HORIZON 2020

Research and Innovation action

Grant Agreement No. 730965

ARICE: Arctic Research Icebreaker Consortium:

A strategy for meeting the needs for marine-based research in the Arctic

Deliverable 7.6

Data Management System and documentation

Submission of Deliverable

Work Package	WP7: Enhancing Virtual and remote access to data
Deliverable no. & title	7.6 Data Management System and documentation
Version	1st draft
Creation Date	06-05-2021
Last change	
Status	☐ Draft ☒ WP lead accepted ☒ Executive Board accepted
Dissemination level	☒ PU-Public ☐ PP- Restricted to programme partners ☐ RE- Restricted to a group specified by the consortium ☐ CO- Confidential, only for members of the consortium
Lead Beneficiary	CNR
Contributors	☐ 1 – AWI, ☐ 2 – SPRS, ☐ 3 - NPI, ☐ 4 - ULAVAL, ☐ 5 – UAF/CFOS, ☐ 7 – CSIC-UTM, ☒ 8 – CNR, ☐ 9 - WOC, ☐ 10 – IOPAN, ☐ 11 – FMI, ☐ 12 - CNRS, ☐ 13 – NERC-BAS, ☐ 14 – DTU-AQUA, ☐ 15 – ARCTIA
Due date	31 March 2021
Delivery date	05 July 2021

Contents

1. Abstract

This report presents the main components and functionalities of the Data Management System (DMS) implemented by ARICE. The strategic targets driving this development have been: (i) the need to simplify the original plan to address challenges arising from changes in involved partners during the project; (ii) the need, in a legacy perspective, to adopt technical solutions and schemes sustainable in time, and (iii) to take in to account the last developments, at least at European level, made by organisations that are data aggregators and/or data providers. The last two points are relevant in order to accomplish ARICE project goals in a very cost-effective way. The use of consolidated software packages in the data management panorama at international level and sustained by large communities (e.g. GEOCAT for GeoNetwork), together with the interaction with EMODnet Physics[1], are the key pillars ensuring that strategic goals will be achieved.

2. Introduction

There is an increasing need and demand from the science community, policymakers, stakeholder's and the general public for timely access to information and data on the Arctic Ocean. One of the main objectives of ARICE is therefore to establish a data management system (DMS) to enhance virtual and remote access to data across this extensive user community. The general goals of this DMS are (i) to provide seamless access to the data generated by ARICE cruises; (ii) to develop and adapt strategies and tools for efficient data access and data dissemination, to help a wider use of data acquired by research icebreakers; (iii) to contribute to the general goal of improving the availability and readiness of data and information about the status of Arctic Ocean. Accomplishing these targets means to implement functionalities for interoperability, standards compliance, user accessibility, metadata discovery, and open access.

The ARICE data management system's original design was provided in Deliverable 7.4, where data products, data flow and system specifications were identified, mainly based on information collected/elaborated and reported in Deliverables 7.1, 7.2 and 7.3. At that time, the system focused on data collected from ARICE cruises from research vessels, to mainly address the first two goals mentioned above. The DMS was intended to provide as well both real time data and delayed data through a unique innovative 3D icebreaker platform.

The transfer of WP7 responsibilities from Arctic Portal led to the elaboration of a new simplified strategy and approach. In the new strategy, the real-time (NRT) data from icebreakers (for which the 3D icebreaker interface is really useful) has been separated from the data provided in delayed mode. By doing this, the functionality of the planned DMS moves to a Metadata Management System (MMS), and related work can be simplified while securing the data produced by ARICE TNA activities. The new ARICE DMS scheme is graphically represented by the figure below that replaces the Figure 1 in D7.4.

1 EMODnet Physics is one of the seven domain specific EMODnet projects supported by DGMare running operational services where ocean physics data and data products built with common standards can be found, viewed and downloaded.

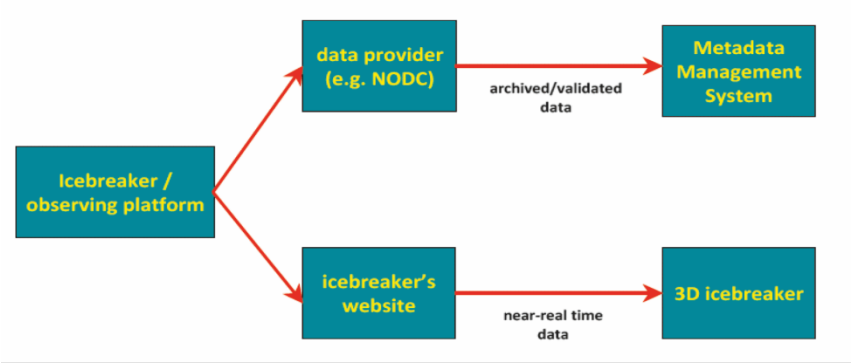

Figure 1 - The new ARICE Data Management Scheme proposed by CNR.

The new plan developed by CNR maintains the metadata-driven approach and the innovative solutions of the 3D icebreaker, but simplifies the work separating the data management system from the 3D icebreaker. The overall idea is to emphasise the different scopes and target users: The data management system is devoted mainly to scientific users and/or experienced stakeholders interested in data; while the 3D icebreaker is devoted mainly to disseminate information on the capacities for research of icebreakers, on their scientific equipment and their activities to a broader audience.

In a medium- and long-term perspective the two access platforms could be easily connected.

3. The ARICE Data Management System

The operability of any DMS depends mainly on the choices made on data flow characteristics as well as on the level of services provided to the users to discover and handle the data, where "handling" in this context means the possibility to better understand the information made available (e.g. by visualisation) and to download the data.

Implementing an efficient discovery functionality following FAIR principles and facilitating machine-to-machine (M2M) interaction means to include an application able to manage spatially referenced resources in the DMS. As reported in detail in the next section, the ARICE DMS integrates for this the open-source application GeoNetwork, which provides powerful functions such as metadata editing and searching, as well as an interactive web map viewer. GeoNetwork is a flexible tool that allows an agile management of catalogues and then makes data discovery easy and user-friendly. However, it is much less powerful and performing when we move to handling the datasets (harvesting and downloading data). To allow the DMS to function as a repository and not only as a catalogue, we need to be able, for example, to offer data in different formats, in response to REST (REpresentational State Transfer) calls in M2M interactions, or to manual queries by an operator/user. Furthermore, although GeoNetwork supports the presentation of interactive maps, it does not allow a direct data representation such as time series or profiles.

Other components are thus necessary to manage, organise and present data in response to different users and through different data disclosure channels, including webGIS solutions. Figure 2 presents the designed architecture that addresses and solves this issue, based on two data servers (ERDDAP and GEOSERVER) and a dedicated webGIS Map viewer application.

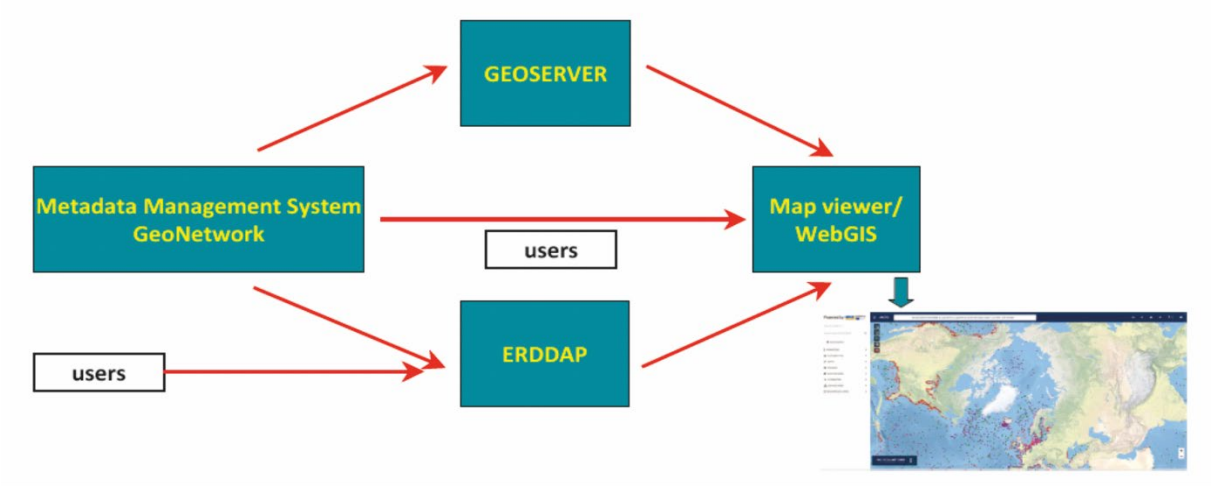

Figure 2 - Components of DMS for data "handling".

ERDDAP acts then as an intermediary data server: when requesting data, ERDDAP reformats the query in the format requested by the remote server, sends the data query to the remote server, obtains the data, and reformats the data back to send the data to the user. In this way, ERDDAP returns the data in the common file format of the user's choice.

GeoServer is a data server created to view and modify geospatial data. The mapviewer is a web tool that allows the interaction between the user and the data. Through the interface, the user can make selections by parameter, theme, etc. and access the selected data.

More information on these components as well as on the catalogue application GeoNetwork will be provided in the next section. Here have to stress that the choices of data server were based on the developments of the data aggregators and service providers landscape at European level (1), and in particular to the development of the Arctic marine portal by EMODnet physics (2) and a map service for SOOS (3).

4. Software applications and system functionalities

Below the characteristics of the ARICE DMS components are presented in detail.

GeoNetwork

The software used by the DMS as catalogue application for managing spatially referenced resources is GeoNetwork Open Source, a mature software that is currently used in numerous Spatial Data Infrastructure initiatives across the world involving FAO, WHO, WFP and many others, including EMODnet Physics.

GeoNetwork provides an easy-to-use web interface to search geospatial data (through their metadata) across multiple catalogs. The search provides full-text search as well as faceted search on keywords, resource types, organisations, and scale, among others. Users can easily navigate across records, refine the search, and find sources or services publishing a dataset, and quickly get the records of interest.

GeoNetwork allows to manage the insertion, editing, and maintenance of metadata in various formats including Dublin Core (usually used for open data portals), ISO19139, ISO19115/119/110 and permits validation of records against ISO and INSPIRE rules. Based on user profiles (e.g. reviewer, editor), a dashboard provides easy access to their information and tasks. Online editing of metadata is based on a powerful template system and directories of information (e.g. contacts, thesaurus). The editor allows uploading of data, graphics, documents, pdf files and any other content type. It supports among others:

• multilingual metadata editing,

• validation system,

• suggestion to improve metadata quality,

• geo-publication of layers to publish geodata layers in OGC services (eg. GeoServer).

It combines the following complementary and integrated functionalities:

- Global library for geospatial data

- Metadata catalogue describing geospatial data and thus enabling users to assess suitability to their analysis needs

- System for searching, editing, and publishing geospatial information

- Service that allows the integration of geospatial data from various sources

An administration console provides quick access to the system configuration. This allows the easy management of user and group accounts, classification systems from the web interface and scheduling metadata harvesting from many sources including:

• OGC-CSW 2.0.2 ISO Profile,

• OAI-PMH,

• Z39.50 protocols,

• Thredds,

• Webdav,

• Web Accessible Folders,

• ESRI GeoPortal

• Other GeoNetwork node.

Built-in schema.org Dataset and Catalogue annotations allow to make datasets searchable through Google Datasets. Figures 3 and 4 provide ARICE catalogue search and product page, respectively.

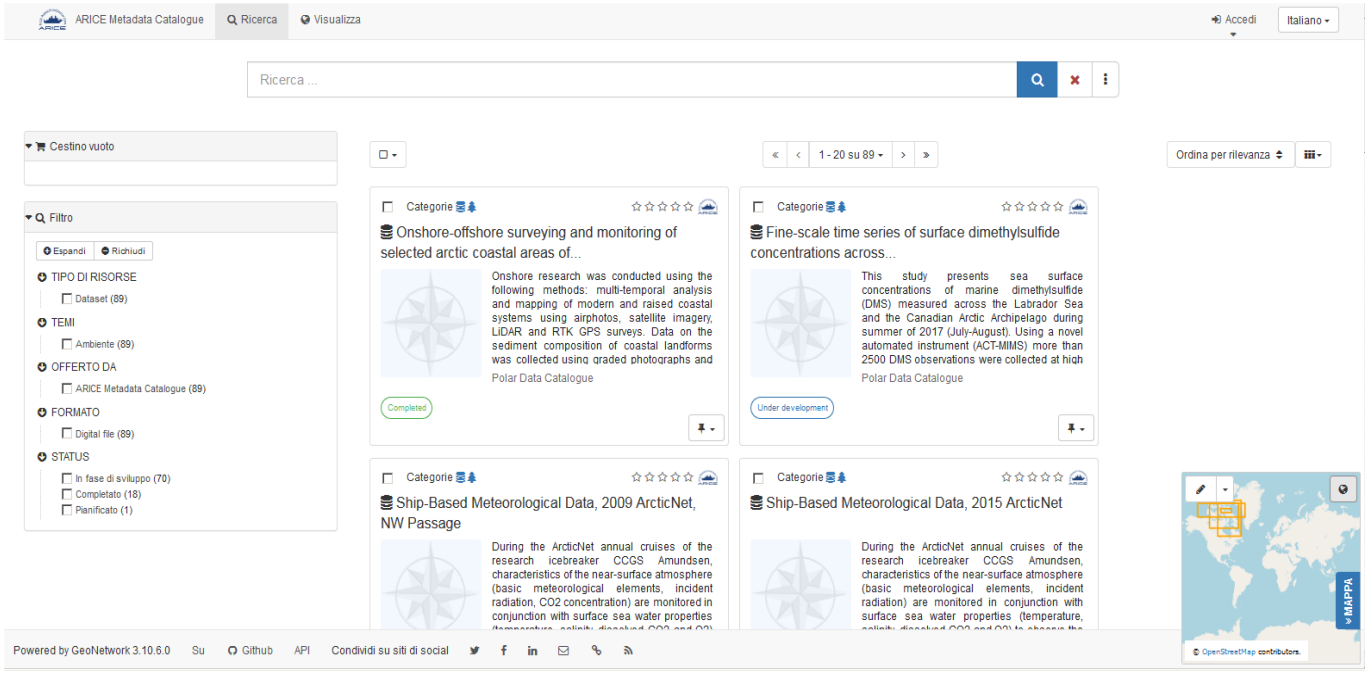

Figure 3 - ARICE catalogue search page

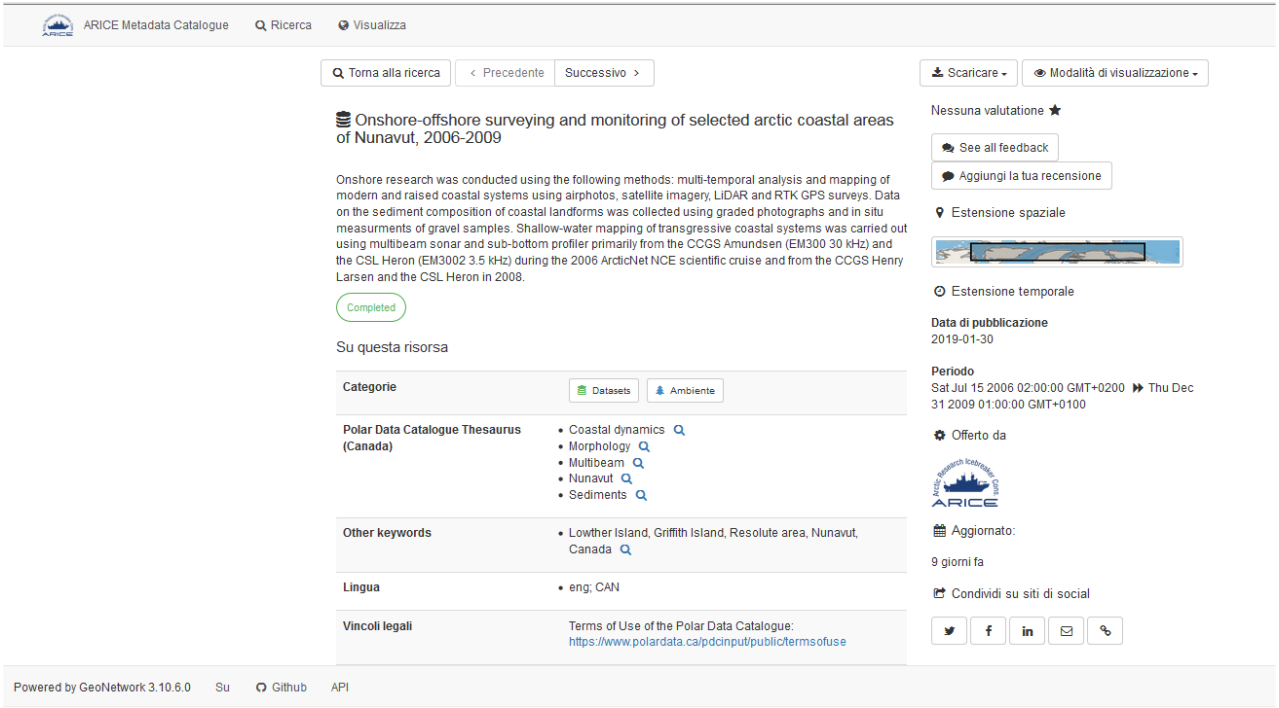

Figure 4 - ARICE catalogue product page

ERDDAP

ERDDAP is a data server that provides a simple, consistent way to download subsets of scientific datasets in common file formats and create graphs and maps. The different scientific communities have developed different types of data servers and each one uses its own protocol and data format. ERDDAP solves the problem by unifying the different types of data servers to have a consistent way to get the data in the desired format. As indicated in the previous section, ERDDAP acts as an intermediary data server: when requesting data, ERDDAP reformats the data query in the format requested by the

remote server, sends the data request to the remote server, obtains the data, reformats the data in the requested format, and sends the data to the user.

ERDDAP can get data from local (on the server's hard drive) and remote (accessed via the web) data sources, can serve many types of scientific data (html table, ESRI .asc and .csv, Google Earth .kml, binary OPeNDAP, .mat, .nc, ODV .txt, .csv, .tsv, .json and .xhtml) and offers several ways to search for it. ERDDAP can also return a .png or .pdf image with a custom chart or map.

ERDDAP provides standardised ways to get data (OPeNDAP and REST), can add and modify metadata, standardises the variable names and units for longitude, latitude, altitude, depth and time in the results, and can represent the data in graphs and maps that can be easily downloaded or embedded in other websites.

ERDDAP is free and open source. It uses Apache-like licenses, and its appearance is customisable to reflect the institution choosing to use it. If an institution already has a web service for distributing data, ERDDAP can access the data via the existing service or via the source files or database. As benefit, people will have another way to access the data and will be able to download it in additional file formats or as graphs or maps.

ERDDAP has been installed by over 85 organisations in at least 14 countries (Australia, Belgium, Canada, China, France, India, Ireland, Italy, New Zealand, South Africa, Spain, Thailand, UK, USA), including: APDRC (Asia-Pacific Data-Research Center, International Pacific Research Center) at the University of Hawaii (UH), NOAA, EMODnet Physics and many others.

GEOSERVER

GeoServer is a data server created to view and modify geospatial data. Is an open source software server written in Java. Designed for interoperability, it publishes data from any major spatial data source using open standards. Being a community-driven project, GeoServer is developed, tested, and supported by a diverse group of individuals and organisations from around the world.

It supports Open Geospatial Consortium (OGC) standards such as Web Feature Service (WFS), Web Map Service (WMS) and Web Coverage Service (WCS). Additional publishing formats and options are available as extensions including Web Processing Service (WPS) and Web Map Tile Service (WMTS).

GeoServer started in 2001 by The Open Planning Project (TOPP), a non-profit technology incubator based in New York. Those involved with GeoServer initiated the GeoTools project, an open-source GIS Java toolkit. GeoTools adds support for shapefiles and Oracle databases, among others. Around the same time as GeoServer was founded, the OpenGIS Consortium (now the Open Geospatial Consortium) was working on the Web Feature Service standard. It specifies a protocol to make spatial data directly available on the web, using GML (Geographic Markup Language), an interoperable data format. A Web Map Service was also created, a protocol for creating and displaying map images created from spatial data. Other projects became interrelated, providing to GeoServer other functionalities. GeoServer can now read data from over a dozen spatial data sources and output to many different formats.

GeoNetwork, ERDDAP and GEOSERVER provide a set of powerful and flexible technologies through which to implement interoperable services. Thanks to them, a wide spectrum of functionalities can be implemented in the ARICE Data Management System:

- A catalogue for requesting scientific data through metadata in the most common formats of use (DIF, ISO, D-CAT, JSON ...).
- Sharing of metadata through modern protocols (Geonetwork protocol, CSW, WAF).
- Adherence to INSPIRE guidelines.
- Use of popular and shared vocabularies.

- Data management in the most common formats of use (SQL database, ASCII files, CSV files, netCDF ...).
- Sharing of data through modern protocols (REST API, OPeNDAP ...).
- Tools for downloading, viewing, and converting data.

Considering that the focus of the M2M services to the users is on making it easier for users to get scientific data, ARICE DMS includes a map viewer, a web tool that allows interaction between the user and the data. Map Viewer is a touch-optimised web application designed to make maps accessible on a large set of devices, including smartphones, tablets, laptops and desktop computers. Map Viewer give access to maps and information simply using the browser. Through the interface, the user can make selections by parameter, theme, etc. and access the selected data.

Implementation of data "handling" functionalities based on ERDDAP, GEOSERVER and map viewer are being achieving through cooperative activities with EMODnet Physics. The next section provides the motivation for this cooperation, and the elements sustaining the envisaged win-win approach.

5. Cooperation with EMODnet

EMODnet Physics integrates and makes available ocean physics data (Real Time, Near Real Time, Historical reprocessed & validated) and derived data products. Operational data are provided mainly by Copernicus CMEMS-INSTAC and EuroGOOS, meanwhile SeaDataNet is the main source for archived and validated data. These European integrators bundle and make available data originating from many oceanographic observing networks and activities. In addition, EMODnet Physics derives data and data products from other European and international providers, such as ICES, PSMSL, GLOSS, SONEL, IOC, PANGAEA, GDAC (Coriolis), OCEAN SITES, ARGO, etc. It links all these ocean data sources into a single discoverable data bank in which metadata are harmonised, and data are harmonised in formats, where possible.

Thanks to an effort started under the ur-EMODnet preparatory action (MARE/2010/02) and successive development (MARE/2012/10; EASME/EMFF/2016/006) and operational (EASME/2019/OP/0003) phases, EMODnet Physics has been successful in designing, organising and running operational services where ocean physics data and data products built with common standards can be found, viewed and downloaded in a way that is free of charge and free of restrictions of use.

Data collected through its federated structure of providers are made available through all interoperable technologies reported in the previous section and hence presented in a EMODnet Physics map viewer. The user can discover the available data and products on the map viewer and then use one of the available M2M interface or services to interact with data automatically. In addition to ERDDAP, GeoServer, GeoNetwork, and map viewer to extend further the users and use of its products, EMODnet Physics also implements web APIs and widgets.

In November 2020, EMODnet Physics held a webinar in which an Arctic Data Portal was launched to improve the availability, quality, timeliness, and accessibility of essential marine in situ data from the Arctic region. Despite the large amount of data included in the EMODnet catalogue and the federated structures of providers, among them observing platforms, there is a large gap in relation to research vessels and icebreakers. This results from a weak connection with the community operating these platforms and the difficulty to fully develop the approach of EMODnet based on data harmonisation for the catalogue through smart adapters procedures, also including, when necessary, suitable quality control and quality filter (QC/QF) procedures.

Considering all above, and in addition EMODnet Physics mandate to support regional ocean observing systems and larger integration of marine observing data, CNR approached EMODnet to verify the possibility to start of mutual beneficial cooperation. A possible Memorandum of Understanding (MoU)

is under negotiation, which shall both fulfil the legacy perspective of ARICE and which shall increase the users for icebreaker datasets. The last ARICE General Assembly approved the proposal made by CNR to develop this idea up to signing a MoU. The actual status of the negotiation is reported as an appendix to this deliverable.

6. Legacy from ARICE

As discussed previously, the goal behind this work is the development of a DMS solution for ARICE that can be supported in the future. The components of the system described in section 4 and the cooperation with EMODnet referred to in section 5 are the pillars through which this strategic objective can be achieved in a reasonable time and budget, also by extending the availability of data for the ARICE community to other data besides those available from icebreakers.

The project web page will represent, the privileged entry point to the ARICE Geonetwork instance and therefore to all metadata: those related to the dozen projects supported through the TNA action, as well as to the much more numerous historical datasets made available by repositories such as PANGEA or the Canadian POLAR DATA CATALOG.

The users will be able to reach the data both through the metadata catalogue and through a webGIS Map viewer application, thanks to the cooperation with EMODnet Physics. This work benefits from their experiences in developing the Marine Arctic data portal as well as the SOOS MAP (4). A dedicated application, as e.g. SOOS, and/or the use the implemented Marine Arctic Data Portal, will be analysed and discussed among the consortium Steering Board with EMODnet. Even the use of the Marine Arctic Data Portal as a dedicated Map Viewer application focusing only on ship data at both poles will be very useful and less dispersive for those looking for and interested only at these data. And, since EMODnet physics is mainly devoted to oceanographic and atmospheric data and less to biological parameters, it would allow to keep the ARICE DMS broader in terms of covered parameters and multidisciplinarity with respect to EMODnet. A dedicated map viewer would provide us with more flexibility to add dedicated filters such as “institution” or “year”, among others, in addition to those already implemented by EMODnet.

Thanks to this cooperation between ARICE and EMODnet, the current status of the Marine Arctic Data Portal Map viewer with respect to icebreakers datasets will be largely improved, since as of now this data is completely missing (see Figure 5).

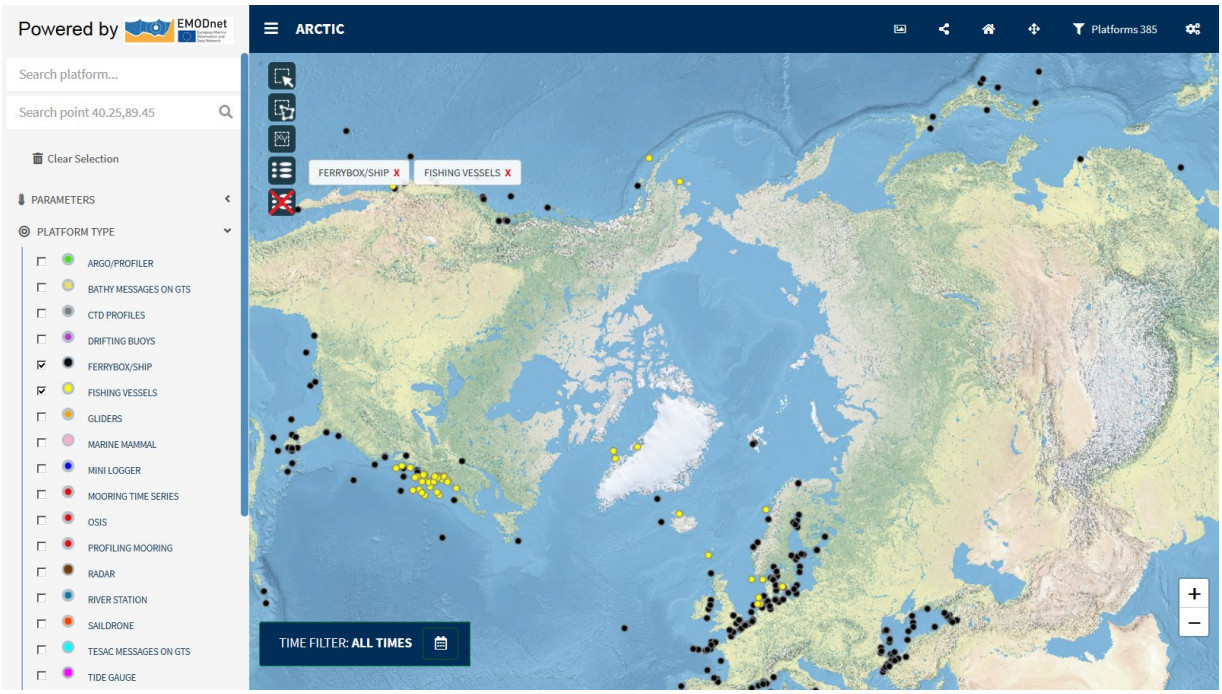

Figure 5 - Marine Arctic Data portal datasets provided by ships. The overall number of records is less than 400, none of them from icebreakers

The current lack of icebreaker datasets is due to the challenges of filtering and classifying these datasets and to the need of developing specific scripts to align the data to the platform needs. For example, the amount of metadata on PANGEA connected to PRV Polarstern are more than 30.000 and include measurements of large numbers of different parameters in different scientific disciplines. A strong interaction between the IT personnel and researchers is fundamental to develop data "handling" functionalities able to respond to the users' needs. This interaction is also important to guide IT personnel in a complex multidisciplinary landscape of parameters.

7. References

1- Mercator Ocean, EUROGOOS, and CMEMS partners, *CMEMS requirements for the evolution of the Copernicus In Situ Component - Version 1*, pp. 96, December 2018 (https://marine.copernicus.eu/sites/default/files/media/pdf/2020-10/CMEMS-requirements-In_Situ.pdf)

2- EMODnet news November 27, 2020 by admin (https://portal.emodnet-physics.eu/arctic-data-portal/)

3- SOOS Annual Report 2019, SOOSmap service, pp. 25-26

(https://www.soos.aq/images/soos/about_us/Annual%20Reports/SOOSAnnualReport-2019.pdf). See also https://www.scar.org/resources/soosmap/

4- https://www.soos.aq/data/soosmap

APPENDIX - Draft ARICE and EMODnet MoU

Background

ARICE – Arctic Research Icebreaker Consortium: A strategy for meeting the needs of marine-based research in the Arctic

The recent changes of the Arctic and the resulting increased economic activity in the region have triggered a societal demand for accurate sea-ice and weather predictions, information on the status of the Arctic Ocean and its marine life, and complex predictions of future scenarios. To address these issues and to develop policy recommendations for a sustainable usage of the Arctic Ocean and its resources, the international Arctic science community must have access to world-class research icebreakers to access the ice-covered Arctic Ocean.

Even if European Arctic research has contributed critical knowledge to identifying the processes behind these rapid changes, datasets from the Arctic Ocean are still insufficient to fully understand and more effectively predict the effects of climate change. The largest gaps in our knowledge and understanding of the Arctic system processes are outside the summer season, when the Arctic Ocean is logistically and technologically extremely difficult to reach.

Currently, the lack of available research icebreakers from Europe and beyond that can operate yearlong in the ice-covered Arctic Ocean, and a weakly coordinated polar research fleet impedes Europe's capacity to investigate the changing Arctic. There is thus an urgent need for providing polar scientists with better research icebreaker capacities for the Arctic, to address the knowledge gaps and to develop policy recommendations for a sustainable usage of the Arctic Ocean and its resources.

This need has now been taken up by the EU funded Arctic Research Icebreaker Consortium (ARICE).

Purpose

This MOU describes the roles of ARICE and EMODnet-Physics in developing an ARICE map viewer – an interactive map-based portal of observing data and platforms to match ARICE needs. It will deliver data and metadata from centrally curated datasets of circumpolar oceanographic observations and platforms.

The above goal will be accomplished by undertaking the following activities:

The CNR ISP on behalf of ARICE will identify datasets for inclusion in viewer.

The CNR ISP data officer will negotiate access to those datasets.

EMODnet-Physics staff will conduct the web development work to serve the data in the viewer and other EMODnet portals, as appropriate.

The CNR ISP data officer will test new developments in the viewer and provide feedback on potential improvements.

If EMODnet is unable to continue to support ARICE, the underlying code and data flows will be transferred to another data centre, with an appropriate host being negotiated with ARICE.

Funding

This MoU is a statement of mutual cooperation and in-kind support between ARICE and EMODnet Physics. It does not include any commitment of funds from either party.

Duration

This MOU is at-will and may be modified by mutual consent of authorised officials from ARICE and EMODnet Physics. This MOU shall become effective upon signature by the authorised officials from ARICE and EMODnet Physics and will remain in effect until modified or terminated by any one of the partners by mutual consent. In the absence of mutual agreement by the authorised officials from ARICE and EMODnet Physics this MOU shall end on December 31st 2024.