The Challenge of Handling Big Data Sets in the Sensor Web

Christian Autermann
c.autermann@52north.org

May 2, 2016
• How can big heterogeneous spatio-temporal datasets be organized, managed, and provided to Sensor Web applications?

• How can views on big data sets and derived information products be made accessible in the Sensor Web?

• How can big observation data sets be processed efficiently?
Accessibility

• How can views on big data sets and derived information products be made accessible in the Sensor Web?

• What are typical request scenarios of observation data for search, download, visualization and processing?

• Are current Sensor Web standards capable of and suitable for handling massive observation data sets?

• Which service standards are appropriate?
  • SOS natural choice, but what about WPS, WCS/WCPS?

• Which conceptual models and encodings, e.g. O&M or NetCDF, are appropriate?
Processing

- How does the underlying storage structure influence performance?
- How does the WPS handle situations in which transferring datasets is hard to achieve?
  - Can the WPS be used as a Rich-Data-Interface for big observation databases?
- How can predefined, parameterized or even interactive analyses be realized?
- How could a query language that enables on-demand analysis of time series data look like?
  - Similar to what the WCPS Language does for coverages
- How could a combined analysis of multiple datasets of different origins be accomplished with such high volumes of data?
Storage

- Data is for the most part organized as files (NetCDF/CSV/TSV) each containing a fixed time interval of multiple sensors
  - allows for fast insertion times
  - expensive to query a single series over an extended amount of time.

→ Requesting a subset of a timeseries vs. requesting the measurements of a phenomenon across multiple sensors at a single point in time

→ How can queries in both request dimensions be realized?

→ How can this be accomplished while retaining fast insertion times?

- Are existing database technologies (e.g. distributed, array or object databases) applicable?
- Does there even exist a solution that offers acceptable trade offs between the different requirements?
Outlook

• Lot of questions, no real answers yet
• Issues should be resolved within the next 3 years
• At least in part in cooperation with AWI, using their data as a case study