was considerable discussion on how to track intellectual property. In general, IPR is handled via informed consent. But it is questionable if people providing data have a full understanding of the meaning and consequences of informed consent.

Further Information

For more information see the Dynamic Data web Section http://www.eudat.eu/dynamic-data or contact:

- Alberto Michelini, INGV [alberto.michelini[at]ingv.it] or
- Sebastian Drude, MPI-PL [Sebastian.Drude[at]mpi.nl]