O2A: A Generic Framework for Enabling the Flow of Sensor Observations to Archives and Publications

Roland Koppe, Peter Gerchow, Ana Macario, Antonie Haas, Christian Schäfer-Neth, Hans Pfeiffenberger
Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research
Bremerhaven, Germany

***Abstract*— The increasing number and complexity of research platforms and respective devices and sensors along with heterogeneous project-driven requirements towards satellite communication, sensor monitoring, quality assessment and control, processing, analysis and visualization has recently lead us to build a generic and cost-effective framework (O2A) to enable the flow of sensor observations to archives. O2A is comprised of several extensible and exchangeable components as well as various interoperability services and is meant to offer practical solutions towards supporting the typical scientific workflow ranging from data acquisition activities until the very last data publication activities. The web-based sensor monitoring component built within O2A offers a dashboard-oriented approach for displaying near real-time and delayed-mode sensor output parameters including simultaneous map and diagram viewing. This module allows project administrators and data specialists to monitor individual sensors in near real-time as well as to view the data values within a wished temporal range and/or geographical coverage. Additional examples of O2A components are the AWI-specific SensorML profile and raw data ingest solutions, the various data workspace areas and dispatcher middleware, the GIS infrastructure, and the ticket and data curation system as central hub supporting the final data publication activity. Finally, in the context of the large-scale multi-disciplinary project "Frontiers of Arctic Monitoring" Project (FRAM), we illustrate how the proposed O2A framework will assist scientists in developing enhanced data products and facilitate data re-use in the future.**

Keywords*— *marine data,sensor observations, data products, interoperability, information system, OGC standards, SensorML 2.0, SOS/SWE technology, data aggregation, data warehouse, monitoring dashboards, GIS, data publication, data dissemination

I. Introduction

Over the last two decades, the Alfred Wegener Institute (AWI) has been continuously committed to develop and sustain an infrastructure for coherent discovery, view, dissemination, and archival of scientific information in polar and marine regions [1], [2]. Most of the data collected by scientists originates from research activities being carried out in a wide range of research platform types operated by AWI: vessels, aircrafts, land-based stations, ice-based stations, moorings, floats, gliders, in-situ ocean floor stations, drones, and ocean floor crawling systems. Archival and publishing in the information system PANGAEA [3], [4] along with DOI assignment to individual datasets is a typical end-of-line step for most data owners.

A workflow for data acquisition from shipborne devices along with ingestion procedures for the raw data into institutional archives has been well-established at AWI for many years [5]. However, an increasing number and complexity of research platforms and respective devices and sensors along with heterogeneous project-driven requirements towards satellite communication, sensor monitoring, quality assessment and control, processing, analysis, and visualization has recently lead us to build a generic and cost-effective framework. This framework, hereafter named O2A, enables the seamless flow of sensor observation to archives and compliance with OGC standards [6], assuring interoperability in international context.

In this paper, we describe the distinct components of our framework as well as the added value of establishing relationship metadata among the various content types. Moreover, we show how our pragmatic sensor characterization effort can be re-used in the context of our sensor monitoring environment. Finally, we illustrate how the distinct O2A components presented in this paper can be used to support the scientific data workflow using the "Frontiers of Arctic Monitoring" project [7] as a use case.

II. Architectural Overview

Our generic data flow framework concept is comprised of several loosely coupled components aiming to provide end-users with a coherent environment for handling the main scientific activities defined in a human workflow (Figure 1). The central underlying component is the 'Data Workspace' (see section VII) which offers an integrated view on different data sources ranging from binary objects stored in file systems to data stored in relational databases or data warehouses. Based on these workspaces, added-value services and portals can be built which are wrapped in standardized sensor descriptions (section III).

The framework presented in this paper is focused on data that originates from heterogeneous devices and sensors mounted in various types of platforms. Each platform offers an automated and specific 'Data Acquisition' system which stores measurements on the platform itself, in proprietary real-time databases or high-volume data in file systems (see Figure 1). High-volume data comes in a variety of specialized data formats (e.g., SEG Y format for seismic). The data is commonly stored in one second intervals or even in higher frequency as typical for data from research aircraft.

978-1-4799-8736-8/15/$31.00 ©2015 IEEE

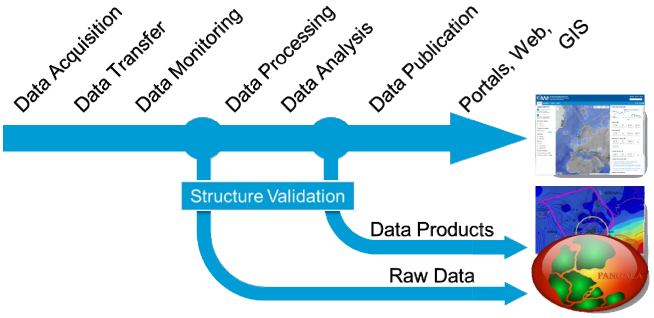


Figure 1: Schematic diagram illustrating in a simplified manner the main scientific activities in our data flow framework.

The second activity is named ‘Data Transfer’ after which the data is ready for proper use via our ‘Data Workspace’ (see section IV). Two distinct scenarios characterize the ‘Data Transfer’ activity:

- Near real-time transfer: measurements of selected parameters are sent directly to our land-based near real-time database in regular intervals (e.g., water temperature, wind speed and platform position). These measurements can be used to monitor devices and respective sensors via monitoring dashboards (see section V).
- Delayed-mode transfer: high volume data stored in proprietary databases and flat files is manually transferred via dump and restore scripts.

The ‘Data Processing’ and ‘Data Analysis’ activities are mainly subject to discipline-specific scientific work aiming to produce a ‘Data Publication’. Under this last publication activity, individual datasets, ready-to-use data products, and respective peer-reviewed publications are commonly generated as standard scientific output.

The scientific work is ideally based on the data stored in the ‘Data Workspace’, but cultural obstacles to such data policies still exist. In the scope of our framework, we plan to provide automated methods for data quality assessment and control as shown in Figure 4 as well as mechanisms for validation and automated generation of enhanced data products (e.g., color codes for along-track data visualization). Finally, we provide a set of solutions for discovery, view, dissemination, archival, and direct data download. This set comprises data portals as cited in [10] and [11] for integrated access to quality-controlled data.

III. Sensor Description and Interoperability Standards

Because the standard scientific workflow from data acquisition to data publishing typically takes over one year, it is crucial that, during this long chain of activities, all relevant information associated with the sensor characteristics is centrally archived and transparent to all for an improved re-usability. Also for example, by knowing relevant characteristics for sensors already deployed (e.g., depth of various sensors fixed on a mooring wire), the design and decision-making processes of future observational experiments can be improved.

In order to provide a mechanism for capturing sensor characteristics and related online resources, we have adopted the OGC standard SensorML 2.0 [12], a standardized mechanism for describing sensor resources related to the Sensor Web Services specifications [13]. With this standard we are able to describe not only the details related to the specific sensor (physical characteristics, positioning within the platform, accuracy, etc.) but also the related events (e.g., sensor calibration, device deployment and recovery) and sensor resources relevant for the scientific workflow (e.g., manufacturer factsheets, user manuals, documentation on sensor calibration, etc.). These resources are archived in our institutional publication repository where a citable digital object identifier (handle) is being minted for each individual item.

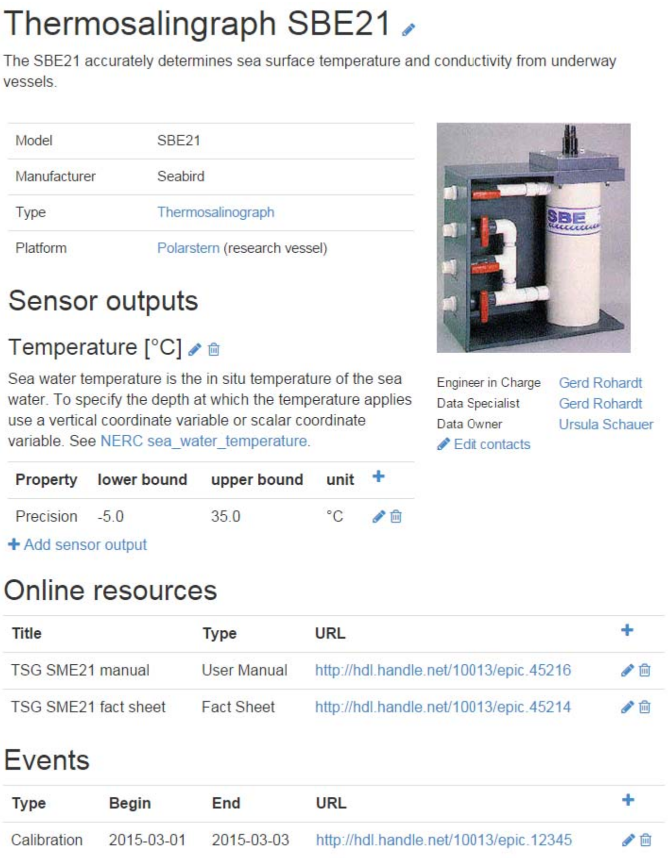
Aiming at the support of scientific activities in our institution and keeping in mind that that the solutions to be built must become effective in the hands of scientists who are usually not fond of complex software systems, we have developed a compact AWI-specific SensorML profile optimized for the practical needs of our scientists. In this profile, besides the standard sensor-related attributes, we are assuring that the needed provenance, lineage, and data governance information is also archived for re-use purposes. Figure 2 depicts our web-based frontend solution using the thermosalinograph SBE21 as example. Because the use of common vocabularies is an important prerequisite towards consistency and interoperability, we have adopted the terms describing platforms, devices, and parameters measured provided by the NERC Vocabulary Server V2.0 [14] which are widely used in the Pan-European infrastructure project SeaDataNet [15].

Another relevant aspect with regard to our AWI SensorML profile is the use of range values associated with individual sensor outputs (e.g., temperature values measured by SBE21 fall in the range -5 to +35 degrees Celsius). Because we insert upper and lower threshold values for describing the range of various sensor outputs, we will be able to re-use these in the dashboard-oriented monitoring solutions (see section V). Alerts and notifications are currently being set manually by dashboard administrators.

IV. Transfer of Data from Field to Workspaces at Land

As illustrated in Figure 1, the scientific workflow starts with data acquisition in the field along with a seamless transfer of data to workspaces. For the purpose of supporting these activities, the “Raw Data Ingest Framework” has been developed in close cooperation with engineers and scientists responsible for the various devices and sensors [5]. RDIF can be seen as an individual component in the data flow framework from sensor to archive described in this paper.

Within O2A, we offer support to both near real-time data transfer and delayed-mode transfer. Given that our research vessel Polarstern commonly operates in Arctic and Antarctic regions and our land-based station Neumayer III is located in the Antarctic, the near real-time data using satellite communication is very cost intensive. Therefore, we are currently transferring data from selected sensors only at 10-min intervals.



Thermosalingraph SBE21

The SBE21 accurately determines sea surface temperature and conductivity from underway vessels.

Model	SBE21
Manufacturer	Seabird
Type	Thermosalinograph
Platform	Polarstern (research vessel)

Engineer in Charge	Gerd Rohardt
Data Specialist	Gerd Rohardt
Data Owner	Ursula Schauer

Edit contacts

Sensor outputs

Temperature [°C]

Sea water temperature is the in situ temperature of the sea water. To specify the depth at which the temperature applies use a vertical coordinate variable or scalar coordinate variable. See NERC sea_water_temperature.

Property	lower bound	upper bound	unit
Precision	-5.0	35.0	°C

+ Add sensor output

Online resources

Title	Type	URL
TSG SME21 manual	User Manual	http://hdl.handle.net/10013/epic.45216
TSG SME21 fact sheet	Fact Sheet	http://hdl.handle.net/10013/epic.45214

Events

Type	Begin	End	URL
Calibration	2015-03-01	2015-03-03	http://hdl.handle.net/10013/epic.12345

Figure 2: Illustration of our web-based sensor description application using as example the thermosalinograph SBE21 mounted on our research vessel Polarstern. Here, temperature is defined as sensor output based on NERC vocabulary P07 [16] and online resources and calibration documentation are being archived in our publications repository using a citable digital object identifier.

Figure 3 shows the simple data model for storing near real-time data in a relational PostgreSQL database [6] with PostGIS [9] extension for geographic analysis and visualization purposes. Individual data points are identified as a chain of identifiers from platform over device, sensor and parameter as well as the timestamp of measurement in Coordinated Universal Time (UTC) [17]. Given that the data volume of near-real time data increases rapidly and fast data access is required, the data table has been divided in partitions per yearly quarter in order to be able to conduct operations like statistics or averages for graphical overviews.

The near-real time database is accessed via a lightweight JAVA web-application that provides REST-based methods for requesting data of selected sensors and parameters in different data formats like CSV or JSON.

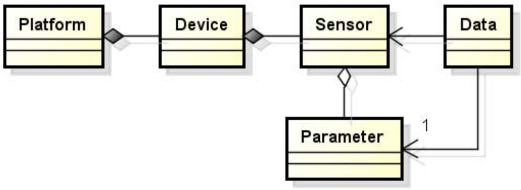


Figure 3: Simplified data model for storing near real-time data

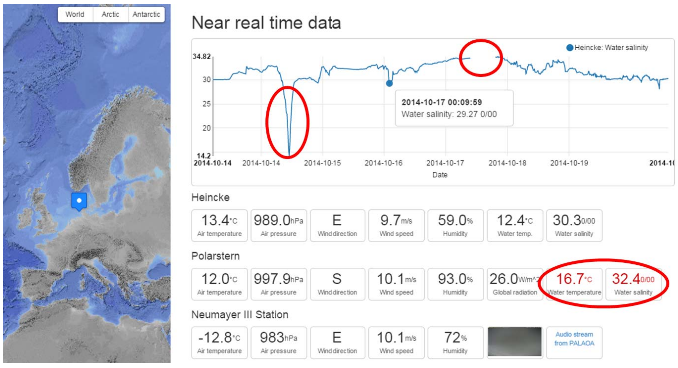
Delayed-mode data is transferred manually via hard disks or tapes being imported in our workspace on land (compare Figure 5). In order to synchronize near real-time data and delayed-mode data, we are using consistent UTC timestamps in combination with the geographic position for each transferred data point. Quality checks are being systematically performed on tracklines from our research vessels and aircrafts which then serve as master positions for geo-referenced measurements. These are stored in PANGAEA along with a citable digital object identifier (DOI). For an example see [18].

V. Monitoring Environment

Our web-based monitoring environment consists of a dashboard-oriented approach for displaying measurements archived in the near real-time database. This monitoring environment offers three main functionalities:

- display of current, raw measurement values; out-of-range values, as defined in SensorML, are highlighted (see section III),
- graphical visualization of time series for sensors and parameters selected by users aiming to assist in the identification of outliers and gaps,
- the locations of individual measurements are shown in a map which is especially relevant for moving platforms like research vessels.

The dashboards shown in Figure 4 were implemented in JavaScript with HTML5 using libraries and frameworks like jQuery [19], Leaflet [20] for maps and D3 [21] for visualizing the data itself. By using this lightweight approach, we are also able to build highly customizable and discipline-oriented dashboards as widgets which can be easily integrated in portal solutions and content management systems.



2014-10-17 00:09:59
Water salinity: 29.27 0/00

Heincke

13.4°C	989.0hPa	E	9.7m/s	59.0%	12.4°C	30.30/00
Air temperature	Air pressure	Wind direction	Wind speed	Humidity	Water temp.	Water salinity

Polarstern

12.0°C	997.9hPa	S	10.1m/s	93.0%	26.0W/m^2	16.7°C	32.40/00
Air temperature	Air pressure	Wind direction	Wind speed	Humidity	Global radiation	Water temperature	Water salinity

Neumayer III Station

-12.8°C	983hPa	E	10.1m/s	72%		Audio stream from PALAOA
Air temperature	Air pressure	Wind direction	Wind speed	Humidity		

Figure 4: Example of a dashboard with map, time series visualization, and highlighted problems with data quality as well as built-in notifications for selected parameters when exceeding the expected range of values.

VI. Visualization Environment

In order to support scientific analysis and publication of static as well as dynamic, geo-referenced data products, we are operating a GIS infrastructure, that focuses on the publication of ready-to-use data products as well as their accessibility. All GIS products are published in OGC compliant WMS [22] or WFS [23] formats.

Due to performance and security issues, we have established two independent GIS systems to support the standard scientific workflow illustrated in Figure 1: one for internal exchange and analysis purposes and the other one for open access publication of GIS data products for which GIS administration rights are required. The internal system is designed to allow scientists to build their GIS products independently just by using their desktop ArcGIS and dropping their GIS projects in a specific internal folder structure that is shared with ArcGIS server.

The core components of our GIS infrastructures consists of ArcGIS for Server [24] and PostgreSQL databases including the Spatial Database Engine (SDE). Data upload to PostgreSQL can be accomplished by using the ArcGIS desktop application. The GIS server is accompanied by an ArcGIS Web Adaptor, delivering proxy and load balancing functionalities. As a result, our GIS infrastructure can be easily extended by adding additional servers in order to manage increasing requests. Moreover, a generic web-based GIS viewer based on Leaflet [20] has been developed providing a user-friendly environment with time sliders, customizable data filters as well as further information associated with GIS projects.

Besides giving assistance to AWI scientists towards developing their own GIS data products, we are engaged in creating base maps by integrating existing data products with AWI data (e.g., GEBCO08 [25]). These base maps are not only being used in individual GIS projects but also in data portals like EXPEDITION [1].

Besides the web GIS solutions mentioned above, lightweight device-specific visualization solutions are being developed in order to be displayed in user-friendly web environments. These solutions are designed in particular to address the challenge of displaying selected sensor output for moving platforms.

VII. Data Flow to Archives – Data Workspace

As illustrated in Figure 5, our 'Data Workspace' concept is comprised of individual and specialized databases hosted in distinct database management systems (DMBS) and file systems. To date, we operate several 'DBMS' like MySQL [26], PostgreSQL [27] and SAP ASE [28] as well as a data warehouse on MySQL/Infobright [29]. The sketched 'File System' is managed as a large cluster of redundant disks and tape storages with a hierarchical storage management system.

A 'Dispatcher' middleware component handles import as well as export data requests from portals or other applications and information systems to the actual data storage systems. The underlying logical file systems can be divided into four areas:

- Area 'S' defines a scratch and cache area with fast and online access to flat files used for download, generic analyses, and visualization purposes. Data originating from acquisition systems or manually transferred is placed in Area 'S'.
- Area 'A' defines a personal area for scientists who are processing high-volume data. Data placed in Area 'A' is for temporary purposes and will not be shared with colleagues.
- Area 'B' defines a project area containing data which is shared in a group or project.
- Area 'C' defines the actual permanent archive. Raw data as well as processed data for long-term preservation are placed here. In contrast to areas 'S', 'A', and 'B', user access is restricted and read-only in Area 'C'.

According to our generic data flow framework approach, data is being transferred from 'DBMS' and 'S' areas over 'A' and 'B' areas to finally reach the access-restricted 'C' area.

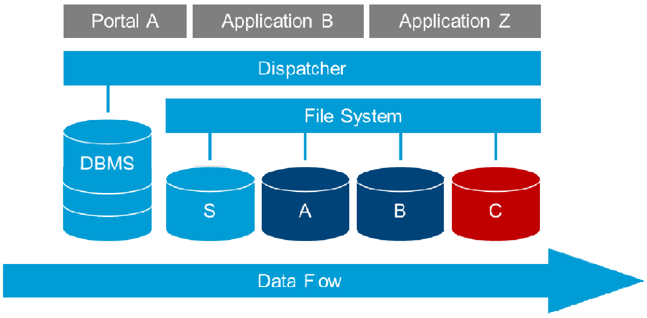


Figure 5: Schematic diagram illustrating of the technical data flow from databases (DBMS) and scratch or cache area (S) over logical areas (A) storing user working data and (B) storing project-related working data to (C) as "read-only" archive area (compare [30]).

Thanks to the data and workflow solutions presented in this paper, the transfer of data from scratch (area 'S') to the archive area 'C' area is performed in a semi-automatic way. The vast majority of all data produced in our institute is being archived at the information system PANGAEA [3] which is jointly operated by AWI and MARUM (Center for Marine Environmental Sciences). As member of the ICSU World Data System [31], PANGAEA follows its data management and archival policies and operates as an open access data library. The three major components associated with the information system PANGAEA are:

- A ticket and a data curation system jointly acting as the 'editorial system'. It is the central hub for data submission and curation, generation and description of metadata, assignment of citable DOIs to datasets, linkage of data and corresponding publications, data ingest, and archival. Any communication between submitting authors and data curators is handled and documented here as well.
- The storage system represents the central component and can handle geo-referenced data of any type. Depending on type, data and associated metadata is distributedly and redundantly being kept in a relational database (tabular and metadata), a data warehouse mirror for efficient retrieval, and on a disk / tape file system (flat files). All information is stored in two copies on tapes which are kept at two separate locations.

- The search and retrieval system allows users to conduct searches for any combination of metadata attributes, geographical or temporal coverage, parameter or species names, authors, research platforms, scientific projects, and so on. The search results can be either retrieved as a whole or further refined for additional parameters.

Open standards are being used in the two interface layers between the PANGAEA components listed above. On the ingest side, data sources like shipborne sensors may readily be fed into PANGAEA. From the retrieval perspective, other systems, publishers, and portals can easily harvest PANGAEA using OAI-PMH [32], with support for ISO19139 metadata format [33] and further web services.

VIII. Use Case and Future Outlook

The generic workflow framework presented here is currently in its pilot phase and is being tested in the context of the large-scale multi-disciplinary "Frontiers of Arctic Monitoring" (FRAM) project funded by the Helmholtz Alliance for the time frame 2015-2019 [7]. In this project, data collected from a wide range of platform types (e.g., research vessels and aircraft, sea ice tethered buoys, moorings, floats, gliders, autonomous underwater vehicles, ocean-based stations, sea floor crawlers, drones, etc.) is being automatically captured from an early stage on, adopting the data flow framework described above. So all scientific devices and respective sensors are being optimally described by FRAM engineers and data scientists using the sensor description solutions mentioned in section III. This includes not only sensor characteristics but also links to online resources like manufacturer descriptions, user manuals, and sensor calibration documents archived in our institutional publication repository (see example for these in Figure 2). In addition, all events associated with each sensor (e.g., maintenance, calibration, device recovery, etc.) are being individually recorded so as to increase the re-usability of the data in the future.

The dashboard-based monitoring environment described in section V allows FRAM engineers and data scientists to keep track of individual sensors in near real-time as well as to view the data values within a selected temporal range and/or geographical coverage. More specifically, the monitoring environment provides valuable assistance to scientists in terms of early detection of malfunction of sensors (e.g., alerts / notifications sent by email / SMS when measurements are out-of-range), filtering of data values for a certain range (e.g., temperature values above a certain range), and data aggregation (e.g., calculation of daily averages).

Another good example on how the distinct O2A components will assist us in developing enhanced data products for FRAM is the use of coloring schemes for parameters measured along tracks of mobile platforms (vessels, aircrafts, gliders, etc.). We are planning to develop a platform-specific averaging algorithm which will be optimized according to platform velocity as well as parameter range and accuracy as described in the AWI-specific SensorML profile (section III)

The FRAM project entails not only near real-time data but also in-situ observations delivered yearly in delayed mode and shipborne data delivered after the end of individual expeditions. To date, the raw data transfer process includes a semi-automatic file structure and a simple quality check. We are currently in the process of designing a flexible solution for extended validation checks and automated quality assessments including the assignment of quality flags as well as dynamic aggregation of data files within a geographic region or time ranges. Moreover, we will be building automated solutions for extracting metadata from data files and annotating these accordingly based on sensor descriptions. By adopting O2A, a seamless transfer of the data (including aggregated data) into the long-term archive PANGAEA including the minting of DOIs will be assured without additional efforts for FRAM data scientists or PANGAEA curators.

Given that each data type is associated with an individual quality control and post-processing technique (e.g., algorithms related to illumination and distortion compensations for seafloor images), we additionally plan an extensive algorithm harmonization effort within the FRAM project. Furthermore, we anticipate the use of Web Processing Services (WPS) [34] to bundle the various existing algorithms in a coherent fashion and thus improve their long-term usability in the future.

Aknowledgment

The authors would like to thank A. Boetius for supporting O2A with financial resources from FRAM project and I. Schewe and G. Rohardt for the fruitful discussions associated with end-user visualization needs. H. Liegmahl-Pieper and C. Wübber have also given us support over many years in regard to the hierarchical storage management system and other key storage solutions.

References

[1] A. Macario, R. Koppe, and H. Pfeiffenberger, "EXPEDITION - An integrated approach to expose expedition information and research results". Proceedings of the International Forum on "Polar Data Activities in Global Data Systems", 2013. hdl:10013/epic.42547

[2] A. Schäfer, and R. Koppe, "The Marine Network of Integrated Data Access and the Data Portal German Marine Research". In: Models of coastal waters in Germany: performance and application examples, Karlsruhe, ISBN: 978-3-939230-28-1, 2014. hdl:10013/epic.44782

[3] http://www.pangaea.de

[4] U. Schindler, M. Diepenbroek, H. Grobe, "PANGAEA® - Research Data enters Scholarly Communication", EGU General Assembly, Wien, 2013. hdl:10013/epic.38690

[5] P. Gerchow, "RDIF@AWI: Raw Data Ingestion Framework at Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research" , Third EUDAT Conference, Amsterdam, 28 September 2014 - 28 September 2014OGC standards, 2014. hdl:10013/epic.44298

[6] http://www.opengeospatial.org/standard

[7] I. Schewe, F. Janssen, O. Boebel, A. Bracher, T. Kanzow, K. Metfies, E. M. Nöthig, U. Schauer, T. Soltwedel, A. Boetius, "FRAM: A multidisciplinary observatory in the North Atlantic - Arctic Ocean transition zone" , Arctic Observing Summit 2014, Helsinki, Finland, 9 April 2014 - 11 April 2014, 2014. hdl:10013/epic.43662

[8] http://www.postgresql.org/

[9] http://postgis.net/

[10] J. J. Morton, V. Ferrini, S. H. O'hara, R. A. Arko, S. M. Carbotte, and B. Coakley, "Rolling Deck to Repository (R2R): Programmatic Quality Assessment and Processing of Marine Gravity and Magnetic Data and Associated Metadata". In AGU Fall Meeting Abstracts, vol. 1, p. 1511, 2011.

[11] R. Riethmu□ller, F. Colijn, H. Krasemann, F. Schroeder, and F. Ziemer, “COSYNA, an integrated coastal observation system for Northern and Arctic Seas”. OCEANS 2009-EUROPE, p. 1-7, 2009.

[12] http://www.opengeospatial.org/standards/sensorml

[13] http://www.opengeospatial.org/standards/sos

[14] http://www.bodc.ac.uk/products/web_services/vocab/

[15] http://www.seadatanet.org/Standards-Software/Common-Vocabularies

[16] http://vocab.nerc.ac.uk/collection/P07/current/

[17] http://www.timeanddate.com/time/aboututc.html

[18] http://doi.pangaea.de/10.1594/PANGAEA.841008

[19] http://jquery.com/

[20] http://leafletjs.com/

[21] http://d3js.org/

[22] http://www.opengeospatial.org/standards/wms

[23] http://www.opengeospatial.org/standards/wfs

[24] http://www.esri.com/software/arcgis/arcgisserver

[25] http://www.gebco.net/data_and_products/gridded_bathymetry_data/

[26] http://www.mysql.com/

[27] http://www.postgresql.org/

[28] http://www.sap.com/pc/tech/database/software/adaptive-server-enterprise/index.html

[29] https://www.infobright.com/index.php/products/mysql-integration/

[30] A. Treloar, D. Groenewegen, C. Harboe-Ree, “The Data Curation Continuum”, D-Lib Magazine, 2007. doi:10.1045/september2007-treloar

[31] https://www.icsu-wds.org/

[32] http://www.openarchives.org/pmh/

[33] https://marinemetadata.org/references/iso19139

[34] http://www.opengeospatial.org/standards/wps