Die Küste, 81 (2014), 19-28

The Marine Network of Integrated Data Access and the Data Portal German Marine Research

Angela Schäfer and Roland Koppe

Summary

The linkage of data beyond disciplinary boundaries is essential for marine research. Sufficient national and international data infrastructures are fundamental for central and easy access to the variety of existing, but distributed datasets in marine science. The Marine Network for Integrated Data Access (MaNIDA) provides a national networked approach in accessing and mining of federated marine research data infrastructures together with data management strategies and data workflows. In that course the MaNIDA consortium conceptualized and developed the "Data Portal German Marine Research" for coherent discovery, view, download and dissemination of scientific data and publications. The data portal is based on a central harvesting and interfacing approach by connecting distributed data sources. Here we inform about the specific details of the portal in terms of content, functionality, services, architecture, interfaces, standards and the contributing data providers.

Keywords

data infrastructures, data portal, marine research data, data access, data retrieval, portal architecture, standards, interfaces, web services, data providers, data management, data workflows, data publication, data citation

Zusammenfassung

Für die marine Forschung ist die Verknüpfung von interdisziplinären Daten essentiell. Um dahingehend einen zentralen und leichten Zugriff auf die existierenden und vielfältigen, jedoch verteilten Daten der marinen Forschung zu ermöglichen, sind gut funktionierende nationale und internationale Infrastrukturen fundamental wichtig.

In diesem Sinne entwickelt das "Marine Network for Integrated Data Access" (MaNIDA) einen nationalen und vernetzten Ansatz für die Auffindbarkeit und den Zugriff auf verteilte marine Forschungsdateninfrastrukturen mit entsprechenden Datenmanagementstrategien und Arbeitsabläufen. Um sowohl eine kohärente Datenauffindbarkeit, Visualisierung und einen einfachen Datenzugriff als auch die Veröffentlichung von wissenschaftlichen Daten und Publikationen zu ermöglichen, entwickelt das MaNIDA-Konsortium das zentrale „Datenportal Deutsche Meeresforschung". Durch ein zentrales, automatisiertes Harvesting-Verfahren und standardisierte Schnittstellen verknüpft dieses Datenportal unterschiedliche und verteilte Datenquellen. In diesem Artikel werden sowohl die speziellen Aspekte des Portals zum Thema Inhalte, Funktionalität, Dienste, Architektur, Schnittstellen und Standards als auch die beitragenden Datenanbieter vorgestellt.

Schlagwörter

Dateninfrastruktur, Datenportal, marine Forschungsdaten, Datenzugriff, Datenabfrage, Portalarchitektur, Standards, Schnittstellen, Webdienste, Datenanbieter, Datenprovider, Datenmanagement, Datenarbeitsabläufe, Datenpublikation, Datenzitierbarkeit

Contents

1 Introduction

In earth system research major achievements in scientific knowledge increasingly depend on the availability of data. Especially the linkage of data beyond disciplinary boundaries is essential for global change research. Major observations and new insights on global environmental change can only be obtained if data is collected over long periods and with easy access in a coherent manner. Compared to other fields of research (astronomy, high-energy physics, genetics) neither sufficient national nor international data infrastructure exists that enables central and easy access to the variety of existing, but distributed datasets in marine science.

The Marine Network for Integrated Data Access (MANIDA 2014), co-financed by the Helmholtz Association and several participating marine research institutes and universities in Germany, provides a networked approach in accessing and mining of federated marine research data infrastructures together with management strategies targeting long-term sustainability. The network aims to create a new paradigm in respect to integration, harmonization and aggregation of quality-controlled marine research data and related data products.

One of the main tasks of MaNIDA is the implementation and maintenance of a sustainable e-infrastructure for coherent discovery, view, download and dissemination of scientific data and publications in the form of a central data portal – the DATA PORTAL GERMAN MARINE RESEARCH (2014). For the first time a large amount of marine research datasets from nationally operated research platforms and monitoring facilities are made searchable and accessible through a single portal.

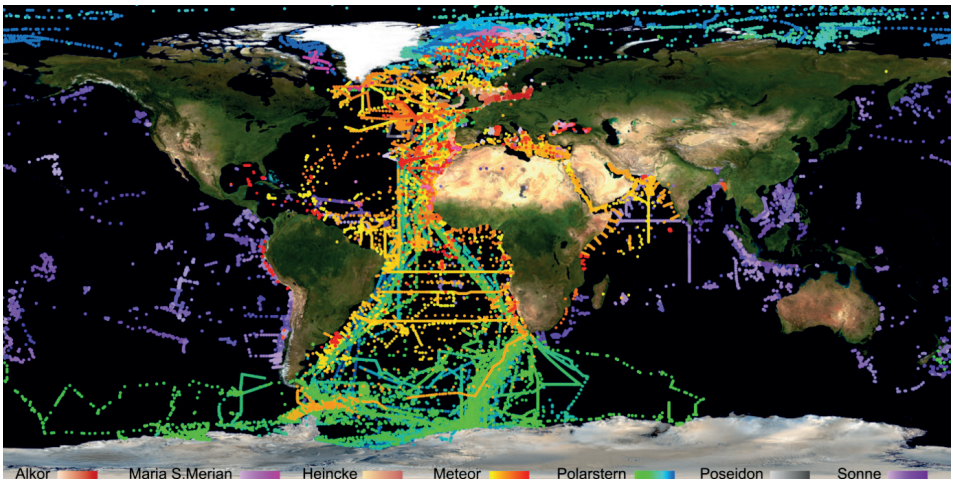

Figure 1: Recent status of sampling stations and cruises by major German research vessels (source PANGAEA 2013).

2 The MaNIDA Consortium

The Marine Network for Integrated Data Access started in February 2012 as an Impulse and Networking Fund project financed cooperatively by the Helmholtz Association and the partner institutions. Currently the consortium consists of five partner institutions and two associated universities that are engaged in marine research and in establishing research data infrastructures. The initial partners are:

- AWI – Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven
- MARUM – Zentrum für Marine Umweltwissenschaften, Universität Bremen
- BSH – Bundesamt für Seeschifffahrt und Hydrographie, Hamburg und Rostock-Warnemünde
- GEOMAR – Helmholtz-Zentrum für Ozeanforschung Kiel
- HZG – Helmholtz-Zentrum Geesthacht, Zentrum für Material- und Küstenforschung GmbH
- CAU – Christian-Albrechts-Universität zu Kiel, Institut für Informatik (associated)
- UHH – Universität Hamburg, Leitstelle Deutsche Forschungsschiffe (associated)

MaNIDA and the "Data Portal German Marine Research" are coordinated by the Alfred Wegener Institute. A steering committee and several cross-partner working groups (comprising more than twenty scientific staff members mainly by own resources) have been installed. They are working jointly on data workflows, harmonization, standards, technical concepts and the practical implementation of the central data portal, interfaces and the structural adaption of the participating distributed data providers.

After the initial phase ending in July 2014 the MaNIDA consortium is open to integrate additional national partner institutions that are willing to engage in establishing joint research data workflows, standards and a common e-infrastructure for marine research data.

2.1 Development of data workflows and data curation

Central search and access to marine research data will only work out in a satisfying manner if data contents are well provided and made accessible in a reliable and re-usable way. Therefore MaNIDA not only focuses on technical solutions alone but on developing data relevant workflows and curation procedures for scientific data as well. This approach comprises

- validated expedition and campaign information on German research vessels and platforms,
- archived, quality controlled primary data, near real time data and data products,
- persistent citability and re-usability of data and
- scientific publications and reports.

In order to achieve these intentions the cross-partner working groups of MaNIDA are also involved in the generation of outlines for best practices in terms of data validation, archival and dissemination of marine research data as well as proposals for decision-makers und committees for marine research infrastructures.

In the long run additional benefit for the research community will be improved procedures for data ingestion, data quality assurance and user support, thereby achieving substantial enhancements in the lifecycle management of marine scientific data. Explicitly, the creation of a Data Curation Center is planned for the next project phase along with the implementation of technical and organizational structures and ways to sustain the infrastructure financially. The main tasks for a joint Data Curation Center for marine research data are:

- support of workflows and user support,
- contact point and editorial system for standardized vocabularies and ontologies and
- structured data management.

In that term MaNIDA reaches out to stakeholders in the academic sector and government agencies so as to create acceptable and reliable workflows.

Figure 2: Simplified workflow schema for well curated research data.

The work of MaNIDA and the central data portal does not only enhance data mining in various disciplines and improve planning of future expeditions, via easy identification of missing data, but also provides means for combining various data types (underway data, post cruise data, satellite gridded products, modeling data, etc.) into enhanced data products.

Hence cross-partner working groups have been established by members according to their expertise as they are involved in the following topics:

- Data portal architecture, interfaces and implementation towards international standards and requirements for participating data providers
- Expedition catalogue – centrally maintained for up-to-date and past expeditions' metadata linked to research data and publications

- Underway data – validation and harmonization via an improved data acquisition and information system (DSHIP 2014) for recording of technical, nautical and scientific data taken at sea on board of all major German research vessels
- Cruise Summary Reports – semi-automated generation and interfacing via DSHIP for national data providers and international alignment of major German research vessels and platforms
- Vocabularies, ontologies, quality assurance – harmonization and establishment of international standards
- Joint data curation center and ticketing system – requirements, concept and establishment
- Adequate handling of marine research data – joint data policies and data management plans

3 The Data Portal German Marine Research

3.1 Content and functionalities

The "Data Portal German Marine Research" offers an integrative "one-stop-shop" framework for coherent discovery, view, download and dissemination of marine research data (research vessels, observatories, gliders, etc.). Its content originates from distinct data and publication repositories offered by the German marine research partner institutions. As Figure 3 illustrates, the portal integrates cruise-related metadata (expeditions), reports, publications, archived data and near real time data as well as data products of the whole marine and earth science community of Germany and their international projects. The portal focuses on published datasets along with necessary metadata and global persistent identifiers (e.g. DOI 2014) for proper citability to facilitate the general attitude of data sharing and data re-use as well as the acknowledgment of the original data provider.

Figure 3: The Data Portal German Marine Research and its integrated content.

Up to now the following search functionalities are implemented in the data portal:

- Generally data and publications are searchable by keywords described in the metadata of datasets collected from data provider repositories. Additional constrains like

temporal or geographical coverage either by regional gazetteers or graphical boundary boxes can be set.
- Also datasets can be found by facetted search and via prepared categories like data providers (repositories), regions (gazetteer), authors, platforms, expeditions/campaigns, projects, devices, parameters and others.
- Along with datasets, data products and near real time data the portal allows for searchable and interlinked access to scientific publications, reports and documentations from institutional publication repositories aligned with accessible datasets.

3.2 Value-added services

Case studies for improved and enhanced data retrieval are developed within MaNIDA and offered as value-added services in the data portal. These case studies are prototypical and will evolve over time with user feedback. So far three additional services are offered by the data portal:

- A module that enables direct access to the data sources of PANGAEA (2014, AWI/MARUM), COSYNA (2014, HZG) und DOD (2014, BSH). Distinct measurements can be requested by defining a temporal and geographical extent for harmonized parameters, e.g. salinity or temperature. The requested and integrated data is then available for download. This direct access service will be extended as more parameters are mapped and made available between the data providers via international vocabularies.
- Validated bathymetry data of the partner institutes are pooled together for the first time with a well-defined extended and harmonized metadata description for central retrievability.
- Near real time data in the German Bight are visualized and retrievable within the data portal by comparison of current vector fields (surface sea water velocity).

3.3 Data providers

The datasets accessible in the "Data Portal German Marine Research" are physically drawn from different sources and are maintained and updated at different institutes. Hence datasets are delivered directly in a provider specific manner – but presently accessible and up-to-date.

Harmonization of data content, parameters, vocabularies, etc. is a major ongoing issue while technically connecting the distributed data holdings of partner institutions that have developed solitarily for the last decades. Together with our scientific community and our data providers we work continuously on standardization, classification and the linking with international initiatives like ICSU World Data System (ICSU WDS 2014), Global Earth Observation System of Systems (GEOSS 2014), Global Biodiversity Information Facility (GBIF 2014), World Register of Marine Species (WORMS 2014) and EU initiatives such as the European Marine Observation and Data Network (EMODNET 2014), SEADATANET (2014) with BODC NERC (2014) and CF-standardizations (CFC 2014) as well as EUROFLEETS (2014).

The participating data providers for the "Data Portal German Marine Research" are:

- PANGAEA (2014) – Data Publisher for Earth & Environmental Science is an information system for long-term archiving and publication of data from earth & environmental sciences operated by AWI and MARUM as an Open Access library for geo-referenced data since 1993. PANGAEA is accredited by the "International Council for Science" (ICSU) as World Data Center and by the "World Meteorological Organization" (WMO). Each dataset can be identified, shared, published and cited by using persistent Digital Object Identifier (DOI). Data are archived as discrete and citable data collections or as supplements to publications. As a data library PANGAEA links primary data related to articles in earth and environmental science journals of the ELSEVIER (2014) portfolio and COPERNICUS PUBLICATIONS (2014) freely available in SCIENCE DIRECT (2014). Currently PANGAEA serves more than 350.000 datasets and 8 billion measurements.
- COSYNA (2014) – "Coastal Observing System for Northern and Arctic Seas" is an operational, integrated observational system that combines observations and numerical modelling. It measures key physical, sedimentary, geochemical and biological parameters at high temporal resolution in the water column and at the sediment and atmospheric boundaries. Data products range from real time or near real time datasets and web services generated from in-situ observation or remote sensing systems to forecasts resulting from operative modelling. COSYNA data management organizes the data streams between observational and central storage systems at the Helmholtz-Zentrum Geesthacht and partner sites, the data documentation and the user interfaces for data retrieval and presentation. Around 15.000 datasets of COSYNA are integrated in the portal.
- DOD (2014) – The German Oceanographic Data Centre was established in 1967 between the German Hydrographic Institute (BSH) and the DFG, the German Research Foundation. The national oceanographic database comprises currently data and information on German cruises and almost 300.000 stations with more than 45 million data values of about 900 parameters. These are mainly oceanographic variables such as temperature and salinity, chemical variables with nutrients, organic, inorganic and radioactive components in seawater, sediments and biota (fish and mussels).
- Aligned with the accessible datasets by the data providers the data portal provides access to scientific publications, reports and documentations. The participating institutional publication repositories are EPIC (2014) at AWI, OCEANREP (2014) at GEOMAR and publications from HZG (2014) and MARUM (2014).

3.4 Architecture

Fig. 4 depicts the underlying architecture of the "Data Portal German Marine Research" The integration of federated content is realized by harvesting and indexing metadata offered by our data providers. The harvesting and indexing approach allows searching of scientific content with high performance based on an APACHE LUCENE (2014) ELASTIC SEARCH (2014) cluster. Since the systems of data providers offer different metadata formats via distinct interfaces the following metadata fields are defined as minimal requirement:

- Persistent identifier
- Title
- Publication date
- Authors, principal investigators
- Begin and end date of data
- Platform or research vessel
- Expedition
- Measured parameters

Furthermore the metadata contains dissemination links to data download services or dynamic services, e.g. Sensor Observation Services (SOS) or Web Map Services (WMS). To present users harmonized (metadata) content a Feature Catalogue is used to align different vocabularies of parameter names to standard names and to annotate original metadata with marine region names based on a standardized Gazetteer.

Next to the metadata delivered by data providers, the portal holds a database on expedition metadata (Expedition Catalogue). The Expedition Catalogue combines metadata about expeditions (e.g., IDs, begin, end, ports, persons) from distinct sources and is therefore used as master catalogue for the portal. The contents of Feature Catalogue and Expedition Catalogue are edited by the mentioned data curation center.

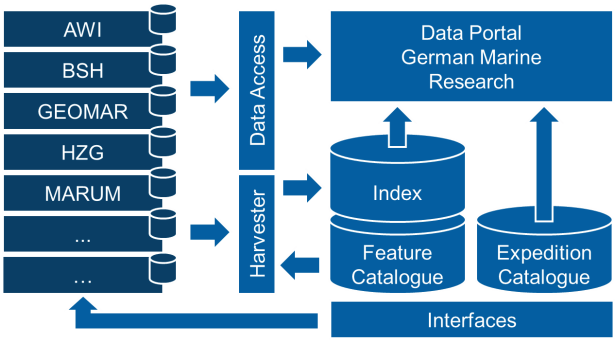

Figure 4: Architecture of the "Data Portal German Marine Research".

Additionally, the portal provides interfaces in form of web services for the Feature Catalogue and the Expedition Catalogue. Therefore data providers can use curated information and metadata for their own catalogues.

3.5 Interfaces and standards

The harvesting approach supports different standards for metadata collection. Three of the supported protocols are:

- OAI-PMH (2014) – Open Archives Initiative Protocol for Metadata Harvesting,
- OGC CSW (2014) – Open Geospatial Consortium Catalogue Service for Web and
- OGC WFS (2014) – Open Geospatial Consortium Web Feature Service.

Metadata is expected in form of XML serializations. If provided INSPIRE (2014) compliant ISO 19115 / 19139 format is recommended. Dissemination of data described by

metadata datasets are defined by the data provider. Typical data delivery formats are tab-delimited (TSV) or comma separated values (CSV), possibly packed as ZIP archives as well as portable document format (PDF) for publications and reports. Provided web services ranges from OPeNDAP (2014) based files, Sensor Observation Service (SOS 2014) to OGC Web Feature Services (WFS) and Web Map Services (WMS 2014).

For direct data access the portal performs concurrent requests to data providers. For data requests and transport the portal supports direct reading SQL access, e.g. the data warehouse of PANGAEA (2014), and Web Feature Service (WFS) requests, e.g. data retrieved from COSYNA (2014) and DOD (2014).

3.6 Terms of data access and good scientific practice

The data portal is based on open technologies and access is freely available for scientists, funding agencies and the public. So far no registration is required and user feedback is very welcome to further improve the functionality and content of the joint data portal.

Data citation is considered as good scientific practice. In this respect the downloaded datasets should always be cited and used with the appropriate citation given by the data provider. To that effect the dataset handle or DOI should always be provided along with the data citation for sustainable tracking.

4 Acknowledgement

The core team of the MaNIDA consortium consists of more than twenty scientific coworkers from all partner institutions and cooperation beyond, who mainly dedicate their own engagement and resources to shape and establish shared data infrastructures and workflows – a service oriented endeavor to the benefit of science. This continuous effort receives our full recognition and thank, especially in the context that data management and data publication usually do not gain sufficient incentives or awards by the scientific rating system of success. Thanks to the efforts of all MaNIDA cooperators the "Data Portal German Marine Research" could be realized and connects distributed data providers and major repositories for earth system and marine research in Germany. The MaNIDA project is mainly funded by the Initiative and Networking Fund of the Helmholtz Association (SO-071) and by a notable portion of own resources of all partners.

5 References

APACHE LUCENE: last visited: 2014-06-14.

BODC NERC (British Oceanographic Data Centre – Natural Environment Research Council): http://www.bodc.ac.uk/products/web_services/vocab/, last visited: 2014-06-14.

CFC (Climate and Forecast Conventions): http://cfconventions.org/, last visited: 2014-06-14.

COPERNICUS PUBLICATIONS: http://publications.copernicus.org/, last visited: 2014-06-14.

COSYNA (Costal Observing System for Northern and Arctic Seas): http://www.cosyna.de/, last visited: 2014-06-14.

DATA PORTAL GERMAN MARINE RESEARCH: http://manida.awi.de, last visited: 2014-06-14.
DOD (German Oceanographic Data Centre): http://www.bsh.de/en/Marine_data /Observations/DOD_Data_Centre/, last visited: 2014-06-14.
DOI (Digital Object Identifier): http://www.doi.org/, last visited: 2014-06-14.
DSHIP Data Acquisition System: http://www.werum.de/plattformen/dship.jsp, last visited: 2014-06-14.
ELASTICSEARCH: http://www.elasticsearch.org/, last visited: 2014-06-14.
ELSEVIER: http://www.elsevier.com/, last visited: 2014-06-14.
EMODNET (European Marine Observation and Data Network): http://www.emodnet.eu/, last visited: 2014-06-14.
EPIC (electronic publication information center): http://epic.awi.de/, last visited: 2014-06-14.
EUROFLEETS: http://www.eurofleets.eu/, last visited: 2014-06-14.
GBIF (Global Biodiversity Information Facility): http://www.gbif.org/, last visited: 2014-06-14.
GEOSS (Global Earth Observation System): https://www.earthobservations.org/geoss.shtml, last visited: 2014-06-14.
HZG – Publication database: http://141.4.217.215/fmi/xsl/publikat/Search.xsl, last visited: 2014-06-14.
ICSU WDS (International Council for Science World Data System): https://www.icsu-wds.org/, last visited: 2014-06-14.
INSPIRE (Infrastructure for Spatial Information in the European Community): http://inspire.ec.europa.eu/index.cfm/pageid/101, last visited: 2014-106-14.
MANIDA (Marine Network for Integrated Data Access): http://manida.org, last visited: 2014-06-14.
MARUM PUBLICATIONS: http://publications.marum.de/, last visited: 2014-06-14.
OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting): http://www.openarchives.org/pmh/, last visited: 2014-06-14.
OCEANREP: http://oceanrep.geomar.de/, last visited: 2014-06-14.
OGC CSW (Open Geospatial Consortium Catalogue Service for Web): http://www.opengeospatial.org/standards/cat, last visited: 2014-06-14.
OGC WFS (Open Geospatial Consortium Web Feature Service): http://www.opengeospatial.org/standards/wfs, last visited: 2014-06-14.
OPENDAP: http://www.opendap.org/, last visited: 2014-06-14.
PANGAEA – Data Publisher for Earth & Environmental Science: http://www.pangaea.de/, last visited: 2014-06-14.
SCIENCE DIRECT: http://www.sciencedirect.com/, last visited: 2014-06-14.
SEADATANET: http://www.seadatanet.org/, last visited: 2014-06-14.
SOS (Open Geospatial Consortium Sensor Observation Service): http://www.opengeospatial.org/standards/sos, last visited: 2014-06-14.
WMS (Open Geospatical Consortium Web Map Service): http://www.opengeospatial.org/standards/wms, last visited: 2014-06-14.
WORMS (World Register of Marine Species): http://www.marinespecies.org/, last visited: 2014-06-14.