September 2019

Newsletter of the Digital Earth Project

Contributions of the Alfred Wegener Institute to Digital Earth

This newsletter presents the project activities of the Alfred Wegener Institute within Digital Earth.

Contact: Stephan.Frickenhaus@awi.de

Data Infrastructure Developments for Digital Earth

As an important technical pillar of Digital Earth AWI computing centre provides data management and cloud processing services to the project partners. We develop project specific extensions to the AWI data flow framework O2A (Observation to Archive). Sensor registration in O2A will support a flexible handling of sensors and their metadata, e.g. for the Digital Earth showcases, methane and soil moisture measurements are in focus for smart monitoring designs and for the access to data in near real time (NRT). Furthermore, data exploration is supported by a rasterdata manager service that can be easily coupled in user´s data workflows with other data sources, like NRT sensor data. In the following we give more details on O2A, its components and concepts.

O2A Data Flow Framework

Over the last two decades the Alfred Wegener Institute (AWI) has been continuously committed to develop and sustain infrastructure for digital science including coherent discovery, view, dissemination and archival of scientific data and related information. In order to address the increasing heterogeneity of research platforms and respective devices and sensors along with varying project-driven requirements, we built and building a generic and cost-effective virtual research environment: the O2A Data Flow Framework.

The O2A architecture comprises of several seamlessly integrated components focusing on special requirements within a data flow. Even the basic O2A infrastructure already exists we are developing, extend and improve these components and underlying technology continuously.

Figure 1 illustrates the different components of the O2A Data Flow Framework:

- **SENSOR** supports description of research platforms, devices and sensors with scientific relevant information. The metadata described in SENSOR makes the entry point for O2A.
- **INGEST** consists of two parts: low-volume near real-time data from research platforms like current positions of vessels or buoys and full resolution high-volume data.
- **DASHBOARD** allows creating highly user-customized dashboards visualizing data with plots or maps from the INGEST component for monitoring purposes.
- **WORKSPACE** provides fast online storage for project-based scientific data and close compute solutions for high performance data science.
- **ANALYTICS** is comprised of GIS solutions, raster data access, support for Jupyer notebooks and technology like Hadoop stack.
- **PORTAL** summarizes portal solutions, which integrate data and information from the other O2A components as well as from O2A-external providers and provide interactive access e.g. to curated maps.
- **REPOSITORIES** can be distinct in PANGAEA data publisher operated by AWI and MARUM for archived and published datasets and our institutional publication repository EPIC.

Editor Concept

For SENSOR metadata editing, O2A follows an editor concept to distribute effort and knowledge in the community. Each institute using SENSOR nominates a person who is responsible for metadata quality, and promoting best practices and tutorials within the respective institute.

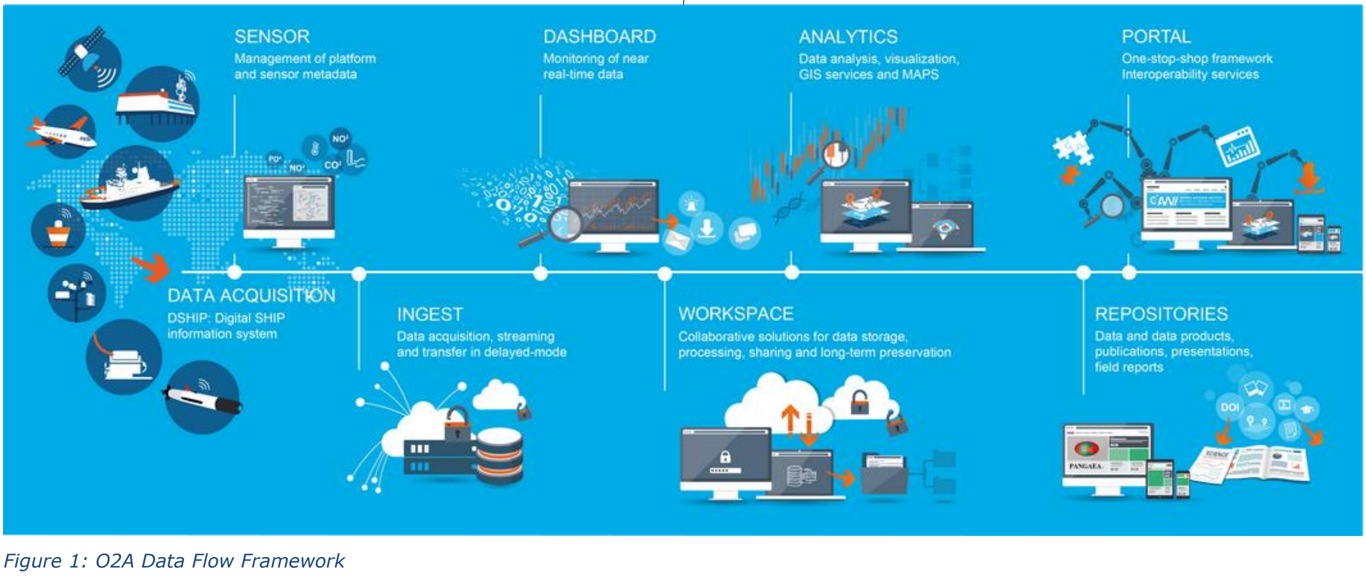
SENSOR
Management of platform and sensor metadata
DASHBOARD
Monitoring of near real-time data
ANALYTICS
Data analysis, visualization, GIS services and MAPS
PORTAL
One-stop-shop framework Interoperability services
DATA ACQUISITION
DSHIP: Digital SHIP information system
INGEST
Data acquisition, streaming and transfer in delayed-mode
WORKSPACE
Collaborative solutions for data storage, processing, sharing and long-term preservation
REPOSITORIES
Data and data products, publications, presentations, field reports
DOI
PANGAEA

Figure 1: O2A Data Flow Framework

Documentation and Support

Information and documentation on O2A components and usage is available in our wiki (https://data.awi.de/o2a-doc), videos and tutorials are available on YouTube (https://data.awi.de/o2a-video). There are also code examples on using O2A web services in GitHub (https://data.awi.de/o2a-code).

Links

- sensor.awi.de
- dashboard.awi.de
- maps.awi.de
- data.awi.de
- pangaea.de

Data Exploration on near real-time data

Using the O2A framework, metadata about platforms, devices and sensors are described in SENSOR. The INGEST component is based on these sensor descriptions and supports several protocols (e.g. FTP) and data formats (e.g. NetCDF) of datasets. INGEST stores harvested data in archive systems and on fast online storage systems. Generalized quality assessment and control procedures can be applied automatically during the INGEST process (download, store, process). Tests like manufacturer range or operation range are bound to specifications in SENSOR. Their implementation closely follows the formulations of the ARGO system.
For numerical near real-time data, O2A provides REST-based web services (https://dashboard.awi.de/data-xxl/) for data access. Data is stored in a PostgreSQL database partitioned by platform and time including simple quality flags. For performance, aggregate statistics are calculated by default, offering fast access to e.g. minutely or hourly averaged values. Examples in R and Python with Jupyter notebooks can be found on GitHub (https://github.com/o2a-data/o2a-data-dws)

Data Exploration for raster data

Within the Helmholtz Data Federation (HDF) project, we are developing solutions and extensions for the raster data manager rasdaman (https://www.rasdaman.com/). Selected data streams from satellites (e.g. precipitation raster data covering Germany) are described in SENSOR and fed into rasdaman via the INGEST component and according processing scripts.
Rasdaman provides fast raster analytics capabilities by supporting OGC standardized web services like web coverage service (WCS). Tools like extended metadata handling, 4D data extraction by profiles and polygons, interpolation and online data computing are available and complete the GIS-based Map Services visualization tools.
In Digital Earth these raster services are prepared for precipitation data and used in the Exploration Framework developments.

Cloud Infrastructure

Also based on the HDF project, AWI develops cloud solutions for data storage and computing. Online file storage is realized on basis of DELL/EMC Isilon systems complemented with a hierarchical tape storage system. Since it is a good idea to do data processing by computing next to the data itself where possible, we are putting to test a self-service marketplace (https://marketplace.awi.de) for virtual appliances based on Docker containers and VMware with fast access to the stored data. For direct data access we are running a JupyterHub instance (https://jupyterhub.awi.de) supporting Python and R notebooks.

Access to the cloud infrastructure is currently offered prototypical upon request.

O2A services and cloud infrastructure are also available on board of Polarstern and used in very large international projects like MOSAiC.

Governance and Digitalization Strategies

For the first time, Digital Earth has greatly enhanced the collaborative perspective of data science, as has happened in data management within the MOSES project and its infrastructures, but also in the modeling activities within ESM. It was noted that the multitude of collaborative activities reflected in the participation of many PIs in various working groups and cross-cutting subjects requires a concerted approach of governance development. This has recently been recognized by the management board of the research field and led to the establishment of the Earth and Environment Data Hub. The approach of Seed Groups in Digital Earth is considered at a more general level, focusing on data infrastructures and services synergistically, i. e., with the goal to streamline developments and to share services. In particular, this is regarded as part of the common digitalization strategy that is an important pillar to our common research program in POF IV. To consolidate the different working groups, DE proposes to establish "Collaborative Governance" as another cross-cutting theme in the Data Hub.
The emerged and to be further consolidated structures have facilitated writing a proposal for extending DE and the Earth System Modeling project ESM, which is funded within the same framework.
Digitalization requires a coordinated and joint effort in Helmholtz. The distributed Data Hub will serve to implement shared data management and analysis platforms in a synergistic way, bridging observation data, modeling data, metadata registration, and data publication. In this context, Digital Earth interfaces with ESM.

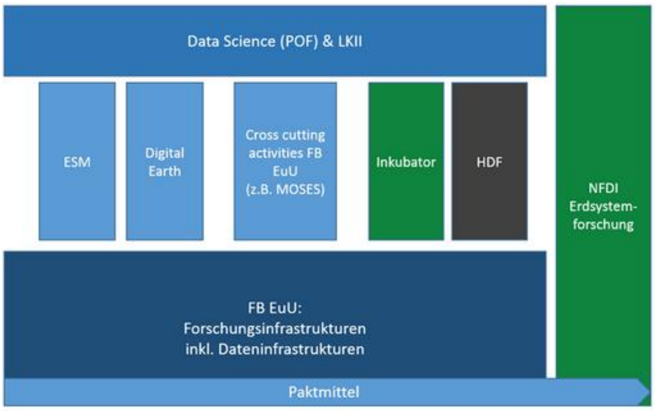
Data Science (POF) & LKII
ESM
Digital Earth
Cross cutting activities FB EuU (z.B. MOSES)
Inkubator
HDF
NFDI Erdsystem-forschung
FB EuU: Forschungsinfrastrukturen inkl. Dateninfrastrukturen
Paktmittel

Figure 2: Schematic view of digitalization contexts

The management of modeling data of ever increasing storage volume requires new approaches to grant access for re-use of (frontier) simulations to a wide community of (non-modeling) scientists. Digital Earth will provide a web-based front-end to directly interact with the model data (regardless of size), i. e., as "virtual campaigns" by extracting essential variables from archived model data interactively (including the use of observational simulators). The required data services will be implemented in the extended ESM workpackages. As depicted above, all digitalization projects of the research field (including the

Data Hub, ESM, and DE) will contribute to the community effort of creating a National Research Data Infrastructure for Earth System Sciences in Germany (NFDI4Earth). The outcome of the collaborative service development for DE and ESM will be available to all, making novel model data available to a wide range of users. Furthermore, the Seed Groups of Digital Earth are strongly coupled to the recently established Data Hub working groups in the research field. These cover distributed data processing, data publication/persistent identifiers, metadata profiles, viewer technologies (data portals), as well as dataflows/sensor management. Group representatives are also involved in the Helmholtz Incubator "Information and Data Science", fostering collaborations in pilot projects like, e. g., Helmholtz Analytics Framework, Pilot Lab Exascale Earth System Modelling, Uncertainty Quantification, and Artificial Intelligence in Cold Regions.

Recent and upcoming events at AWI

A networking workshop between AWI and UFZ took place 22.07.2018 at AWI computing centre, focusing on the following subjects:

Use of AWI´s Sensor management "as a service" by UFZ. A special test run is foreseen for the upcoming MOSES campaign Elbe 2020. The sensor component of the O2A framework will be tested for actually applied MOSES sensors in combination with and linkage to the actual data storage system at UFZ. This is enabled via O2A's APIs, registry, and metadata input. The preparation phase for this test run is scheduled until end of 2019.
Consequently, UFZ will appoint a Sensor "chief editor" responsible for all UFZ-assigned sensor components for terrestrial data acquisition, especially for the combination of sensor network and collection of soil moisture and cosmic ray neutron sensing data.
In this context, a suitable method of data visualiziation of the 2020 campaign will be elaborated in cooperation with the participating Helmholtz centers. Amongst other things, this will be discussed within the EuU pilot project by the assigned science-infrastructure tandems of UFZ, HZG, AWI, and GFZ who are responsible for MOSES data management.

Furthermore, the necessity of a cross-sectional working group for governance issues has been pointed out for a common understanding of Helmholtz center-specific internal procedures and to establish a common understanding of administrative methods, procedures and business models. This would include accounting modalities, financial flows, interpretation and application of public service regulations, and legal aspects. Hence, experts from the administrative departments of each center should be involved. This suggestion will be discussed and decided during the next EuU pilot meeting of the tandems and data management working group at 20.09.2019 in Berlin.

A brief outlook of the development or rather design of the EuU pilot and its DataHUBs and cross-sectional topics should be given during the upcoming DE interim meeting for the benefit of a wider audience.

Hands-on workshop on Data Flow: DATA INGEST and SENSOR, December 4-5, 2019

Data Science Analytics with Sensor/NRT/Rasterdata: cloud-based access and processing, Spring 2020 – actual date will be announced