

# Implementing FAIR in a Collaborative Data Management Framework

**Angela Schäfer<sup>1</sup>**, Norbert Anselm<sup>1</sup>, Janik Eilers<sup>1</sup>, Stephan Frickenhaus<sup>1</sup>, Peter Gerchow<sup>1</sup>, Frank Oliver Glöckner<sup>1</sup>, Antonie Haas<sup>1</sup>, Isabel Herrarte<sup>1</sup>, Roland Koppe<sup>1</sup>, Ana Macario<sup>1</sup>, Christian Schäfer-Neth<sup>1</sup>, Brenner Silva<sup>1</sup>, and Philipp Fischer<sup>1</sup>

<sup>1</sup>Alfred Wegener Institute - Helmholtz Centre for Polar and Marine Research, Computing and Data Centre, Bremerhaven, Germany

EGU2020-19631, presented on 06 May 2020

<https://doi.org/10.5194/egusphere-egu2020-19631>

EGU General Assembly 2020

© Author(s) 2020. This work is distributed under the Creative Commons Attribution 4.0 License.



# Implementing FAIR in a Collaborative Data Management Framework



**Abstract** Today's fast digital growth made data the most essential tool for scientific progress in Earth Systems Science. Hence, we strive to assemble a modular research infrastructure comprising a collection of tools and services that allow researchers to turn big data into scientific outcomes.

Major roadblocks are (i) the increasing number and complexity of research platforms, devices, and sensors, (ii) the heterogeneous project-driven requirements towards, e. g., satellite data, sensor monitoring, quality assessment and control, processing, analysis and visualization, and (iii) the demand for near real time analyses.

These requirements have led us to build a generic and cost-effective framework **O2A (Observation to Archive and Analysis)** to enable, control, and access the flow of sensor observations to archives and repositories.

By establishing O2A within major cooperative projects like **MOSES** and **Digital Earth** in the research field Earth and Environment of the German Helmholtz Association, we extend research data management services, computing powers, and skills to connect with the evolving software and storage services for data science. This fully supports the typical scientific workflow from its very beginning to its very end, that is, from data acquisition to final data publication.

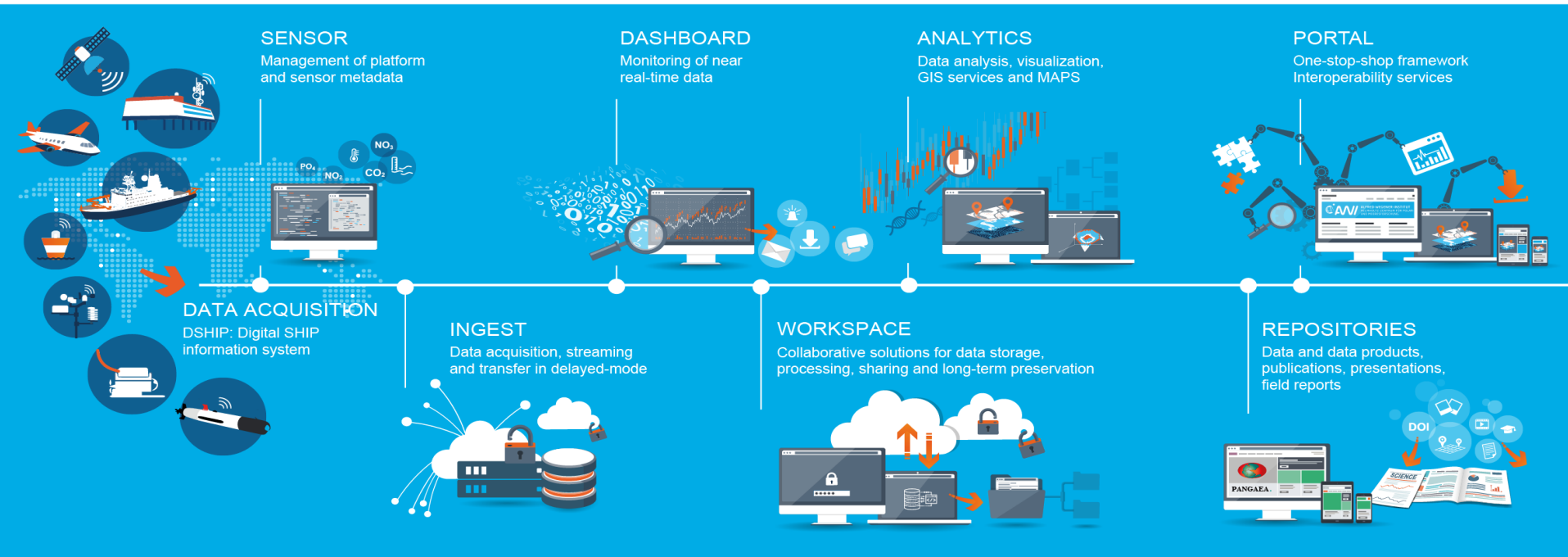
The key modules of O2A's digital research infrastructure established by AWI to enable Digital Earth Science are implementing the **FAIR** principles:

- **Sensor Web**, to register sensor applications and capture controlled meta data before and alongside any measurement in the field
- **Data ingest**, allowing researchers to feed data into storage systems and processing pipelines in a prepared and documented way, at best in controlled NRT data streams
- **Dashboards**, allowing researchers to find and access data and share and collaborate among partners
- **Workspace**, enabling researchers to access and use data with research software in a cloud based virtualized infrastructure that allows researchers to analyze massive amounts of data on the spot
- **Archiving** and **publishing data** via repositories and Digital Object Identifiers (DOI).

# O2A - Data Flow Framework



## from Observation to Archives and Analysis



AWI's generic and sustainable framework enables data flow and management of sensor observations to archives. This framework builds upon international standards for metadata and data interoperability and is meant to assist scientists in developing enhanced data products while implementing the **FAIR** principles



[SENSOR.awi.de](https://sensor.awi.de) creates an authoritative repository for managing platform, device and sensor-related information and data according to the principle

**Metadata first !**

- **Autocompleting of station metadata** for German research vessels using [PANGAEA](#) expedition catalogue
- Metadata **versioning** including minting of [PIDs for individual versions](#) along with citation
- [Provenance metadata](#) for repositories are linking datasets to sensor description, under [event information handles](#)
- Management and **discovery** of standard sensor-specific information (manufacturer information, inventory number, history of actions, precision, photos, etc.) and **data-specific information** (calibration reports, physical characterization of individual sensors, etc.)
- Bookkeeping of **ownership and affiliation** of devices
- **Machine-friendly access** to metadata ([OGC](#) compliant output standards)
- Sensor properties for **monitoring of data streams** in near real-time ([DASHBOARD](#))
- **Re-use** in portals and national and international service infrastructures and [data.awi.de](#)

## Parent Device Type

[Buoy \(617\)](#)  
[Mooring \(259\)](#)  
[Station \(129\)](#)  
[Pack\\_Ice \(119\)](#)  
[Vessel \(78\)](#)

[See more...](#)

## Subdevice Type

[temperature sensor \(200\)](#)  
[MicroCAT \(156\)](#)  
[meteorological package \(79\)](#)  
[CTD \(61\)](#)

[See more...](#)

## Expeditions

[M160 \(15\)](#)  
[moses\\_stern1\\_littorina\\_büs\\_hel \(5\)](#)  
[moses\\_stern1\\_littorina\\_hel\\_büs \(5\)](#)  
[moses\\_stern2\\_uthoern\\_cux\\_hel \(4\)](#)  
[moses\\_stern2\\_uthoern\\_hel\\_bhv \(4\)](#)

[See more...](#)

## Parameters

[temperature \(696\)](#)  
[latitude \(548\)](#)  
[longitude \(548\)](#)  
[date and time \(465\)](#)  
[pressure \(320\)](#)

[See more...](#)

## Action type

[Deployment \(873\)](#)  
[Information \(295\)](#)  
[Total failure \(259\)](#)  
[Recovery \(193\)](#)  
[Mount \(130\)](#)

[See more...](#)

## Action

[PS122-1 \(36\)](#)  
[Initial configuration \(32\)](#)

Select some filters.

Sort by relevance

Sort alphabetically

Has no Parent: ☐

Info	Long Name	Short Name	URN	Item State	Tools
	Temperature, Water	SBE38	vessel:heincke:tsg:sbe38	public	↑
	Conductivity-Temperature-Depth Probe	underway_ctd_240	vessel:polarstern:underway_ctd_240	public	↑
	temperature sensor	107er_071cm	station:bayelva:cr1000_soil09:107er_071cm	public	↑
	Air temperature and relative humidity	HMP	laboratory:mcan2018:majcs:hmp	public	↑
	Humidity and Temperature Sensor HMT333	HMT333	aircraft:polar5:hmt333	public	↑
	ibutton temperature	Dart_Temps_18	pack_ice:ice_ps:dart_temps_18	public	↑
	ibutton temperature	Dart_Temps_19	pack_ice:ice_ps:dart_temps_19	public	↑
	ibutton temperature	Dart_Temps_20	pack_ice:ice_ps:dart_temps_20	public	↑
	ibutton temperature	Dart_Temps_6	pack_ice:ice_ps:dart_temps_6	public	↑
	Infrared temperature	ts_3285_UCB	pack_ice:ice_ps:flux_tower_12m_ucb:ts_3285_ucb	public	↑
	Temperature Sensor No 1 with Magnet for Logger DK 312 No 3	T_Sensor_DK312_HSVA_3_1	vessel:polarstern:t_logger_dk312_hsva_3:t_sensor_dk312_hsva_3_1	public	↑
	bottom temperature recorder	Tpop_60	mooring:tpop_60	public	
	bottom temperature recorder	Tpop_67	mooring:tpop_67	public	
	bottom temperature recorder	Tpop_72	mooring:tpop_72	public	
	bottom temperature recorder	Tpop_57	mooring:tpop_57	public	

Showing 1 to 15 of 1517 entries

« 1 2 3 4 5 6 7 8 9 »



A **raw data ingest framework** enables near real-time transfer of data collected from ship-mounted sensors, buoys, moorings, etc. into data storage solutions on land.

- Describe your sensor and sensor output data via [SENSOR.awi.de](https://sensor.awi.de)
- Copy your sensor data (+/-internet connection) on central storage while using **conventions of file- and directory name and formats**. The data will be imported to AWI's central temporary disk space.
- Archiving these sensor data to repository [PANGAEA](https://pangaea.de) happens while your new sensor data will be checked against the file- and directory-naming convention you described in [SENSOR.awi.de](https://sensor.awi.de) and will receive a DOI.

The **NRT INGEST software** can periodically download your data from remote devices and store it on a central storage. The software installed at AWI collects e.g. remote sensing products from landside servers to the central storage/NRT-DB at AWI computing center.

- **Recipes** can be defined to gather your sensor's data. These can be scheduled to run at certain time intervals.
- There are **defined and supported protocols** to download data from remote servers for ingestions (FTP/FTPS/SFTP; HTTP; IMAP; RSYNC)
- **Recipes for postprocessing routines** can be defined via scripts to be executed after each complete file download. Scripts need to be installed on the INGEST software's server.
- **Automation of download** can be added in [SENSOR Web](https://sensor.awi.de) to your described sensor via Contact Form for data ingest in near real-time.



## Spatial Data Formats

speciation are needed for the automated dataflow of spatial data products: O2A GeoCSV, GeoTIFF, NetCDF.

## NRT Data Format

- INGEST and DASHBOARD Web Services need a **common data format** to exchange time-series data: SENSOR's parameter codes (URNs, Uniform Resource Names) for the data columns in the files.
- The data is stored in a defined tabular plain-text
- If a sensor provides a binary or plain-text format, a **driver (converter)** will have to be **programmed to convert the data to the NRT Data Format**.
- A number of drivers is already available

## Archiving Directory Structure

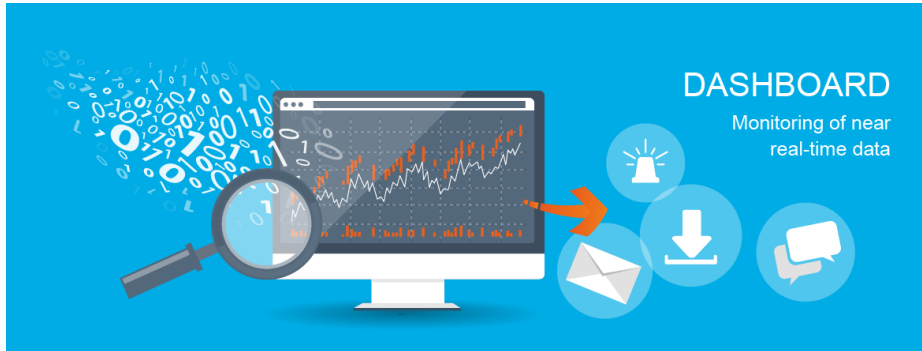
O2A data structure is defined by the metadata description in [sensor.awi.de](http://sensor.awi.de).

## Quality flagging

[EGU2020-15961 Automatic quality control and quality control schema in the Observation to Archive by Brenner Silva et al.](#)

- The INGEST performs automatic quality control to deliver quality-flagged data in the DASHBOARD.
- The automatic quality control requires the observation properties in SENSOR for each corresponding sensor and each quality control test.
- The input data is in NRT format, where each column of observations is under a unique sensor-URN.
- At ingest, the quality control algorithm assesses the input data for correctness and validity of observations.
- After requesting SENSOR for observation properties, the INGEST performs the defined tests.

# DASHBOARD and Monitoring



[DASHBOARD.awi.de](https://DASHBOARD.awi.de) enables monitoring of near real-time data using highly customizable dashboards and offers

- [REST services](#)
- [Aggregation and download services](#)
- [Software repository for access, download and visualization using Jupyter Notebook.](#)

- **2D graph widgets** depicting changes in a parameter with time; X-axis may also display another parameter
- **Heatmap widgets** depicting changes in a parameter with time and depth/altitude including contour plot option; ideal for displaying mooring and CTD data
- **Map widgets** including **color-coded parameters** for displaying data from ship-mounted instruments
- Statistical analysis of selected parameter
- **Direct data download**
- Machine-friendly **data access via REST interface**
- User-friendly configuration of **aggregation and display settings**
- **Re-use widgets** across distinct dashboards
- **Publishing** of individual dashboards (opened to all) or **sharing** within groups

# DASHBOARD and Monitoring



Dashboards ▾



Coastal underwater observatories

Login



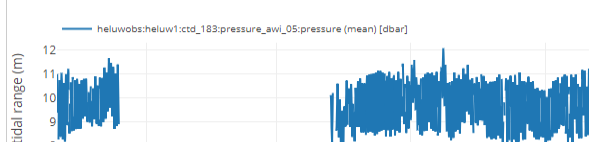
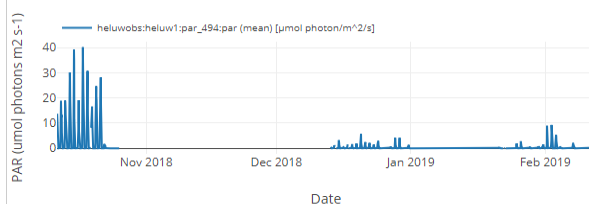
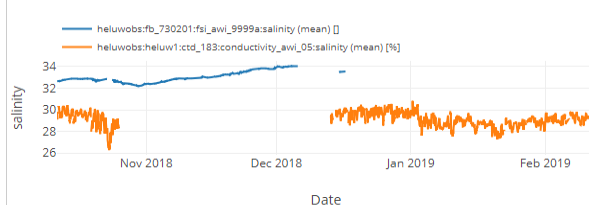
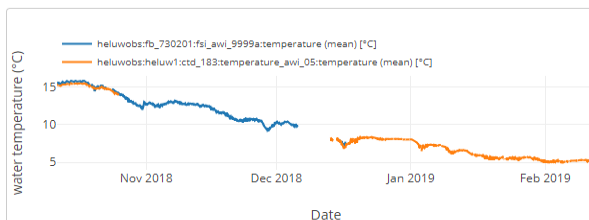
## The AWI O2A data frame workspace - an online analysis tool for remote sensor operation and big data access

Focus on specific dates or values by horizontal or vertical left-click dragging a zoom window. Double-click for focus reset.

Click on sensor names in the plots to select/deselect data.



### High resolution near-realtime data from the North Sea / Helgoland



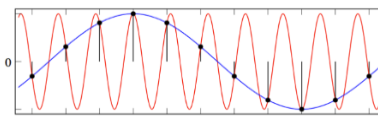
### Antarctica - North Sea - Arctic

- a comparative approach -

Climate change affect ecosystems differently depending on their geographic location. The North Sea and the polar coast in the Arctic and Antarctica are identified as hot spots in climate change and serve as "canary in the coalmine" to predict the future effects of climate change world wide.

### Why high resolution sensor data?

The Nyquist theorem describes the challenge (or problem) of sampling in proper time scales to avoid alias correlations.

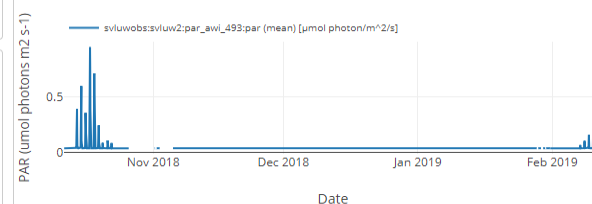
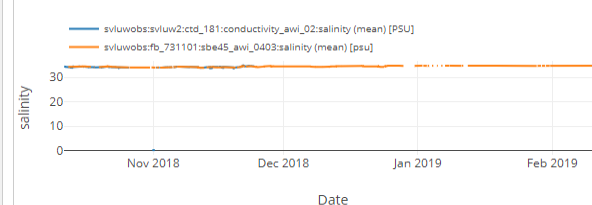
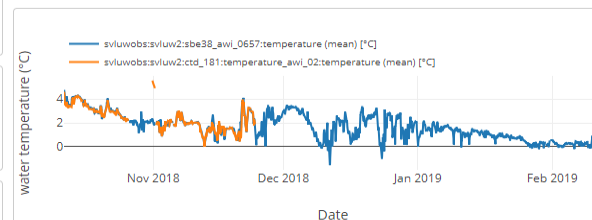


### Why remote controlled?

Polar and North Sea coastal areas often have limited access due to bad weather conditions and/or harsh environmental conditions e.g. during the polar night. Over the last years, fully automatated and remote controlled sensor technology and even complex underwater setups have been developped world wide and nowadays form a backbone of this kind of "tele(-connected)-ecology" especially in extreme environments.



### High resolution near-realtime data from Svalbard / NyAlesund



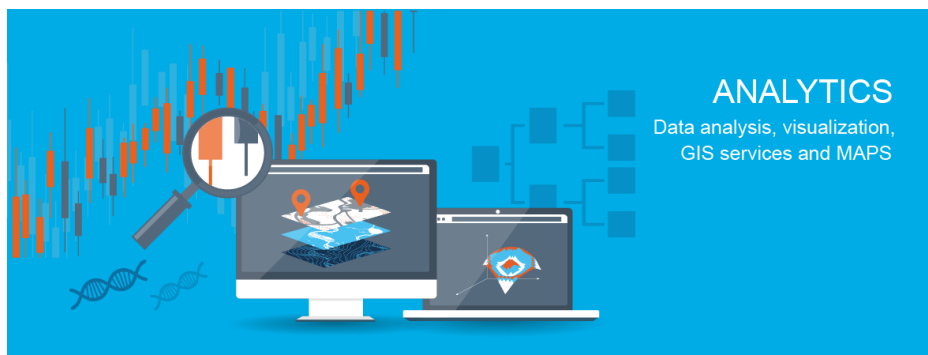


**WORKSPACE** provides solutions for

- data storage and computing for high end processing via cloud-like marketplace and virtualization infrastructures integrating **ANALYTICS**
- archiving and big data on-line storage and multiple tape storage and libraries at various locations.

**ANALYTICS** supports scientists in data analysis and visualization activities with infrastructure and tools

- **JupyterHub** is a web-based application for interactive work on Jupyter Notebook documents to combine source code, output, text and images while accessing multiple data sources within the O2A data framework and desktops combined with e.g. R or Python interpreter.
- **rasdaman** provides powerful array analytics for multidimensional large-scale datasets to extract data subsets e.g. via geographical features and perform queries in a federated environment
- [gitlab.awi.de](https://gitlab.awi.de) provides an open-source software development platform supporting controlled access to software projects
- [O2A github](#) **code repository** is a software development platform for open access examples on how to download, perform basic analysis and visualize data stored in various back ends using Jupyter Notebook



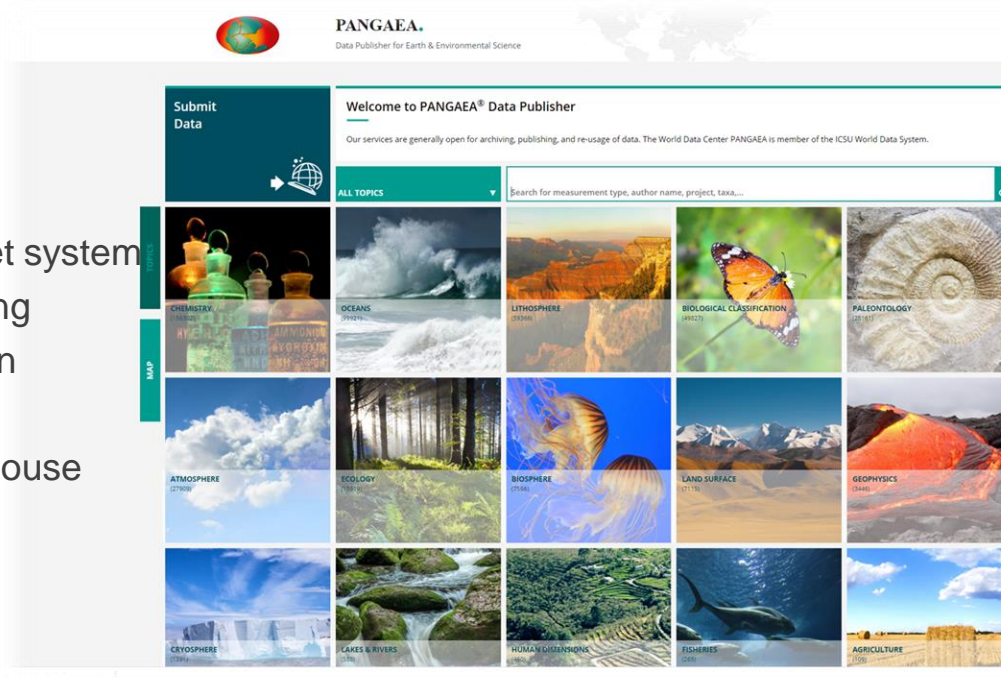
# Archiving and Publishing Data



**REPOSITORIES** offer archival and long-term preservation solutions for data, data products, publications, presentations and field reports.

**PANGAEA.de**  
repository for data and data products

- **Browse** by discipline, project and region
- **Search** in all metadata fields
- Metadata access [OAI-PMH Base](#)
- [Data download services](#)
- User-friendly **data submission** using a ticket system
- **Quality control** of datasets prior to publishing
- **DOI minting** including recommended citation
- [Tools](#) for data access and visualization
- **Machine-to-machine access** to data warehouse
- **Long-term** data archival
- Optional assignment of **data collections**
- **FAIR** principles



# Archiving and Publishing Data



**EPIC.awi.de**  
repository for publications and presentations

- **Browse** by item type, author, platform, organizational unit ...
- **Search** in all metadata fields including **full-text**
- Archival of a **large range of genre** from field reports to peer-reviewed articles
- **Handle minting** for all items
- **Preservation** of documents
- **Open access** to full items, optional embargo for PDFs
- Interoperability with **openAIRE** and **ORCID**

**LATEST PUBLICATIONS**  
Articles, Books, InBooks and Field reports

Article	Book	InBook	Report
Book, M. and Sabatini, R. (2020) <i>Environmental consequences of environmental origin: understanding and managing the risks</i> . Marine Biology, 167(4), pp. 1-14. doi:10.1007/s00227-020-0364-6. doi:10.1007/s00227-020-0364-6. doi:10.1007/s00227-020-0364-6	Book, M., Rymasz, M., Piel, V., Williams, T., Ranga, T. and Espersen, S. S. (2019) <i>Book: 362 International Report: Using time and submarine ice and Ocean Dynamics</i> . International Ocean Discovery Program. doi:10.1007/978-1-4939-9832-2. doi:10.1007/978-1-4939-9832-2. doi:10.1007/978-1-4939-9832-2	Book, M., Chelton, D., Venzke, A., Lippert, C., Sengler, T., Teyssie, A., Vigna, T. and Zhang, X. (2020) <i>Optimal processes in the Arctic atmosphere: a review</i> . A. K. and T. M. (eds.). In: Physics and Chemistry of the Arctic Atmosphere. Springer. ISBN: 978-3-030-03640-0. doi:10.1007/978-3-030-03640-0. doi:10.1007/978-3-030-03640-0	Book, M., Rymasz, M., Piel, V., Williams, T., Ranga, T. and Espersen, S. S. (2019) <i>Book: 362 International Report: Using time and submarine ice and Ocean Dynamics</i> . International Ocean Discovery Program. doi:10.1007/978-1-4939-9832-2. doi:10.1007/978-1-4939-9832-2. doi:10.1007/978-1-4939-9832-2

17277 PUBLICATIONS      17567 PRESENTATIONS      2824 THESES      1354 REPORTS

**ORCID**  
Get a unique identifier

**BEST PRACTICES**  
Learn about your options

**DATA PORTAL**  
Discover publications in expedition and data context

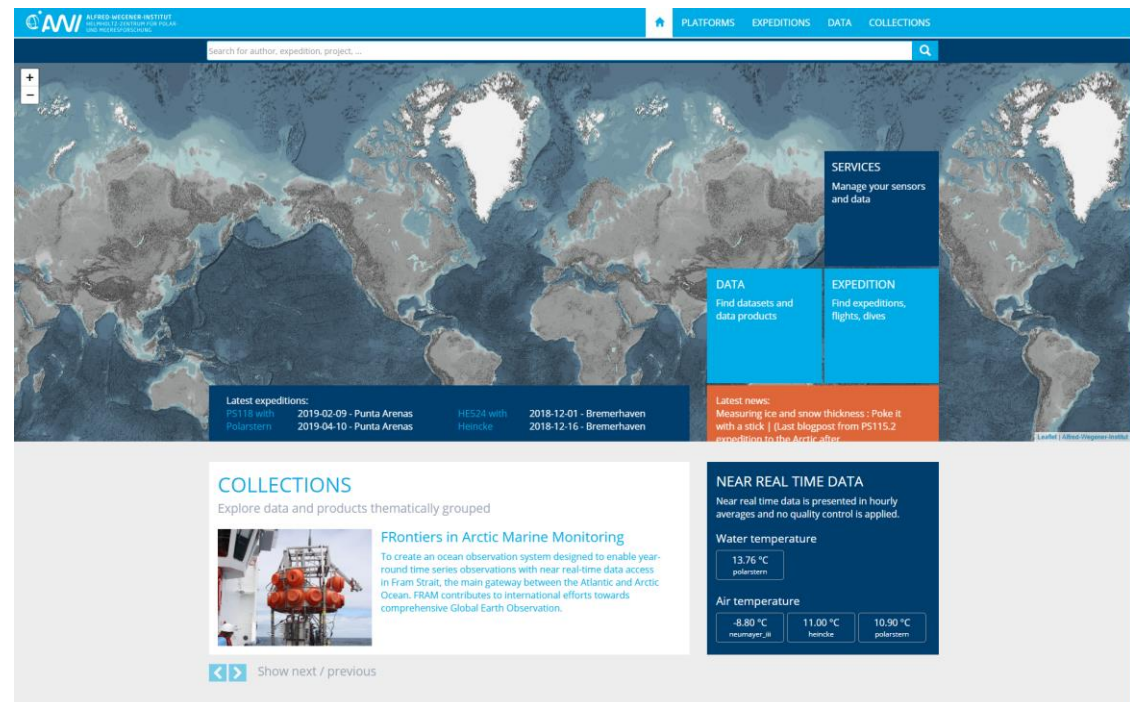
**PANGAEA**  
Publish your data and data products

# Archiving and Publishing Data



Portal [DATA.awi.de](https://data.awi.de) is a user-friendly one-stop-shop framework for integrated discovery of various types of scientific content, from near real-time data to quality-controlled WebGIS map products [maps.awi.de](https://maps.awi.de).

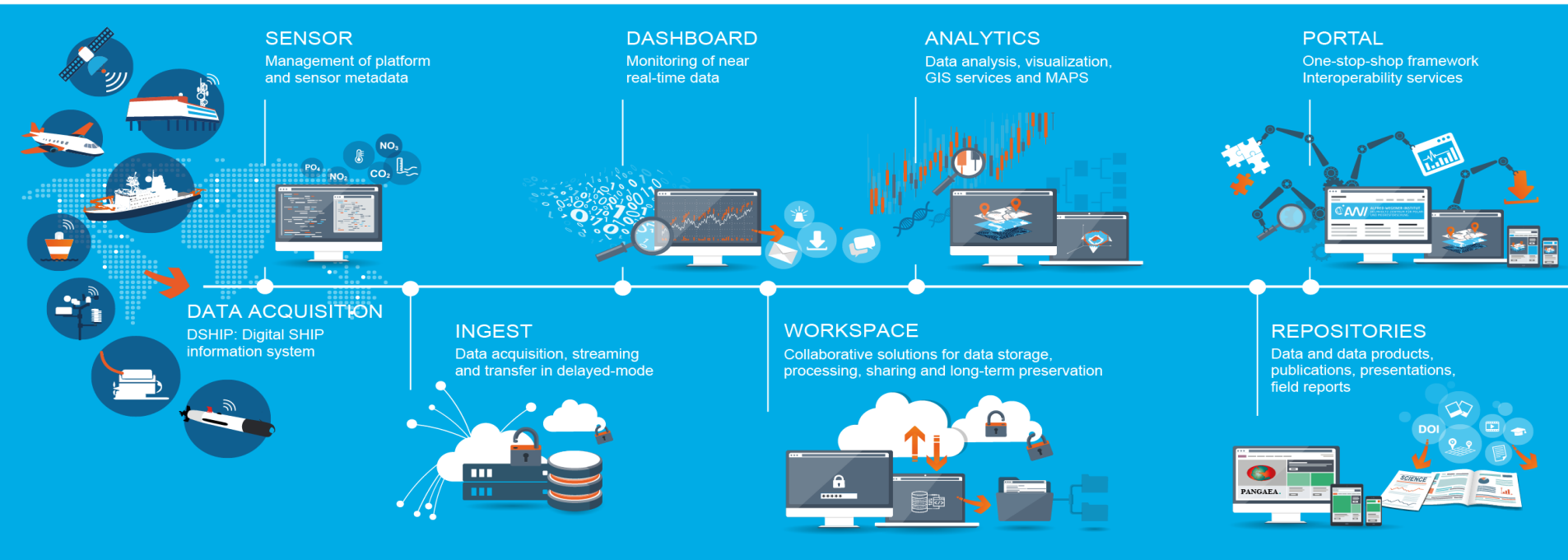
- Discovery of information on research **platforms, devices and sensors**
- Discovery of **campaign specific metadata** including its **scientific output**
- Discovery of **thematically grouped data, data products in collections** and WebGIS maps
- **Facetted search** allowing hit results to be filtered for repository type, project, platform, etc.



# Implementing FAIR in a Collaborative Data Management Framework



## O2A - Data Flow Framework from Observation to Archives and Analytics



AWI's generic and sustainable framework enables data flow and management of sensor observations to archives. This framework builds upon international standards for metadata and data interoperability and is meant to assist scientists in developing enhanced data products while implementing the **FAIR** principles