

Data flow, standardization, and quality control

presented by Brenner Silva with
the Software Engineering Team**,
and the Computing and Data Centre***
of the Alfred-Wegener-Institute
Bremerhaven, Germany*

*Contacts: * presenter: brenner.silva@awi.de, ** lead: roland.koppe@awi.de, *** head: stephan.frickenhaus@awi.de*

Context

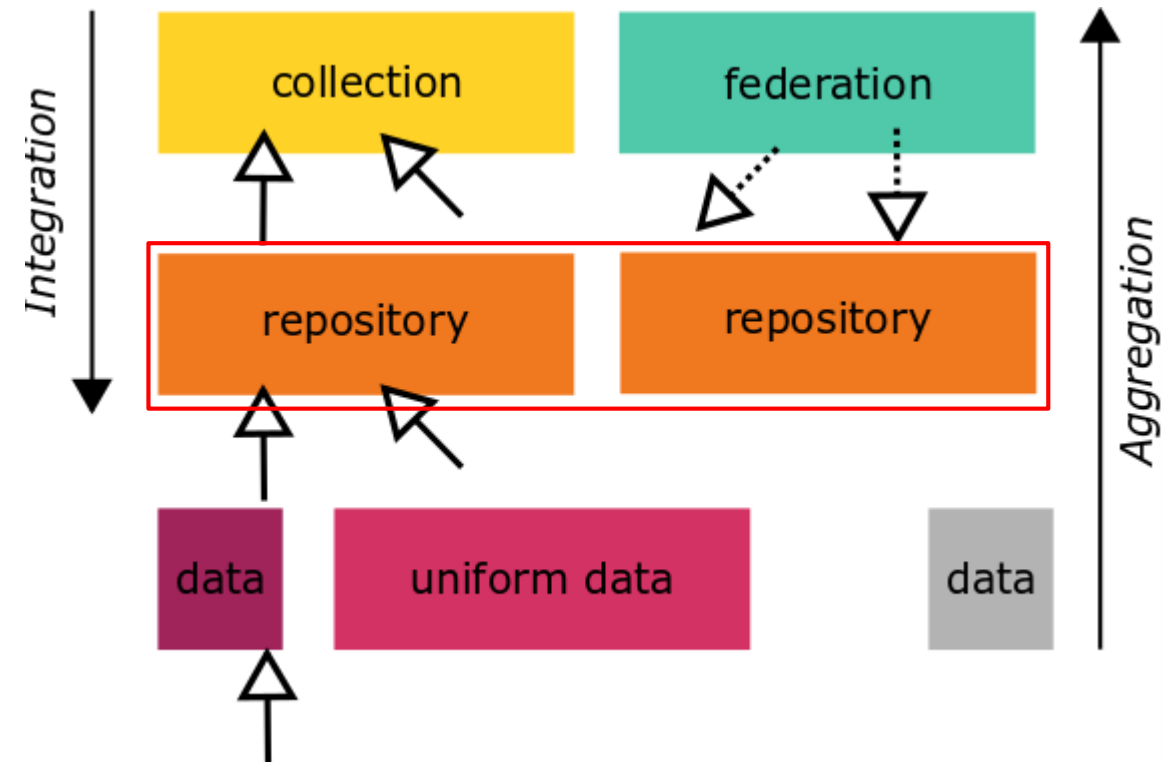
Defining data and data services

Registered single or uniform measurements

Repository: high level of integration for researcher, easy of use, standardization of incoming data.

Collections: catalogue of multiple organizations, lower level of integration, standardized and indexed metadata.

Federations: “connects once, access all”; control and maintenance of data remains with data providers.



Services

List of products, solutions or standards in use for integration and interoperability with **data repositories**.

Type	Name	URL	
OGC:WFS	LfULG, Saxony	umwelt.sachsen.de	
OAS:REST	Waterways, WSV	www.pegelonline.wsv.de	
OGC:WCS	rasdaman, AWI	data.awi.de/rasdaman	
OPenDAP	COSYNA, HZG	opendap.hzg.de	
THREDDS	Copernicus, EU	my.cmems-du.eu	
ftp	DWD, Germany	opendata.dwd.de	

Standards or solutions used in data service applications

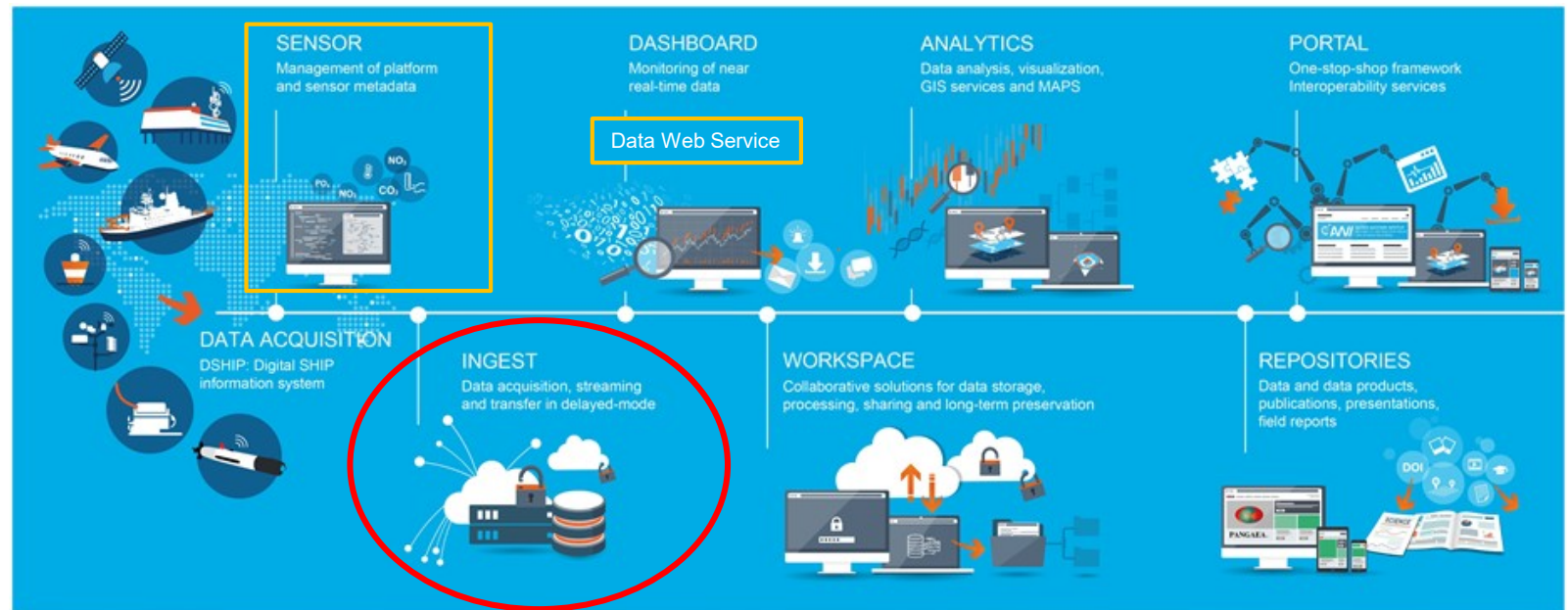
Solution / Protocol	Features / Functionality	
FTP	universal, slow transfer	
HTTP	generic, typing, compression, for small packages	
WSDL / SOAP (HTTP/SMTP)	extensible, web services specification via XML	
REST / HTTP	simple, customizable requests via HTTP, JSON	
OPenDAP / HTTP	selective data retrieval	
THREDDS / OPenDAP	versatile, multiple protocols and a single output (netcdf)	
OGC WCS / HTTP, SOAP	complete, multiple format encoding and transfer protocols	
.. WebDAV, SRM, XROOTD, RFIO, S3, Swift, CDMI ..	authoring, multicast, scalability, ...	

O2A - Observation to Archive and Analysis

- The data-flow framework operational and developed at the AWI, Bremerhaven.
- For interoperability the O2A currently uses OGC standards and REST architecture to support:

- SENSOR management
- near-real-time ingestion
- quality control
- data monitoring and request via:
 - Dashboard
 - Data Web Service
- analytics
- knowledge base registration

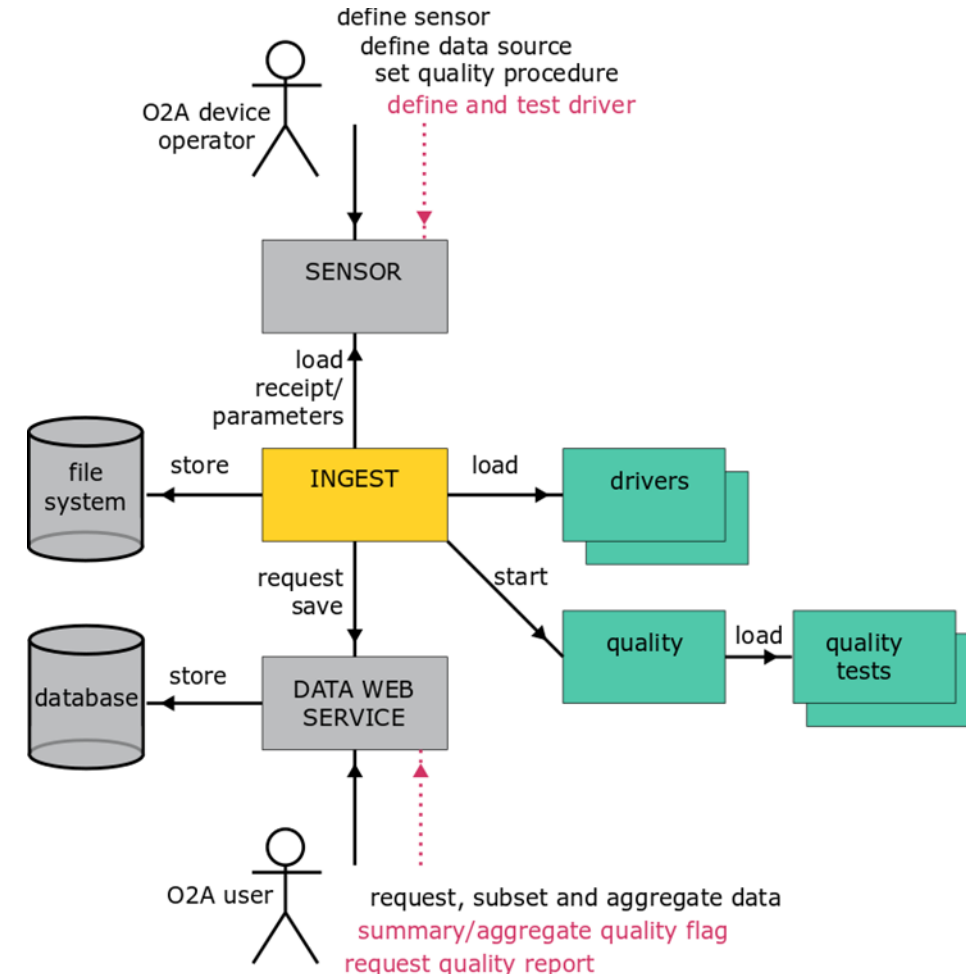
<http://data.awi.de/o2a-doc>



Ingest

The ingest performs automatic quality control to deliver quality-flagged data to the dashboard. Drivers are used to access data of specific formats (e.g. regarding data loggers of different instruments like weather stations, buoys, ferry boxes, CTD).

The quality control (QC) requests observation properties from the sensor REST-API 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 builds a table of devices and parameters to assess the input data for correctness and validity of observations.



Modules

illustrative script for data flow

```

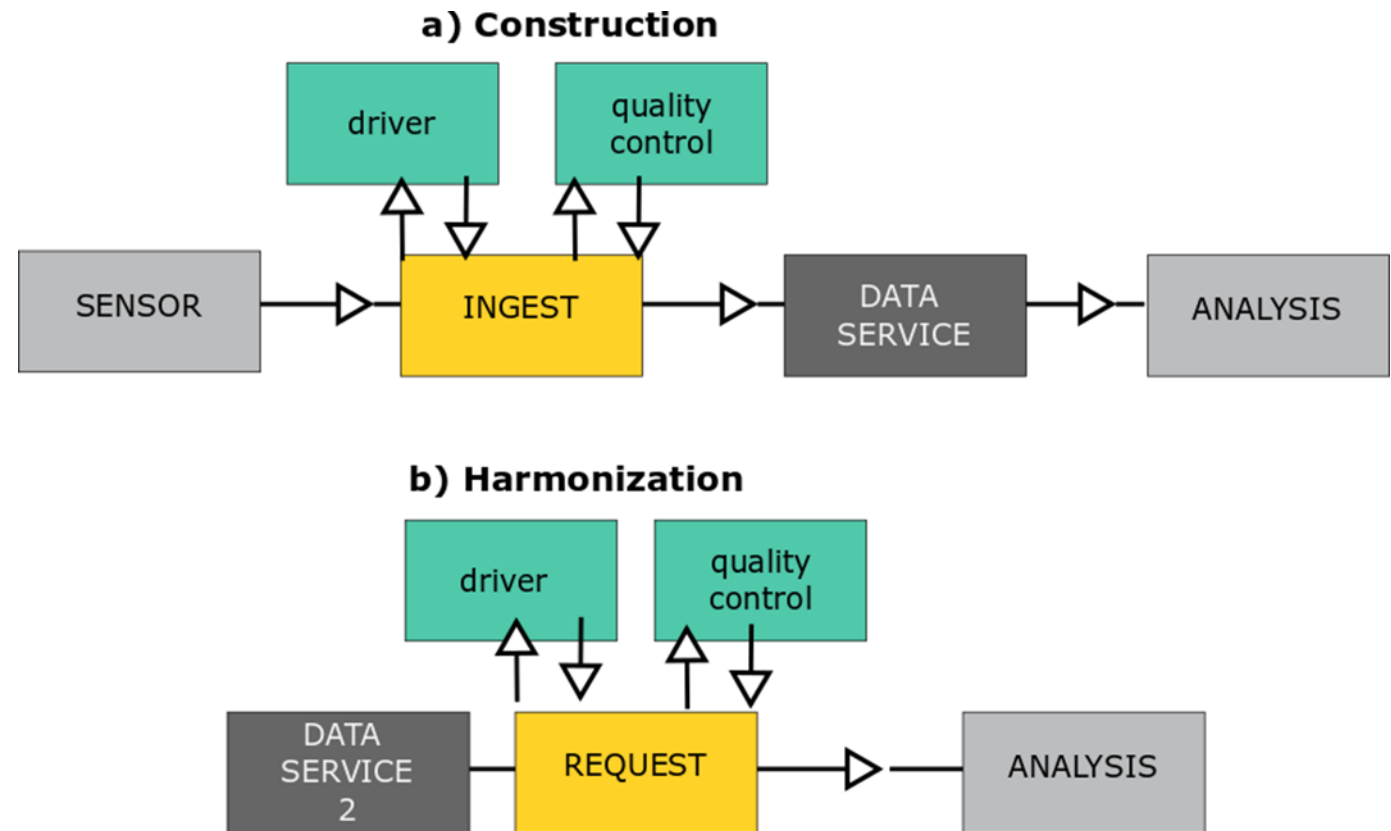
import dws      # for data request
import driver   # for transformation
import quality  # for quality tests
  
```

```
data = dws.request(source)
```

```
data = driver.format(data)
```

```
data = quality(data)
```

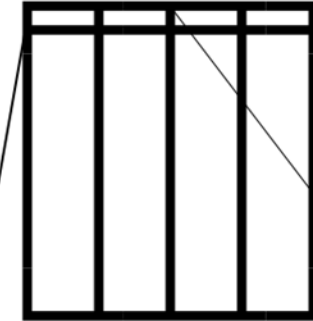
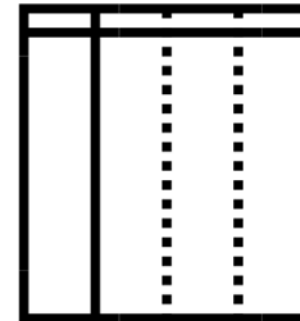
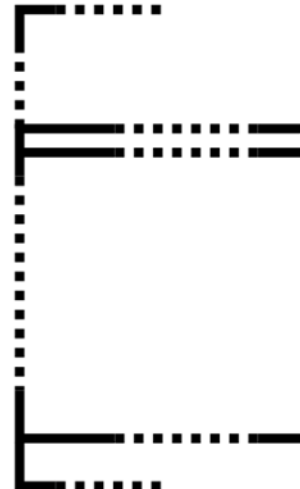
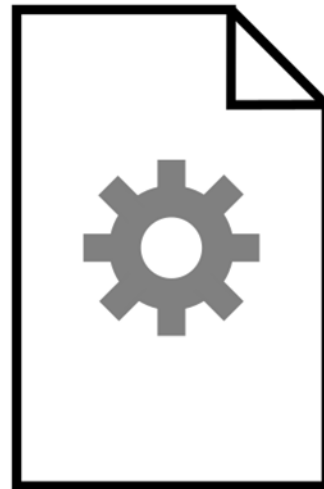
```
analyse(data)
```



Driver – for textual tabular data

Some features:

- **detect** encoding
- **apply** a data model to detect tabular data (separator, valid lines and valid rows)
- **validate** timestamp
- **map** header columns for sensor identification
- **store** in NRTformat



datetime	sensor : urn
DD-MM-YYYY HH:MM	station:name:sensor

Quality control

Quality assurance and optimization of sensor metadata are based on ingest procedure and quality tests.

The starting point for the currently implemented quality control tests is the ARGO real-time quality control (Wong et al. 2019).

Please refer to the left table or go to Quality Flagging for current status of the O2A. Currently, the flagging scheme in use is ordinal and of primary level (UNESCO 2013).

Test name	Description	Property required	Ancillary data required	Status
Operation temperature range	Test for temperature conditions (air and/or surface temperature) under which the instrumentation is deployed	Operation Temperature	Temperature observation	operational
Manufacturer range	Test if value is within the limits of the instrumentation (e.g. due to construction, material or filter) as given by manufacturer	Manufacturer range	None	operational
Operation range	Test if value is within a specific range valid for the location where the sensor is deployed	Operation range	None	operational
Gradient test	Test for gradient, i.e. absolute distance from the median value of neighboring (n=5) observations	Gradient Threshold	None	operational
Spike test	Test for spikes, i.e. distance from the median value subtracted by the standard deviation of neighboring (n=5) observations	Spike Threshold	None	operational
Range function	Test for physical relationships, or interdependency, among observations	Thresholds array	None	development
Geo location	Test for valid geographic location of moving and stationary sensors	Latitude, longitude and altitude ranges	Location of observations or Sensor Event	development

Quality flagging

A secondary level of the QF has been developed to represent the processing history into a **Quality Code**. In addition, a **Quality Score** can be used either to assess the **Quality Flag** or, as in the approach of the FZJ (Kaffashzadeh *et al.*, 2019), to indicate the plausibility of each observation.

a) Quality flags at primary level

Flag	Meaning
0	No QC performed
1	Good data
2	Probably good data
3	Probably bad data that are potentially correctable
4	Bad data
5	Missing value
6	Primal error
7	Not used

b) Illustration of observation including quality information

datetime	sensor : urn	Quality Flag	Quality Score	Quality Code	Uncertainty
DD-MM-YYYY HH:MM	station:name:sensor	[0-9]	[0..1]	[0-n]	[-1..1]

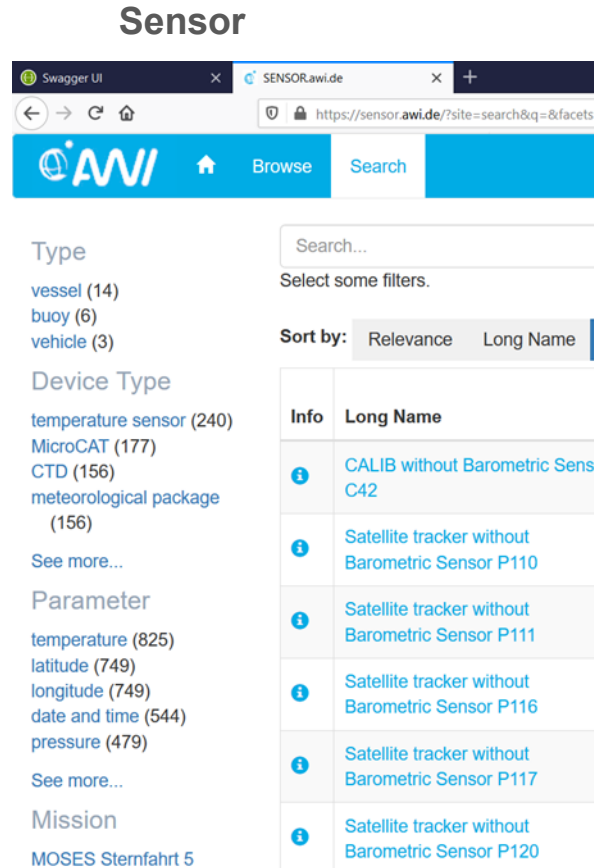
Sensor, DWS, and Dashboard

Sensor access and manage metadata.

The Data Web Service (DWS) gives a near-real-time overview of the harmonized data the current quality flag (QF).

Dashboard is for monitoring. The example on the right shows data from the Müglitz River collected by the UFZ (Nixdorf E. and Ködel U. 2019).

Sensor



Search...

Select some filters.

Sort by: Relevance Long Name Short Name URN Has no Parent:

Info	Long Name
	CALIB without Barometric Sensor C42
	Satellite tracker without Barometric Sensor P110
	Satellite tracker without Barometric Sensor P111
	Satellite tracker without Barometric Sensor P116
	Satellite tracker without Barometric Sensor P117
	Satellite tracker without Barometric Sensor P120

Type

- vessel (14)
- buoy (6)
- vehicle (3)

Device Type

- temperature sensor (240)
- MicroCAT (177)
- CTD (156)
- meteorological package (156)

See more...

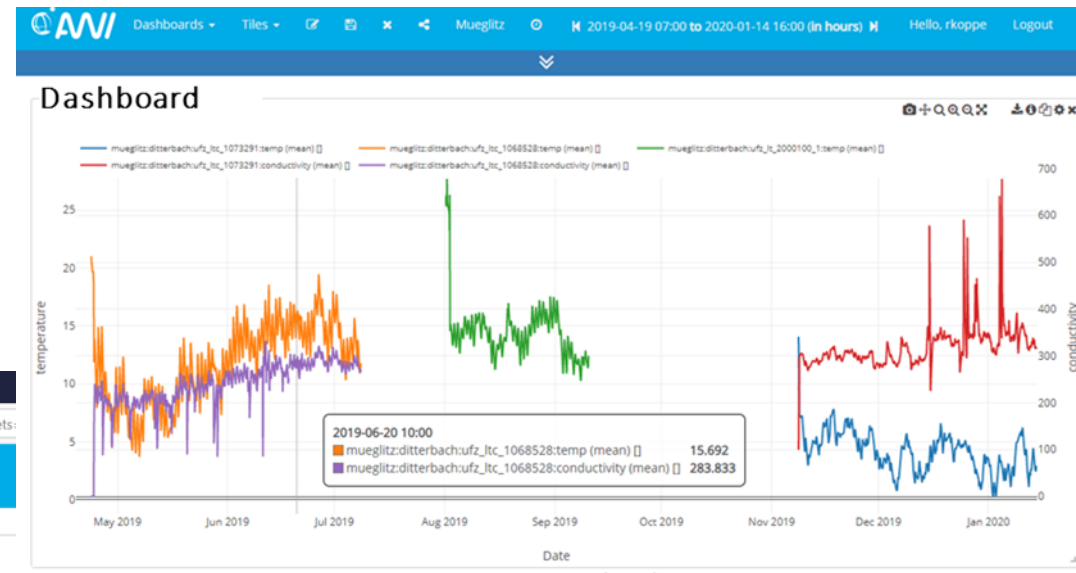
Parameter

- temperature (825)
- latitude (749)
- longitude (749)
- date and time (544)
- pressure (479)

See more...

Mission

- MOSES Sternfahrt 5



Data Web Service

Welcome to the O2A (near real-time) data services.

Our services provide open access to near real-time data streams described in sensor.awi.de. Monitoring of data streams is supported with dashboard.awi.de. Find code examples for R and Python on GitHub <https://github.com/o2a-data> and have a look into our O2A data flow documentation, especially for [quality flagging](#). An API for this service is provided [here](#).

Code	Age	Quality Flag
0 data streams selected. Click to show / hide selected streams.		
3222 data streams found.		
<input type="checkbox"/> buoy:2013s2:bar_pre	6 hours	982.10 hPa 1
<input type="checkbox"/> buoy:2013s2:bat_vol	6 hours	14.00 V 0
<input type="checkbox"/> buoy:2013s2:dis_to_sur_1	6 hours	0.18 m 1
<input type="checkbox"/> buoy:2013s2:dis_to_sur_2	6 hours	0.14 m 1
<input type="checkbox"/> buoy:2013s2:dis_to_sur_3	6 hours	0.21 m 1
<input type="checkbox"/> buoy:2013s2:dis_to_sur_4	6 hours	0.18 m 1
<input type="checkbox"/> buoy:2013s2:lat	6 hours	-70.68 degree 1

Type

- station (114)
- laboratory (102)
- vessel (42)
- test-station (31)
- test (17)
- buoy (15)
- ctd (2)

show more...

Parameter

- temperature (378)
- quality flag (207)
- concentration (195)
- depth (184)
- conductivity (134)
- pressure (132)

Sensor and DWS

The Sensor and the Data Web Service uses the REST architecture to offer open access to metadata and data. Useful for building client applications (e.g. driver and quality modules).



SensorWeb REST services

This is the API of the REST service endpoints for managing sensor metadata.

backend_Collections_Operations

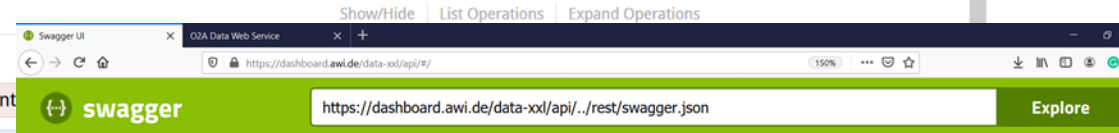
backend_Contact_Operations

DELETE /sensorsAdmin/contacts/deleteSystemCont
GET /sensorsAdmin/contacts/getAllSystemCont
GET /sensorsAdmin/contacts/getAllContactRefer

Parameters
 Parameter Value
contactID (required)

Response Messages
 HTTP Status Code Reason
 500 Internal Server Error .
 Try it out!

PUT /sensorsAdmin/contacts/putNewSystemCo
PUT /sensorsAdmin/contacts/modifySystemCon



Data Web Service ^{1.6.0}

[Base URL: /data-xxl/rest]
<https://dashboard.awi.de/data-xxl/api/./rest/swagger.json>

The data web service allows accessing and storing near real-time and delayed mode data. Have a look at the [overview](#) for more information about content and usage.

GET /data Loads data according to given query parameters.

POST /data Saves data specified in the body

Data must be tab-separated values with newlines. Containing a header of sensor codes. First column must be named datetime in ISO. The inserted data will be precalculated into statistics for different resolutions (e.g. HOUR). In this precalculation will only values count in with a qualityflag <= 3. The calculated qualityflag will be 0 if one of the values have 0 qf or the highest qf in resolution-step

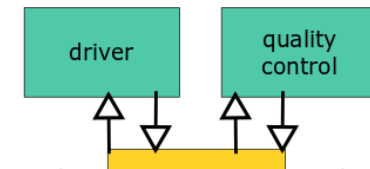
Parameters Try it out

Name	Description
body	Example Value Model
(body)	"string"
	Parameter content type
	text/tab-separated-values;charset=UTF-8

Take home

1. The landscape of standards and protocols is diverse. When considering integration, functionality levels the approach.
2. Driver and quality control aim to increase modularity in the O2A repository. Further tests are required for assessment, documentation, and the transferability.
3. Next use-case (AWIpev) aims to assess quality control implications on data aggregation of time-series.

Type	Name	URL
OGC:WFS	LfULG, Saxony	umwelt.sachsen.de
OAS:REST	Waterways, WSV	www.pegelonline.wsv.de
OGC:WCS	rasdaman, AWI	data.awi.de/rasdaman
OPenDAP	COSYNA, HZG	opendap.hzg.de
THREDDS	Copernicus, EU	my.cmems-du.eu
ftp	DWD, Germany	opendata.dwd.de



Acknowledgement

Literature review and development of driver and quality modules were carried out within the **Digital Earth** Project <<https://www.digitalearth-hgf.de>> with the **Software Engineering** Team (lead by Dr. Roland Koppe, <roland.koppe@awi.de>) and the **Computing and Data Centre** (head by Prof. Dr. Stephan Frickenhaus, <stephan.frickenhaus@awi.de>) of the **Alfred-Wegener-Institute** (AWI), Bremerhaven.

Thank you