Dynamic Data at the 3rd EUDAT conference- *Bringing Data Infrastructures to Horizon 2020*

24 & 25 September 2014, Amsterdam, the Netherlands

The Dynamic Data Session was composed of the following three presentations

- Is Dynamic Data a Special case? – Daan Broeder, MPI for Psycholinguistics
- Handling Dynamic Data from Sensors – Peter Danecek INGV
- Dynamic Data in the Humanities – Marc Kemps-Snijders, Meertens Institute

As an introduction Daan Broeder gave an overview of the Dynamic Data activities in the EUDAT project: the WG meetings in Barcelona and Rome. Also the EUDAT 2020 proposal foresees a special Dynamic Data WG and in EUDAT2020 workflow and directive aspects of Dynamic Data play a prominent role.

Dynamic Data, a Special case? This presentation gave on overview of the Dynamic Data (DD) terminology. The use-cases until now considered in EUDAT are sensor data from earth sciences and crowd sourcing, corpus collection and surveys from the Humanities. The use-cases seemed to indicate that incomplete DD converging to a stable status are an important aspect, but later discussions showed that that was not necessarily the case. The challenge with DD for EUDAT is especially in the replication scenario where it is difficult to update already replicated DD. If the DD sets are not too big, representing the DD by a series of versions seems to be sufficient to solve most questions. In the case of big DD sets, the use of stored deltas, the differences between subsequent versions, seems more efficient. In both cases the relations between the versions must be maintained, different mechanisms exist as using metadata, PIDs or virtual collections. The use of deltas can also enable efficient processing if linear transformations are required. Referring to DD using Handle type of PIDs can be achieved using Handle System templates.

Handling Dynamic Data from Sensors: Peter Danacek presented the DD that result from the use of remote sensors for data acquisition in the seismic monitoring network. There is a need for real-time processing and early availability of data to the scientists. The monitoring network is especially challenging: large number of sensors, use of compression, small data fragments (30M), high varying data rate (200 updates/s), out of order package arrival. There is a need to solve accountability, traceability, reproducibility and citability for two related problems: (near-) real time processing and data delivery from a query system. There was investigation into temporal databases and version control systems and, in the context of the EUDAT DD WG, the bi-temporal concept was introduced that distinguishes between observation and state time. Open questions are the use of PIDs (discussion follows on the possibilities of the Handle System) and how to provide for versioning of Data Objects.

Dynamic Data in the Humanities Marc Kemps-Snijders presented the concept of Dynamic Data in the Humanities as it occurs in the Nederlab project that collects digital Dutch written sources from several providers, in total 37M documents and 12000 M words, in a single Virtual Research Environment (VRE). He refers to a few examples from this project. A data curation example refers to a similar concept as the bi-temporal concept mentioned by Peter Danecek, where data produced in the distant past gets archived asynchronously and must be curated also at unpredictable times, resulting in different versions for the data-set in the archive or DB. The use of “editions” (compilations and

aggregations of existing data) in the humanities provides an additional challenge. From the Nederlab example he concludes that efficient versioning is the key for DD management by maintaining version history, use of appropriate time-stamps and a clear policy for phasing out data. Data integrity may be compromised when dealing with heterogeneous collections. Especially in the case of overlapping collections when for the apparent same data item different data enrichment processes may have been applied. Special harmonization tools are needed. Discussion follows about the use of EUDAT services, how to archive, how to maintain version history.

Session Outcomes:

- **that non-converging dynamic data should definitely be a use case**
- **The use of deltas and PID part identifiers are important mechanisms for efficiency.**