SCIENCE AND TECHNOLOGY ORGANIZATION
CENTRE FOR MARITIME RESEARCH AND EXPERIMENTATION

Reprint Series **CMRE-PR-2019-121**

From data to decision: an overview of CMRE's information supply chain

Alessandro Berni, Giampaolo Cimino, Elena Camossi, Simone Giannecchini

June 2019

Originally published in:

Proceedings of the OCEANS 2015 MTS/IEEE Conference, 18-21 May 2015, Genoa, Italy, doi: 10.1109/OCEANS-Genova.2015.7271408

About CMRE

The Centre for Maritime Research and Experimentation (CMRE) is a world-class NATO scientific research and experimentation facility located in La Spezia, Italy.

The CMRE was established by the North Atlantic Council on 1 July 2012 as part of the NATO Science & Technology Organization. The CMRE and its predecessors have served NATO for over 50 years as the SACLANT Anti-Submarine Warfare Centre, SACLANT Undersea Research Centre, NATO Undersea Research Centre (NURC) and now as part of the Science & Technology Organization.

CMRE conducts state-of-the-art scientific research and experimentation ranging from concept development to prototype demonstration in an operational environment and has produced leaders in ocean science, modelling and simulation, acoustics and other disciplines, as well as producing critical results and understanding that have been built into the operational concepts of NATO and the nations.

CMRE conducts hands-on scientific and engineering research for the direct benefit of its NATO Customers. It operates two research vessels that enable science and technology solutions to be explored and exploited at sea. The largest of these vessels, the NRV Alliance, is a global class vessel that is acoustically extremely quiet.

CMRE is a leading example of enabling nations to work more effectively and efficiently together by prioritizing national needs, focusing on research and technology challenges, both in and out of the maritime environment, through the collective Power of its world-class scientists, engineers, and specialized laboratories in collaboration with the many partners in and out of the scientific domain.

NOTE: The CMRE Reprint series reprints papers and articles published by CMRE authors in the open literature as an effort to widely disseminate CMRE products. Users are encouraged to cite the original article where possible.

From Data to Decision

An Overview of CMRE's Information Supply Chain

Alessandro Berni, Giampaolo Cimino, Elena Camossi
NATO STO Centre for Maritime Research and Experimentation (CMRE)
La Spezia, Italy
Alessandro.Berni@cmre.nato.int

Simone Giannecchini
Geosolutions S.a.S.
Massarosa, Italy
Simone.Giannecchini@geo-solutions.it

***Abstract*— Proper data curation is a prerequisite for the creation of long-lasting value from scientific datasets and for the generation of authoritative data-driven decision support tools. Availability of detailed information on instruments calibration and other useful metadata are essential for effective data use and re-use. High quality and standardized metadata, in turn, allows effective data discovery through web-based data catalogues, linking interoperable data repositories and paving the way to federated data infrastructures on a global scale, to ensure that data are made available to scientists, decision makers and automated computing workflows. This paper presents the approach undertaken by the Centre for Maritime Research and Experimentation (CMRE) to provide a "common enterprise bus" in which scientific data and added-value products are the principal currency.**

I. Introduction

Research bodies around the world have acquired over time, and continue acquiring, significant amounts of highly valuable data (e.g. environmental, acoustic, geophysical, etc.). Some data are required in real time while some other can be of interest also in the future, as reference data sets for the testing and validation of new models and algorithms. The Centre for Maritime Research and Experimentation (CMRE), an executive body of NATO's Science and Technology Organization (STO), has a long-standing experience in scientific data management, to include collection from remote sensing platforms or autonomous underwater vehicles, data curation and discovery within "big data" repositories, and delivery of data-driven predictive models and decision support tools.

The collection and sustainment of scientific data is an expensive process, and there is a compelling business case to capture, in a reliable way, the elements related to data quality, data formats and sensor calibration information, together with the contextual information that enables the full understanding of the way the experiment was conducted, to allow subsequent analysis and validation of models. For this reason increased emphasis is being put on the development of coherent approaches aimed at the improved understanding of the current and forecast physical environment, delivering new value to stakeholders through the introduction of innovations in the fields of data acquisition, brokering and mediation, catalogue services, sensor web enablement and model web.

A number of basic building blocks have already been deployed, and others are being developed, to realize a vision in which "information supply chains", linking data acquisition to predictive models, are used to deliver added-value elements such as decision support systems informed by environmental and contextual information. The overall concept is aligned to implement the "observe, orient, decide, and act" (OODA) loop [1], an important concept in strategic decision-making which also finds application in the treatment of big societal challenges, such as those addressed by the Global Earth Observing System of Systems (GEOSS) [2].

Verification of the effectiveness of the approach is obtained through testing of the prototype tools in realistic situations (e.g. during the execution of at-sea experiments, synthetic interoperability exercises and field trials). Adherence to NATO (e.g. Allied Geographical Publications, STANAGS) and international standards (ISO, Open Geospatial Consortium – OGC) is key to experimentation of the tools with allied nations and with other international partners.

In this paper, the authors present the experiences made by CMRE in the development of an integrated data management infrastructure for scientific data and products. With a forward-looking perspective, the following objectives are being taken into account, to guide further integration and development:

- Provide essential functions related to the operation of CMRE's scientific data repository, by operating a shared infrastructure for data management that covers the whole data/information lifecycle, from acquisition to publishing and disposition, addressing issues of data provenance and identification, quality assurance, data lineage, and e-publication.
- Serve as CMRE's Virtual Research Environment, by enabling the definition and the execution of documented workflows (processing chains) for scientific data processing and modeling and simulation using the data available from the data repository as source.
- Sustain capacity-building, fostering opportunities for internal and external cooperation through a shared infrastructure providing processing capabilities and data management capabilities.
- Provide experiment support and reproducibility, by integrating the full experimental processing chain (including sea-going elements such as NATO's Research Vessels and Autonomous Vehicles) to provide effective development and sharing of trusted

This work was performed in the scope of project *Decisions in Uncertain Ocean Environment*, for Allied Command Transformation (ACT).

observational data and data-derived products, and producing as output e-publications that combine the typical format of the scientific report with links to processing and analysis services and data, to enable readers in NATO and Nations to gain access to unprecedented insight on the processes behind CMRE's scientific discovery.

- Provide Interoperability with Global Open Data Repositories, by adhering to NATO and international standards on data and metadata interoperability, to enable CMRE and NATO to take advantage of emergent data sources that individual Nations and international bodies, such as the Group of Earth Observations (GEO), have established, and continue developing, making available large amounts of highly valuable data (e.g. environmental, acoustic, geophysical, etc.) on a global scale.

The paper is organized as follows. Section II discusses interoperability, the importance of standards for data and metadata, and how compliance to standards is addressed at CMRE. Section III describes CMRE's approach to realize standard-based web processing services. Finally, Section IV concludes the paper outlining future developments.

II. The importance of interoperability

The understanding of the physical environment on a global scale passes through the acceptance and implementation of interoperability arrangements between resource providers, to specify the technical aspects of activities such as data collection and storage, as well those related to the dissemination of data, metadata and added-value products. Preference is typically given to non-proprietary international standards, focusing on component interfaces, to minimize impact to, and dependencies from, internal system implementations.

In spite of all progress made in recent years, the ocean is still an under-sampled dimension and no single institution is positioned to satisfy observational needs on a global scale. The adoption of standard data formats and protocols is therefore a powerful enabler, as suggested in 0, to expand the range of available data sources through the integration of external data repositories and catalogues.

From an information standpoint, interoperability has to be deployed at different levels, to include machine encoding and format encoding, as well as semantics and ontologies. In this section we will present some of the activities undertaken by CMRE to produce interoperable metadata and data in near real time, and to provide automated workflows connecting sensors and models in a seamless way.

A. Real time metadata generation

Metadata standards are complex in nature, and producing and validating records in compliance to a standard can be a demanding task. As manual processes are time-consuming and subject to human errors and quality issues, an investment was made in the creation of a continuous deployment service based on Python scripts and versioning software to extract automatically metadata from the most common data formats, sensor payload types and model output variants.

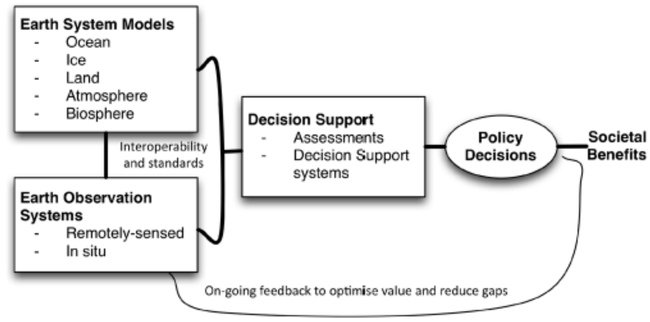

Fig. 1. GEOSS Observations and Models as example of OODA loop (adapted from 0)

During any scientific activity, be that a sea trial or a synthetic experiment, data are funnelled through various data repositories, using common protocols such as HTTP or FTP. A set of appropriate acquisition scripts, adapted to the variety of data that need to be considered, is launched at periodic intervals using a web-based scheduler (Jenkins) [3], to extract metadata and to load them in a catalogue (GeoNetwork) [4], while source data is stored in a geospatial web server (GeoServer) [5].

Quantity and quality of data types and formats vary from one experiment to another. Some changes are just unavoidable as they relate, e.g. to a new geographic area or to a new type of instrument. Some others could be related to human factors, such as differences in data formats, measurement units, changes in the name of the variable measured, change of scientific approach, etc. Machine-independent data formats such as NetCDF [6] and its extension NetCDF-CF [7], are extremely valuable to limit the spread of such inconsistencies, however, in some cases this variability needs to be handled directly at the level of the acquisition system.

CMRE's integrated scientific data management system includes a production-quality Version Control System (Git) [8] to handle different versions of the acquisition scripts, in order to deal with different "variants" of incoming datasets. Each experiment has a dedicated branch in the Git version control tree, and the scheduler has a dedicated task to manage the experiment. At the start of the experiment the appropriate code is checked out from Git and then executed. Execution starts from a main script that, in turn, lunches a dedicated acquisition script for each data type. In a nutshell, each acquisition script performs the following actions:

1. Connect to the data repository;
2. Check if a new data has arrived since the last iteration; if new data is available, download the dataset;
3. Extract the data, perform data transformations (if necessary) and load the result into Geoserver;
4. Generate the metadata (as ISO 19115 [9][10] and ISO 19139 [11], extended with NATO AGeoP-8 [12]) from the dataset files and load the newly created standards-based XML file into GeoNetwork.

Once the cycle is completed, data have been loaded in Geoserver and become available for use (and re-use), e.g.

through OGC Web Services (OWS) [13]. Discovery can be made by users, or by processing services, such as models, that require fresh METOC information for their initialization, by querying the data catalogue.

Scripts developed to date allow the near real time extraction of metadata for the following data types:

- AVHRR METOP remote sensing (Cloud Coverage, Sea Surface Temperature)
- AVHRR NOAA remote sensing (Cloud Coverage, Sea Surface Temperature)
- Research vessel navigation log files (GPS, Radar, AIS)
- CTD (Water Column Properties)
- Glider Data (Telemetry, Water Column Properties)
- ADCP (Currents)
- Surface Drifters (Sea Surface Temperature, Surface Currents)
- Oceanographic model output (NetCDF format)

B. *Data catalogues*

A data catalogue that can be queried with standard protocols, such as the Catalogue Service for the Web (CSW) [15], allows the different entities in the "information supply chain" to discover the "best" data to achieve a given purpose.

An example use case is provided in Fig. 2. A data consumer (e.g. a model) issues a CSW request to the catalogue (1) to find, e.g., all "wind speed" grids in an area of interest.

The catalogue responds (2) with matching metadata, enabling the model to choose the most relevant data source from those available. A OWS request is then sent to the geospatial server (3), which provides the required layers (4).

Still, protocols such as CSW are suitable for machine-to-machine communication, but not for human-machine interaction. This makes it difficult to address situations in which a scientist needs to "explore" available datasets answering questions such as "what data is available about that geographical area?" or "what types of data were collected during that experiment?"

To address that use case a user-friendly portal, built on the Comprehensive Knowledge Archive Network (CKAN) [17], is used as an interface for data discovery by humans, to perform search actions in a simplified and effective manner (Fig. 3), with no need for background information such as the details of the underlying archival structure or data provenance and quality.

The use of the CKAN portal allows the sharing of the results of raw and derived data, such as model results, for all kinds of users, including those that do not possess big technical skills. Data are published on few minutes after they become available in the CSW catalogue, through recourse to frequent harvesting, and users can use simplified search functionalities such as fuzzy search, faceted browsing, and spatial-temporal filtering of datasets to find the information of interest. Preview of related data is provided with a WMS query [14] to the geospatial server. A web link is also provided to download full datasets, with the possibility of enforcing fine-grained role-based access control.

C. *Data Interoperability*

Finding the references to the datasets of interest through a catalogue is not sufficient, as data have to be provided in a format accepted by user's tools. Although standards do not extend to the point of detailing all possible data formats, scientific communities have developed and adopted best practices and open formats such as NetCDF from Unidata [6]. The NetCDF format is machine-independent, can be handled through a variety of open source tools and its popularity spreads several communities, from Earth Observation to Oceanography. Still, being the format open and self-describing, inconsistencies may arise, resulting in detriment to data usability. To address such issues, harmonization initiatives such as Climate and Forecast conventions [7] are being undertaken, to augment NetCDF by defining a common vocabulary for climate and forecast applications, standardising the names of variables and defining spatio-temporal properties of data.

Another degree of freedom is left in the granularity of datasets. In most cases, there is no need to archive directly the single observations, and some forms of aggregation, e.g. by trajectory or time series, provides a better match to requirements for the implementation of models and catalogues.

Data management software such as Unidata THREDDS [18] are also available, to convert data on the fly and to support subset selection of datasets and aggregation, as a way to address efficiently issues related to data aggregation needs.

III. PROCESSING SERVICES

Web Processing Service (WPS) [16] is an OGC Specification providing XML-based Web Service interface to enable clients to remotely execute computation jobs via HTTP. The standard provides a way to define input and output parameters (with specific focus on geospatial data types) and supports long-running transaction.

Often tasks executed via WPS are computationally demanding and result in implementations that do not scale efficiently: tasks are usually executed locally to the WPS process implementation, and the only way to scale is to add WPS replicas.

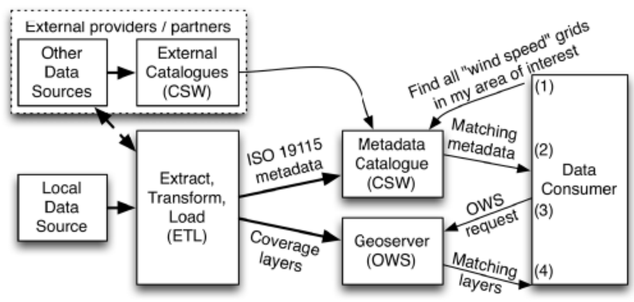

Fig. 2. **Augmentation of data consumer capabilities through CSW data discovery**

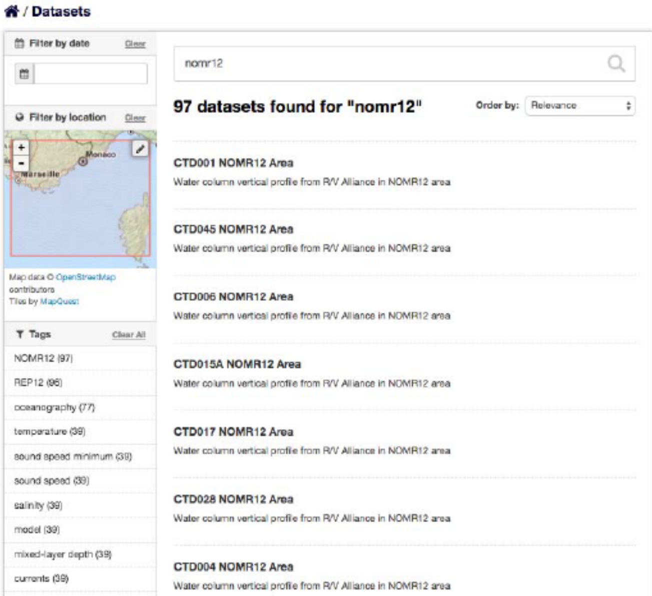

Fig. 3. CMRE's data catalogue interface

CMRE, starting from the Geoserver WPS implementation, implemented a remote execution model for WPS computation jobs where the WPS server is connected to a set of remote machine (computational nodes) via XMPP [19].

The computational nodes host a service script (Python-based) that handles the communications with Geoserver WPS and the set of algorithms to be invoked via WPS, which can be of any type, as long as they can be invoked from a command line interface.

Once a new request is received, it is directed to one computational node using a simple load-balancing algorithm. The execution is then performed on the designated computational node, thus unloading the WPS server.

During the execution, progress information and errors/warnings, should something go wrong, are sent via XMPP to the WPS server. XMPP is also used to detect the presence of a new computational node and to register the algorithms that it supports (together with relevant input-output parameters) into the WPS server.

This approach scales very well and allows discovery of new computational nodes even in complex distributed architectures and cloud computing frameworks.

Registration of a new functionality on WPS is just a matter of deploying the code on the computational nodes and defining the input-output parameters. The specific WPS operations ("Get Capabilities", "Describe Process" and "Process Execution") will be updated automatically.

As the number of computational nodes grows, manual installation of the service scripts and the algorithms on the computational nodes can soon become inefficient. Scalable results can easily be obtained through the use of automatic distributed application deployment tools such as Chef [20], Docker [21] or Puppet [22].

1) ROMS on the cloud

An experimental evaluation has been carried out in the context of the Global Earth Observation System of Systems – Application Implementation Pilot 7 (GEOSS-AIP7) deploying a parallel implementation of the Regional Ocean Modeling System (ROMS) [23] on a cloud environment, to evaluate the potential of high performance computing solutions to support the sustained delivery of a continuous forecasting service capability. The different steps of the model execution (setting of the environment, getting input data from catalogue, model initialization, run in operational mode and ensemble mode) have been deployed on a cloud environment, leveraging WPS as service interface. The experiment will be extended in GEOSS-AIP8 to address data assimilation with data coming OGC Sensor Observation Service [24].

IV. Conclusions and Future Work

The elements developed so far can be considered as the fundamental building blocks of a comprehensive data management architecture to enable Centre work and collaborative research with partners.

From a functional point of view, future evolutions will include core infrastructural services to guide quality assessment, and fine-grained access control, together with higher-level functions for the creation of new knowledge e.g. through data discovery, processing, and workflow automation.

The data management system will be developed to include Virtual Research Environment extensions, to provide collaborative tools and information-driven functionalities for data acquisition, processing and data-driven analysis (as implementation of feedback loops), as well as interfaces to modeling & simulation systems, to enable combination of "real world" with "synthetic worlds" (Fig. 4).

Fig. 5 illustrates the information viewpoint of the system. The information drivers of the system (in green) result from the application of information-driven services (in blue). NATO policies for Information Management (NIMP), procedures and licensing provide the normative background required to balance "need to share" with "need to know" (e.g. respect of intellectual property caveats).

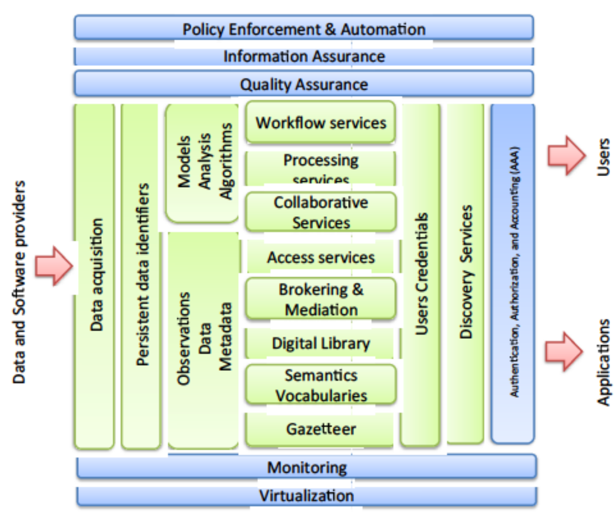

Fig. 4. Integrated Scientific Data management Architecture

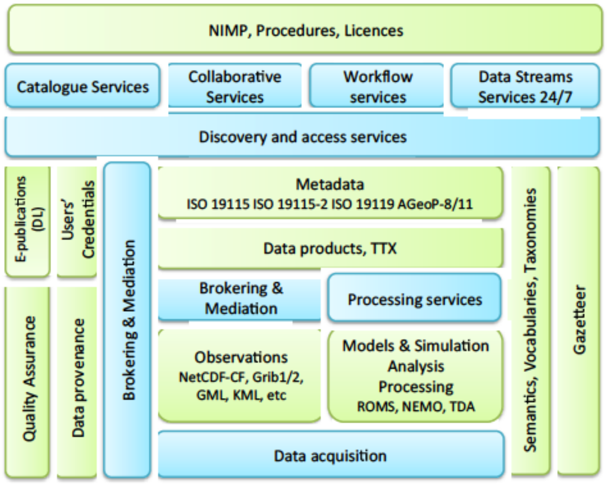

Fig. 5. Information viewpoint of Integrated Scientific Data Management

One of the key objectives of the envisioned developments is the federation with global networks of ocean observations systems, to pave the way towards the development of global ocean modelling services for continuous monitoring and forecasting of ocean conditions.

Relevant initiatives in this field include the Global Ocean Observing System (GOOS) [25] is a permanent facility for ocean observations, modelling and analysis and is part of GEOSS [5]. The GEOSS portal leverages the GI-Cat brokering catalogue, [26] a suite of brokering services to federate Earth Observation catalogues. GI-Cat can access a number of catalogue service software[1], and relies on ISO 19115 as internal metadata model for harmonization. Analogously, it can serve metadata harvestable by various catalogue interfaces. A similar brokering approach is used by the National Oceanic and Atmospheric Administration (NOAA) geoportal [27].

Other global and European oceanographic catalogues developed in the context of the GEOSS/GOOS initiative have resulted by European Union projects: e.g., the pan-European infrastructure for ocean and marine data management SeaDataNet [28], the European Marine Observation and Data Network (EMODnet) [29] and the Copernicus Marine Service and its pre-operative service MyOcean [30]. These systems adopt ISO 19115 and its XML specification ISO 19139 for metadata harmonization, and are designed to be compliant to the European Directive establishing the implementation of an Infrastructure for Spatial Information in the European Community (INSPIRE) [31][32].

CMRE intends to continue and to develop further the engagement with the international community, to share technical approaches and experiences towards the improvement of our understanding of the marine environment, transforming raw data into effective tools to support informed decisions.

Acknowledgment

The authors thank Dr Rich Signell of USGS for all the encouragement, support and helpful technical discussions.

[1] GI-cat supported catalogue services: http://essi-lab.eu/do/view/GIcat/GIcatDocumentation

References

[1] F. Osinga, "Science, Strategy and War: The Strategic Theory of John Boyd", Routledge, 2007
[2] Global Earth Observation system of Systems (GEOSS), "10-Year Implementation Plan Reference Document", GEO 1000R / ESA Sp-1284, February 2005
[3] Jenkins CI, https://jenkins-ci.org (accessed in March 2015)
[4] Geonetwork, http://geonetwork-opensource.org (accessed in March 2015)
[5] Geoserver, http://geoserver.org (accessed in March 2015)
[6] Unidata Network Common Data Format (NetCDF) http://www.unidata.ucar.edu/software/netcdf/ (Accessed in March 2015)
[7] Climate and Forecast conventions and metadata http://cfconventions.org/ (Accessed in March 2015)
[8] Git, http://git-scm.com/ (Accessed in March 2015)
[9] ISO 19115-1:2014 Geographic information -- Metadata -- Part 1: Fundamentals
[10] ISO 19115-2:2009 Geographic information -- Metadata -- Part 2: Extensions for imagery and gridded data
[11] ISO/TS 19139:2007 Geographic information -- Metadata -- XML schema implementation STANAG 2586 Ed01
[12] AgeoP-08, NATO Geospatial Metadata Profile (NGMP)
[13] Open Geospatial Consortium Reference Model OGC 08-062r7 ver 2.1, December 2011, http://www.opengis.net/doc/orm/2.1 (Accessed in March 2015)
[14] OGC Web Map Service, http://www.opengeospatial.org/standards/wms (Accessed in March 2015)
[15] OGC Catalogue Service, http://www.opengeospatial.org/standards/cat (Accessed in March 2015)
[16] OGC Wep Processing Service, http://www.opengeospatial.org/standards/wps (Accessed in March 2015)
[17] The Open Knowledge Foundation: CKAN, The Open Source Data portal http://ckan.org (Accessed in March 2015)
[18] Unidata THREDDS Data Server (TDS) http://www.unidata.ucar.edu/software/thredds/current/tds (Accessed in March 2015)
[19] XMPP Stamdards Foundation, http://xmpp.org (Accessed in March 2015)
[20] Chef, http://www.chef.io (Accessed in March 2015)
[21] Docker, http://www.docker.com (Accessed in March 2015)
[22] Puppet, http://puppetlabs.com (Accessed in March 2015)
[23] Regional Ocean Modeling System (ROMS) www.myroms.org (Accessed in March 2015)
[24] Open Geospatial Consortium OGC® Sensor Observation Service Interface Standard, Version: 2.0. Document OGC 12-006, 2012-04-20. Available at http://www.opengis.net/doc/IS/SOS/2.0 (Accessed in March 2015)
[25] Global Ocean Observing System (GOOS) http://www.ioc-goos.org/ (Accessed in March 2015)
[26] ESSI-Lab Gi-CAT http://essi-lab.eu/do/view/GIcat (Accessed in March 2015)
[27] National Oceanic and Atmospheric Administration (NOAA) http://www.ngdc.noaa.gov/geoportal/catalog/main/home.page (Accessed in March 2015)
[28] SeaDataNet www.seadatanet.org (Accessed in March 2015)
[29] EMODnet www.emodnet.eu/ (Accessed in March 2015)
[30] MyOcean http://www.myocean.eu/ (Accessed in March 2015)
[31] Infrastructure for Spatial Information in the European Community (INSPIRE) http://inspire.ec.europa.eu (Accessed in March 2015)
[32] MyOcean2 and the INSPIRE directive? Current state of play http://www.myocean.eu/automne_modules_files/pmedia/public/r1985_9_myocean2_focuson_article_inspire_directive_1st_august2014.pdf (Accessed in March 2015)

Document Data Sheet

Security Classification		Project No.
Document Serial No. CMRE-PR-2019-121	**Date of Issue** June 2019	**Total Pages** 5 pp.

Author(s)

Alessandro Berni, Giampaolo Cimino, Elena Camossi, Simone Giannecchini

Title

From data to decision: an overview of CMRE's information supply chain

Abstract

Proper data curation is a prerequisite for the creation of long-lasting value from scientific datasets and for the generation of authoritative data-driven decision support tools. Availability of detailed information on instruments calibration and other useful metadata are essential for effective data use and re-use. High quality and standardized metadata, in turn, allows effective data discovery through web-based data catalogues, linking interoperable data repositories and paving the way to federated data infrastructures on a global scale, to ensure that data are made available to scientists, decision makers and automated computing workflows. This paper presents the approach undertaken by the Centre for Maritime Research and Experimentation (CMRE) to provide a "common enterprise bus" in which scientific data and added-value products are the principal currency.

Keywords

Metadata, ISO standards, oceans, interoperability, geospatial analysis

Issuing Organization

NATO Science and Technology Organization
Centre for Maritime Research and Experimentation

Viale San Bartolomeo 400, 19126 La Spezia, Italy

[From N. America:
STO CMRE
Unit 31318, Box 19, APO AE 09613-1318]

Tel: +39 0187 527 361
Fax:+39 0187 527 700

E-mail: library@cmre.nato.int