Data flow, harmonization, and quality control

Brenner Silva, Philipp Fischer, Sebastian Immoor, Rudolf Denkmann, Marion Maturilli, Philipp Weidinger, Steven Rehmcke, Tobias Düde, Norbert Anselm, Peter Gerchow, Antonie Haas, Christian Schäfer-Neth, Angela Schäfer, Stephan Frickenhaus, Roland Koppe, and the Computing and Data Centre of the Alfred-Wegener-Institute

Earth system cyberinfrastructures include three types of data services: repositories, collections, and federations. These services arrange data by their purpose, level of integration, and governance. For instance, registered data of uniform measurements fulfill the goal of publication but do not necessarily flow in an integrated data system. The data repository provides the first and high level of integration that strongly depends on the standardization of incoming data. One example here is the framework Observation to Archive and Analysis (O2A) that is operational and continuously developed at the Alfred-Wegener-Institute, Bremerhaven. A data repository is one of the components of the O2A framework and much of its functionality depends on the standardization of the incoming data. In this context, we focus on the development of a modular approach to provide the standardization and quality control for the monitoring of the near real-time data. Two modules are under development. First, the driver module transforms different tabular data to a common format. Second, the quality control module that runs the quality tests on the ingested data. Both modules rely on the sensor operator and on the data scientist, two actors that interact with both ends of the ingest component of the O2A framework (http://data.awi.de/o2a-doc). We demonstrate the driver and the quality control modules in the data flow within Digital Earth showcases that also connect repositories and federated databases to the end-user. The end-user is the scientist, who works closely in the development approach to ensure applicability. The result is the proven benefit of harmonizing data and metadata of multiple sources, easy integration and rapid assessment of the ingested data. Further, we discuss concepts and current development that aim at the enhanced monitoring and scientific workflow.

Computing and Data Centre of the Alfred-Wegener-Institute: Stephan Frickenhaus (speaker)