Grant agreement no: 654310

Ocean Data Interoperability Platform

Deliverable D3.2: Definition of ODIP Prototypes 2

Workpackage	**WP3**	**ODIP Prototypes**
Author (s)	Dick M.A. Schaap	MARIS
Author (s)	Simon Claus	VLIZ
Author (s)	Jonathan Hodge	CSIRO
Author (s)		
Author (s)		
Author (s)		
Authorized by	Helen Glaves	NERC
Reviewer		
Doc Id	ODIP II_D3.2	
Dissemination Level	PUBLIC	
Issue	1.0	
Date	31 August 2017	

Document History				
Version	**Author(s)**	**Status**	**Date**	**Comments**
1.0	Dick M.A. Schaap (MARIS) with contributions by Simon Claus (VLIZ) and Jonathan Hodge (CSIRO)	DRAFT	31 August 2017	First draft

Contents

Executive Summary

In the framework of the ODIP II project a number of ODIP II prototypes are worked out in order to evaluate and test selected potential standards and interoperability solutions for establishing and demonstrating improved interoperability between the regional infrastructures and towards global infrastructures. As a first strand of prototyping for the ODIP II project it was agreed to continue and expand the three prototype projects that had been successfully initiated and implemented in the predecessor ODIP project. The specifications for the expansion have been formulated in the period between the 1st ODIP II Workshop and the 2nd ODIP II Workshop and documented in the ODIP II Deliverable D3.1.

During the 2nd ODIP II Workshop brainstorming started concerning the formulation of a number of additional prototypes, such as focus on i) the concept of collaborative workspaces on the cloud for functions such as computing, analyzing, accessing data, and visualization, and ii) a possible prototype for marine biology. At the 3rd ODIP II Workshop several ongoing virtual research laboratories projects were presented and following the brainstorming a new ODIP II prototype was defined and agreed as:

- ODIP II prototype 4: the Digital Playground.

The marine biology community in ODIP II also continued their brainstorming and recently have agreed on the following new ODIP II prototype:

- ODIP II prototype 5: integration of data management for biological and physicochemical marine data.

This ODIP II Deliverable D3.2 gives an extract of the examples and considerations that were presented and discussed at the Workshops concerning big data, cloud computing and collaborative workspaces as well as concerning biological data management. As follow-up specifications and associated actions are formulated for developing the two additional ODIP II prototype projects which already have been put into motion and will be implemented in the remaining ODIP II project duration.

1. Introduction

The “Extending the Ocean Data Interoperability Platform” project (ODIP II) is promoting the development of a common global framework for marine data management by developing interoperability between existing regional e-infrastructures of Europe, USA and Australia and towards global infrastructures such as GEOSS, IOC-IODE and POGO.

This is done in practice by organising four international workshops over the three years lifetime of the project to present, compare and discuss approaches and standards applied. The workshops involving relevant domain experts provide insights into commonalities and differences and contribute to identify opportunities for the development of common standards and interoperability solutions. As a follow-up ODIP prototypes projects are formulated and worked out in order to evaluate and test selected potential standards and interoperability solutions for establishing and demonstrating improved interoperability between the regional infrastructures and towards global infrastructures. A complication has arisen for ODIP II in comparison to ODIP that only EU partners have achieved extra funding, while contributions from USA and Australian partners must be brought up from their own institute funds. This gives even more emphasis on the approach that actual ODIP II prototype developments should be done largely by leveraging on the activities of current regional projects and initiatives of the ODIP II partners. Therefore ODIP II prototype projects must be formulated taking these constraints into account. This also implicates that additional developments outside of ongoing projects should be done largely by the European ODIP II partners. However luckily the European partners will start from 1st November 2016 with the EU supported SeaDataCloud project as successor to the SeaDataNet II project and this will give extra synergy options.

As a first strand of prototyping for the ODIP II project it was discussed and agreed between partners during the first 2 ODIP II Workshops to expand the three prototype projects that had been successfully initiated and implemented in the predecessor ODIP project. The specifications for the expansions have been documented in the ODIP II Deliverable D3.1.

During the 2nd and the 3rd ODIP II Workshop brainstorming took place concerning the formulation of a number of additional prototypes, in particular concerning i) big data, cloud computing and collaborative workspaces, and ii) marine biological data management. This ODIP II Deliverable D3.2 starts with summarising related sessions. These sessions have provided the basis for the formulation of two additional ODIP II prototype projects, which are specified with associated actions in the final chapter.

2. Brainstorming during ODIP Workshops

Brainstorming for additional ODIP II Prototype projects took place during the 2nd and 3rd ODIP II Workshops. The basis has been provided by the sessions on big data, cloud computing and collaborative workspaces as well as on marine biological data management.

2.1 Big Data and Challenges

Relevant aspects for the definition of Big Data are 'Volume', 'Velocity' and 'Variety' as illustrated in the image below.

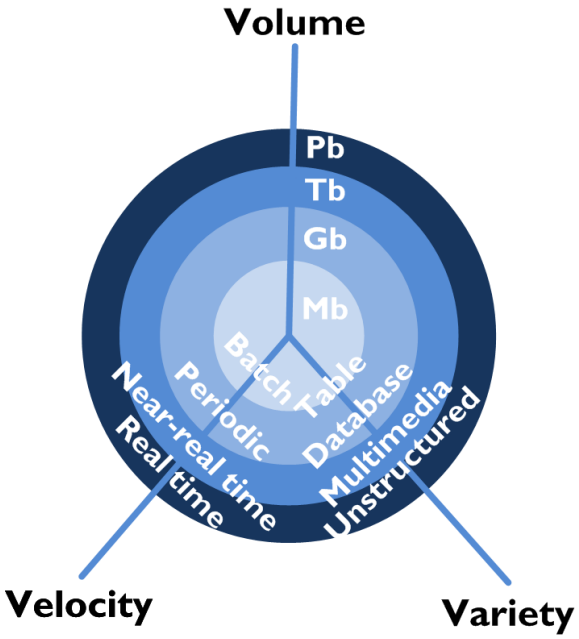

Image: The expansion of data volume, velocity and variety together create the Big Data paradigm.

- **Volume:** Data volumes are growing in all scientific disciplines as new and more connected instrumentation is developed and deployed. In the ocean sciences this growth in data volumes has been characterised by the change in observations from solely ship-board activities to the inclusion of remotely sensed data and the deployment of autonomous vehicles, including Argo floats and glider fleets. Oceanographic models also create large volume output datasets.
- **Velocity:** One of the most revolutionary aspects of connected devices is the speed at which data can be collected, transmitted, processed and integrated. The development of subsea observatories, connected to shore, has driven the push to true real-time data in oceanographic applications. Similarly, increased bandwidth from satellite communications systems means that there are now options to retrieve data in near-real time from research vessels and autonomous vehicles. This latter improvement will grow as low-powered wide area network transmitters become ruggedized for the marine environment. These emerging near- and real-time technologies improve the ability to detect and react to events in the ocean, for example detection of subsea earthquakes feeding tsunami warning systems.

- **Variety:** There are several Big Data applications where the variety of data is not an issue: for example the Twitter dataset contains only (a very large number) of messages with a length no greater than 140 characters. However oceanographic data is much more complex. For example further analysis of a water bottle sample taken at a CTD station on a research cruise might produce tens of measurements for different parameters.

The **Volume** aspect of Big Data presents challenges in working with datasets which are too large to be processed on traditional desktop machines. In order to effectively process the data, workflows must be provided in which the code is moved from a local host to a remote machine so that it is close to the data. **Velocity** challenges the traditional mode of operation of the National Oceanographic Data Centres; namely delayed mode, quality controlled data delivery. Therefore, there is a social-barrier to be overcome in terms of publishing data in real-time. Some of the technical challenges to be overcome include development of quality control algorithms which can act in real-time, identifying and developing algorithms for detecting events of interest in specific oceanographic applications, and ensuring there is minimal data transfer lag in the systems used. The **Variety** challenge can be approached by making use of Controlled Vocabularies for the marine domain. These allow data descriptions to be understood globally, both by humans and computers. The comprehension of dataset descriptions by computers is key for the development of information products in automated workflows.

Examples of Big data projects

2.1.1 NOAA's Big Data Project (USA)

NOAA's Big Data Project has an innovative approach to publishing NOAA's vast environmental data resources by positioning them near cost-efficient high-performance computing, analytic, and storage services provided by the private sector. This way a sustainable, market-driven ecosystem is created that lowers the cost barrier to data publication. The first task of the Big Data Project was the re-creation of the National Center for Environmental Information archive of Level II RADAR data from 1991 to present and on-going. The NEXRAD data are now freely available through cloud infrastructure. It has centralised storage, and makes use of standards (OGC services, NetCDF, THREDDS) to provide interoperable services. Using the cloud allows scalable computing resources. Moreover it allows to build and support workflows, using various applications added by machine imaging and containerisation.

2.1.2 Galway Bay subsea cabled observatory (Europe)

The Marine Institute (Ireland) has been developing a streaming data system for its subsea cabled observatory, deployed in Galway Bay in the summer of 2015. The streaming data system deployed uses open source technologies developed for large-scale web applications with a real-time data component. Key amongst these technologies is the Apache Kafka message queue, which was open sourced by LinkedIn, allowing for buffered storage of data at various points in the data flow and for chunks of the data stream to be easily reprocessed. This system is also taking advantage of emerging Internet of Things standards, namely MQTT, to provide Web Sockets interfaces to the data allowing real-time pushing of data to update graphical displays. This includes streaming data services with SOS-OM-JSON output that are used as part of model workflows.

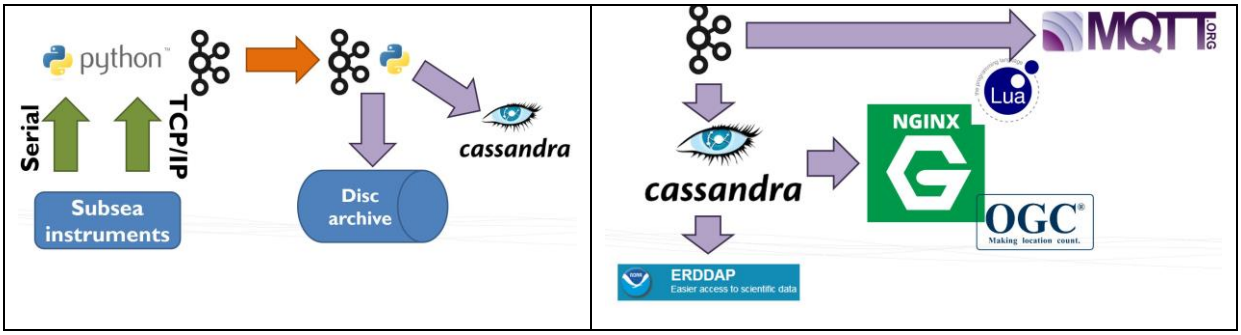

Image: Architecture of the Galway Bay subsea cabled observatory (part 1: collection; part 2: access and publishing)

2.1.3 Data Quality Strategy at the National Computational Infrastructure (Australia)

NCI manages 10+ PB data collections from climate, coasts, oceans and geophysics through to astronomy, bioinformatics and the social sciences. It wants to enable transdisciplinary access to its data holdings. Therefore it has implemented a Data Quality Strategy (DQS) to simplify access to data. Key elements of the DQS are to maximize benefit of NCI's collections and computational capabilities and to ensure seamless interoperable access to these datasets. The goal is to combine data and to visualize them. Collections are being accessed and utilised from a broad range of options: direct access on filesystem, Web and data services, data portals, and virtual labs. The data concern many disciplines and within these disciplines span a wide range of features: gridded, non-gridded (i.e., trajectories/profiles, point data), coordinate reference projections, resolutions. A major challenge was application of community-agreed data standards to the broad set of Earth systems and environmental data that are being used.

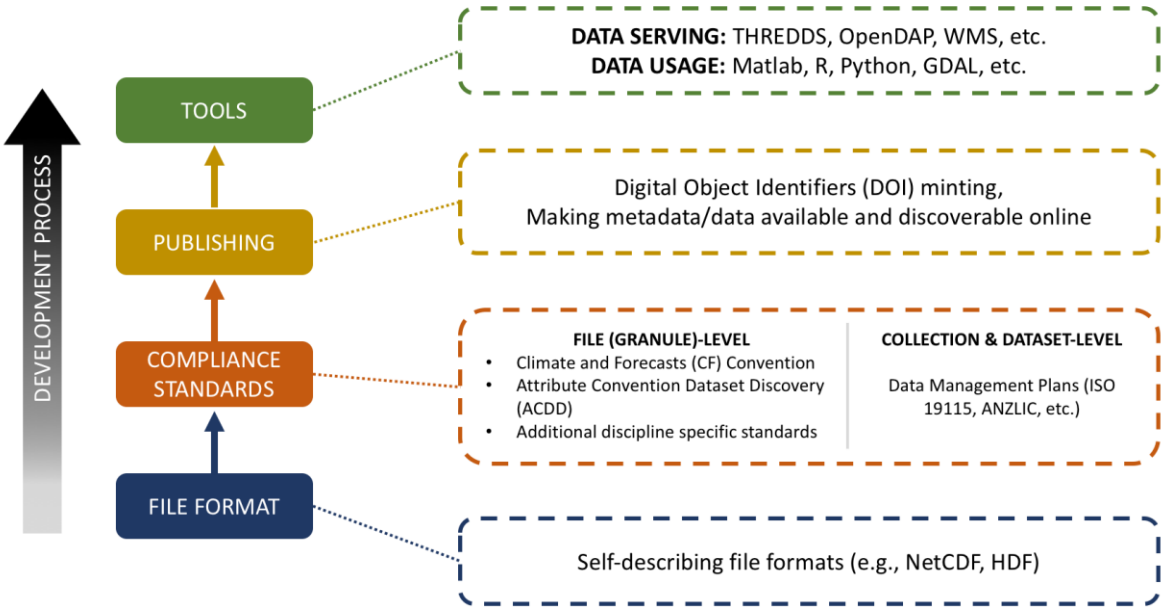

Image: Data Quality Strategy (DQS) at NCI

The DQS provides processes for: 1) underlying High Performance Data (HPD) file format, 2) close collaboration with data custodians and managers (planning, designing, and assessing the data collections), 3) quality control through compliance with recognised community standards, 4) Data assurance through demonstrated functionality across common platforms, tools, and services. The following image gives the architecture that was established at NCI for the National Environmental Research Data Interoperability Platform (NERDIP).

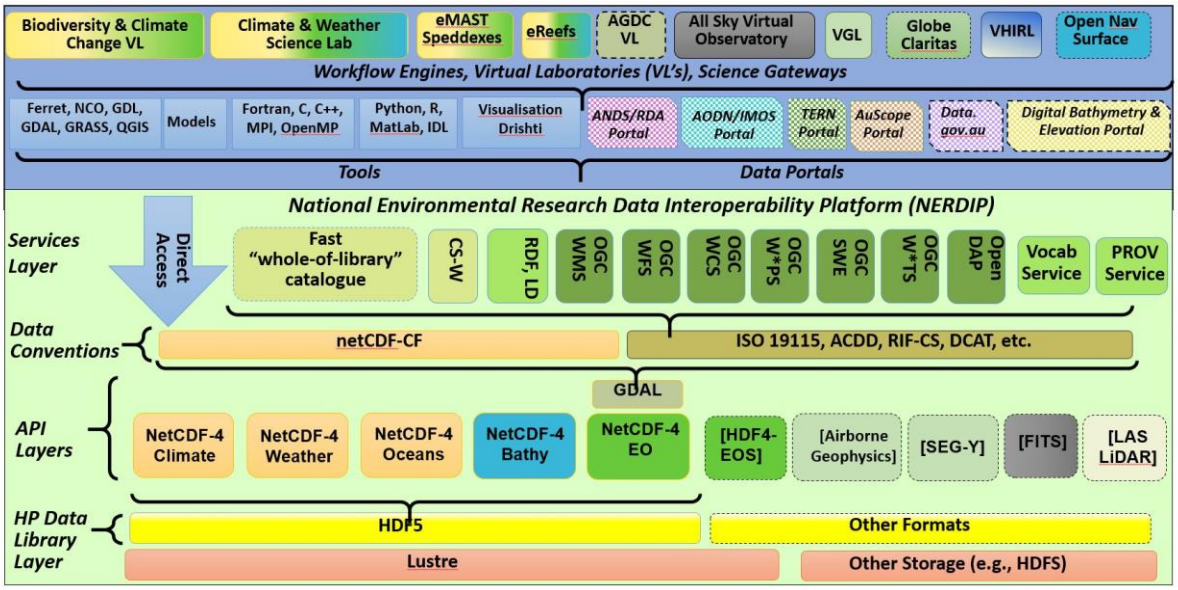

Image: Architecture of the National Environmental Research Data Interoperability Platform (NERDIP)

2.2 Examples of virtual research environment projects

Virtual Research Environments (VREs) provide the IT infrastructure to enable researchers to collaborate, share, analyse and visualise data over the internet.

2.2.1 The EVER-EST project (Europe)

The EVER-EST Horizon 2020 project is developing a virtual research environment (VRE) focussed on the requirements of the Earth Science community. Within the earth sciences there are major challenges such as climate change research and ensuring the secure and sustainable availability of natural resources and understanding natural hazards which require inter-disciplinary working and sharing of large amounts of data across diverse geographic locations and science disciplines to work towards a solution. The EVER-EST virtual research environment is building on a number of e-infrastructures which have been created under European Commission funding in recent years. Other work packages are validating this emerging infrastructure using appropriate use cases, which are deployed by four virtual research communities (VRCs) for i) Natural hazards, ii) Supersites, iii) Land monitoring, and iv) Sea monitoring. The EVER-EST e-infrastructure will provide users with a suite of tools and services, that will enable them to:

- Discover, access, assess and process both existing and new heterogeneous Earth Science datasets including the associated information and preserved knowledge held by distributed data centres

- Share data, models, algorithms, scientific results (including traceability of workflows and processes that would facilitate reproducibility of modelling and simulations) and their own experiences within a community or across communities (including those in other domains beyond Earth Science)
- Capture, annotate and store the workflows, processes and results from their research activities
- Work together in a real-time environment that facilitates the sharing of expertise, information and data resources overcoming the limitations of traditional working practices e.g. need for physical meetings or the transfer of large datasets between users
- Ensure the long-term sustainability and preservation of data, models, workflows, tools and services developed by existing communities of practice that can potentially be re-used in the future by other users either for validation of existing research or for new applications.

The EVER-EST VRE Main Gateway will provide a user-friendly interface to access the research e-infrastructure. This Gateway will be a public point of access allowing users to understand the services and functionalities available in the EVER-EST VRE. It will also provide each EVER-EST VRC with its own specific and customised user interface. The EVER-EST Virtual Research Environment will be implemented as a Service Oriented Architecture (SOA) based on loosely coupled services tailored to the requirements of Virtual Research Communities involved. These services can be differentiated as those that are generic and those that are specific to the requirements of the Earth Science domain.

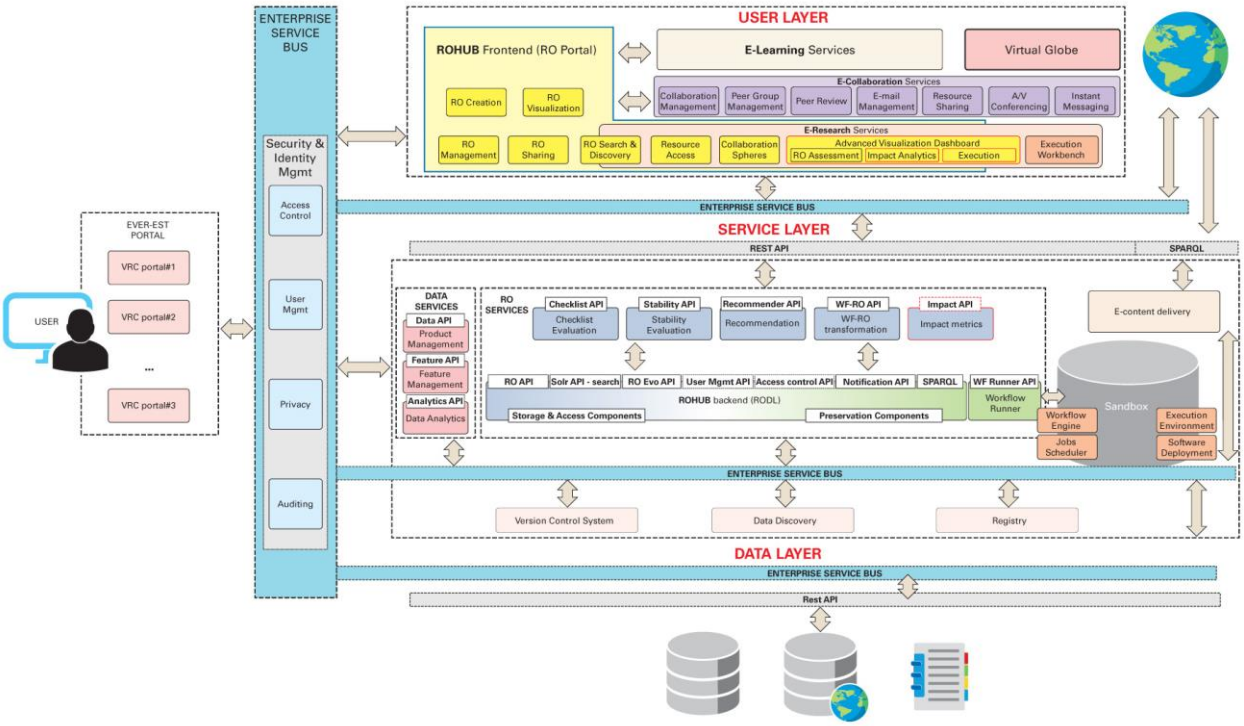

Image: Architecture of the EVER-EST Virtual Research Environment

2.2.2 SeaDataCloud – VRE development (Europe)

The SeaDataCloud Horizon 2020 project is successor to the SeaDataNet II project and it concerns upgrading and further developing the standards and services of the pan-European SeaDataNet infrastructure for marine and ocean data management. This includes upgrading and expanding the architecture of the SeaDataNet infrastructure, inter alia by adopting cloud and High Performance Computing technology. For the latter the SeaDataNet members have entered into a strategic and technical cooperation with the EUDAT consortium. EUDAT is a European network of computing infrastructures that develop and operate a common framework for managing scientific data and providing an interoperable layer of common data services. SeaDataNet will cooperate with the EUDAT e-infrastructure service providers to build upon the state of the art in ICT and e-infrastructures for data, computing and networking.

One of the activities concerns incorporating a Virtual Research Environment (VRE) to facilitate collaborative and individual research from public, academic and private institutes concerning using, handling, analysing and processing ocean and marine data into value-added data products, which can be integrated, visualised and published using OGC and high level visualisation services. Thus the cloud environment will host a number of advanced services, seen as a packaged collection of processing services and that can be connected to subsets of the SeaDataNet data resources. A variety of advanced services will be offered by the Ocean Data View (ODV) and the Data-Interpolating Variational Analysis (DIVA) software for which online versions will be developed. The ODV service will include functionality for validating and harmonising large collections of data sets for specific data types, which will contribute to preparing data products such as temperature and salinity climatologies. The VRE will be set up in such a way that additional advanced services can be included without too much effort. The VRE development is still in an early stage of functional analysis.

2.2.3 Climate Information Portal for Copernicus (CLIPC) (Europe)

The CLIPC FP7 project has developed the CLIPC portal which provides access to Europe-wide climate and climate impact data, from scientifically trusted sources, along with the supporting information required for its effective and meaningful use. This “one-stop-shop” portal facilitates users in their search to answer questions related to climate change impact. It has been developed to accommodate the needs and demands of diverse users across Europe to the largest extent possible. The CLIPC portal has important advantages over other European climate information portals:

- Access to different data types: The portal includes data from satellite and in-situ observations, climate models and data re-analyses, transformed data products enabling impact assessments, climate change impact indicators, and socio-economic data that are important to assess vulnerabilities;
- Access to a large variety of data sources**:** The climate data search service allows users to search for climate datasets in several important international infrastructures. The portal complements existing services, but focuses on datasets providing information on climate variability on decadal to centennial time scales from observed and projected climate change impacts in Europe;
- Data provenance**:** CLIPC ensures that the provenance of data products is well-documented, by providing access to intermediate data products and documentation on the technical quality of data, on metrics related to scientific quality, and on uncertainties in and limitations of the data;
- Enhanced functionality**:** Users can store the results of their searches in their own environment (MyCLIPC) and combine the information with other data files for their specific purposes. Various postprocessing options are available; for example, the

novel Impact Indicator Toolbox allows users to combine, compare and rank indicators and generate new ones.

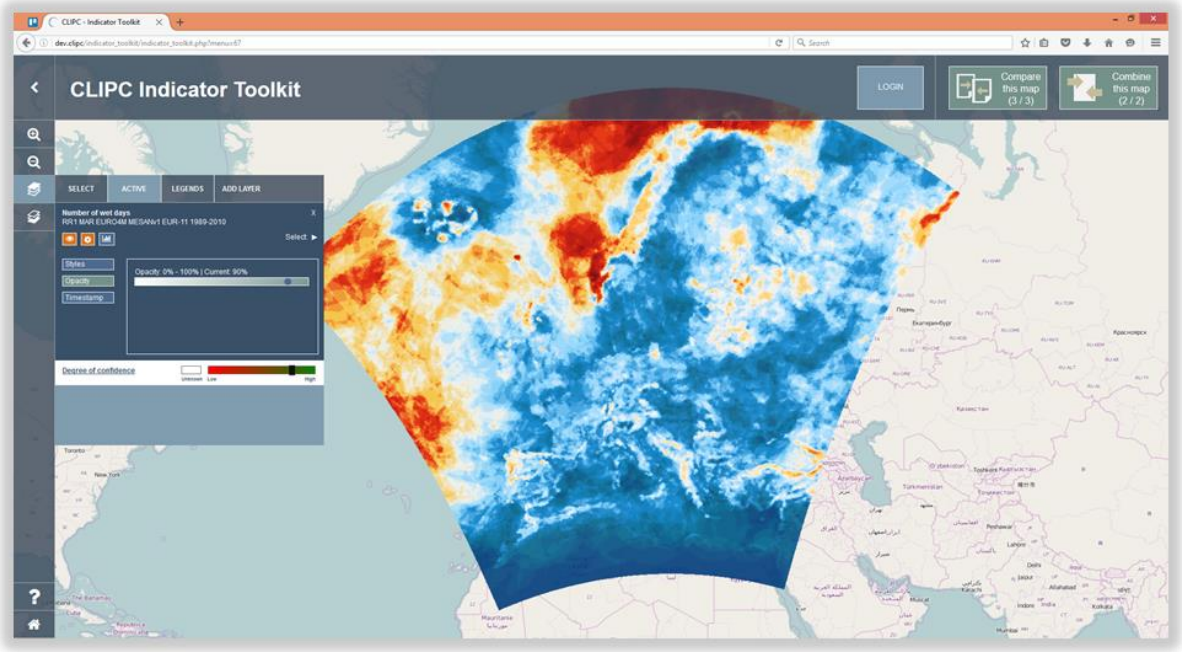

Image: CLIPC Impact Indicator Toolbox

The CLIPC toolbox enables specialist and non-technical users to assess possible impacts of climate change in an effective and trustworthy way, by combining climate and climate impact indicators. The user can view and explore impact indicators calculated for different climate change and socio-economic scenarios. Available datasets for the indicators can be selected and combined with each other using built-in operators and normalization functions. In addition, the toolkit allows users to perform decadal averaging “on the fly” to time series of indicators, and spatial averaging of these results across the regions of Europe. Toolkit results can be saved and retrieved from a personal data-basket. Users can also compare the selected datasets via a map view, compare the supporting metadata, or “combine” two datasets into a new dataset. With this functionality users can add up climate impacts, or create a difference map. This opens up many new possibilities for climate change impact and vulnerability analysis.

2.2.4 Nectar Research Cloud, MARVL, and Australian Marine Sciences Cloud (Australia)

The Nectar Research Cloud is a national computing service with 30.000 virtual machines connected by AARNET. About 10000 users, mostly from universities, are connected and use the Nectar Cloud. It provides computing infrastructure, software and services that allow Australia’s research community to access and share computational models, tools, data, and collaboration environments. Nectar Cloud allows researchers access at any time from any location and to easily collaborate nationally and internationally. Nectar is providing a platform on which already several VREs have been developed. The Nectar Virtual Laboratories (https://nectar.org.au/labs/) are rich domain-oriented online environments that draw together research data, models, analysis tools and workflows to support collaborative research across institutional and discipline boundaries. The Virtual Laboratories are built by and for the Australian research community.

Marine Virtual Laboratory (MARVL):
One of the Nectar Virtual Labs is the Marine Virtual Laboratory (MARVL). MARVL focuses on setting up a coastal ocean model. It allows a non-specialist user to configure and run a model, automating many of the modelling preparation steps needed to bring the researcher faster to the stage of simulation and analysis. MARVL allows researchers to:

- Efficiently configure a range of different community ocean and wave models for any region around Australia, for any historical time period, with model specifications of their choice, through a user-friendly web application;
- Access data sets to force a model and nest a model;
- Discover and assemble ocean observations from the Australian Ocean Data Network (AODN);
- Run the assembled configuration in a cloud computing environment, or download the assembled configuration to run on any other system of the user's choice.

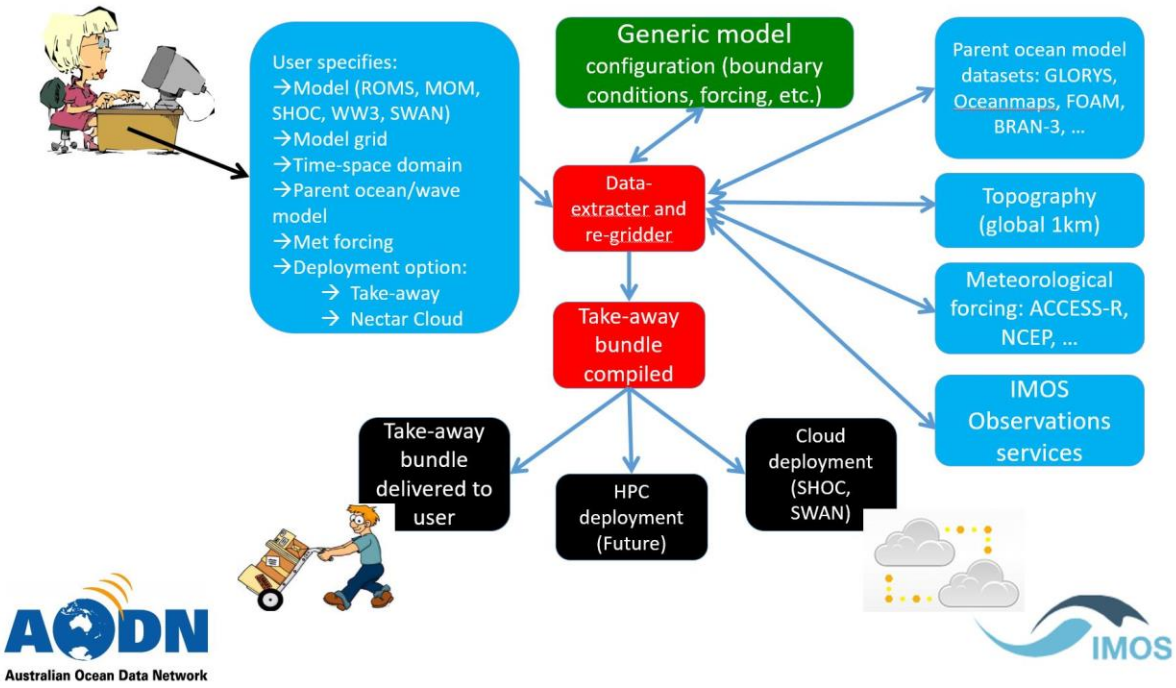

Image: set-up of the Marine Virtual Laboratory (MARVL)

Science Clouds:
In partnership with National Collaborative Research Infrastructure Strategy (NCRIS) domain capabilities, also three science communities are being established: the Biosciences cloud, the Ecosystems Sciences cloud, and the Marine Sciences cloud. These community focused cloud platforms will extend, upgrade and integrate existing NeCTAR, RDS, BPA, TERN and IMOS infrastructure. The Science Clouds will ensure that research communities have improved access to shared data, tools, platforms and computing resources according to each community's domain-specific needs.

One of the components of the Marine Sciences cloud is virtual desktop support for marine and climate scientists. This consists of:

- CoESRA – Collaborative Environment for Ecosystem Science Research and Analysis (developed by NCRIS capability TERN – Terrestrial Ecosystems Research

Network). This provides a desktop environment (a VRE) where the user can interact with the applications and also load his own data. The desktop environment can sit close to big data sets and uses them without having to download them;

- Enabling 'on the fly' creation of a scalable VM containing suite of tools including
  - Jupyter notebook, QGIS Browser, Rstudio, Panoply, Kepler Scientific workflow
- Enabling links to personal and public data repositories

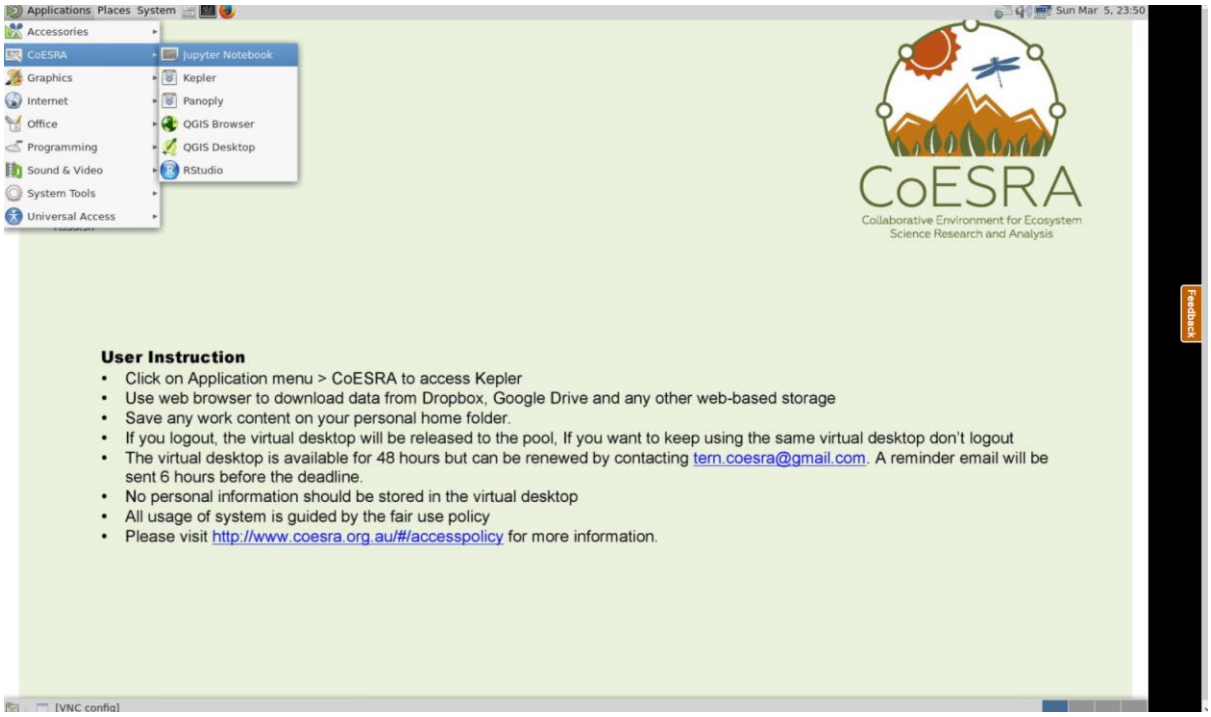

Image: CoESRA desktop interface as part of Australian Marine Sciences cloud

This way the Australian Marine Sciences Cloud is creating for the marine research community a rich cloud resource to maximise usage and skills development relating to:

- Existing data sets, their location and utilisation;
- Data management - access, storage, interoperability, and reuse;
- New data and knowledge such as climate change and biodiversity information;
- Online modelling and analysis tools

The Virtual Desktop provides an easily-accessible environment that comes pre-installed with all the necessary tools (such as rstudio and ipython) to significantly reduce the timeframe for researcher access. Before Nectar's Marine Sciences Cloud Virtual Desktop was established, experienced researchers using a handful of tools took around half a day to be in a position to perform analysis. This timeframe potentially extended to weeks for less experienced people to identify the required software and configure the environment to process data.

Another component of the Marine Sciences cloud is a national service for annotating and analysing underwater imagery. The annotation and analysis service includes SQUIDLE+ for analysis and classification, and GlobalArchive that provides a platform for sharing, cataloging and exploring archived annotation data.

The Marine Sciences cloud development has a long term benefit to MARVL, the Marine Virtual Laboratory, in that the analytics tools developed will provide a natural step up for the analysis of MARVL simulations.

2.2.5 eReefs (Australia)

eReefs is an Australian collaborative project to develop an information system for monitoring the Great Barrier Reef and predicting future changes. It provides nested modelling of hydrodynamic (3D water flows, Temperature, Salinity), geochemical (nutrients, Chl), ecosystem, fisheries models with other local ones along the coast for better understanding of the ecosystem. CSIRO is one of the developing partners. eReefs commenced in January 2012 and is a six year $30 million collaborative project that combines government commitment to Reef protection, world-class science innovation and contributions from leading Australian businesses. Using the latest technologies to collate data, and new and integrated modelling, eReefs produces powerful visualisation, communication and reporting tools. This information will benefit government agencies, Reef managers, policy makers, researchers, industry and local communities.

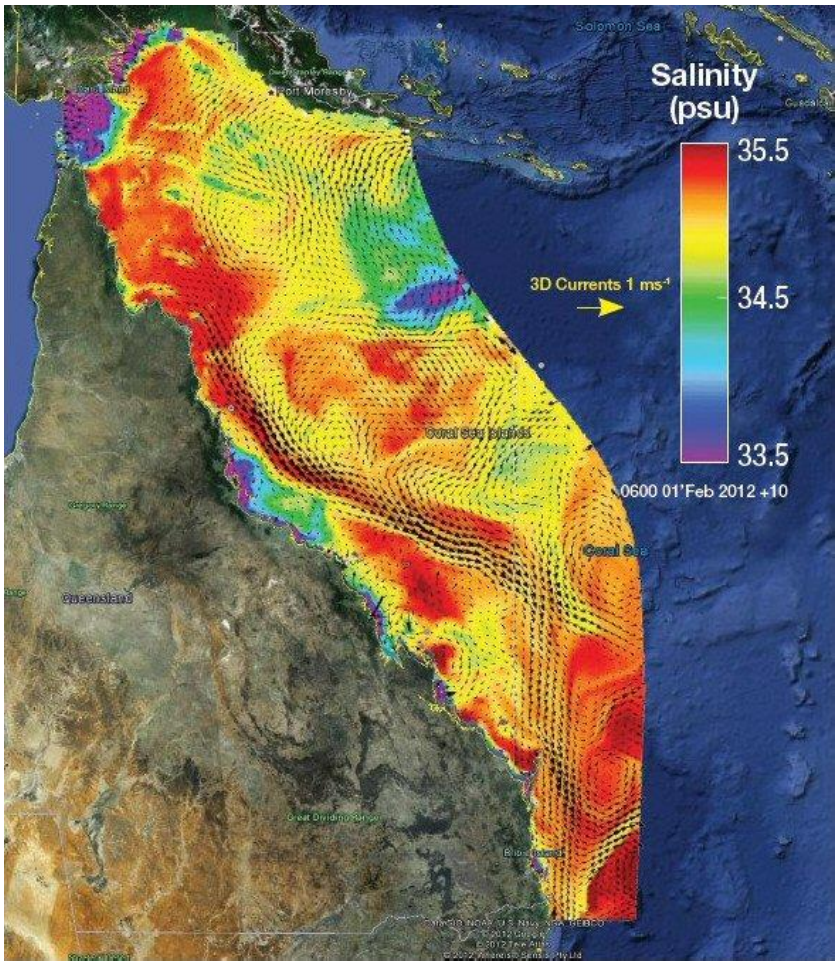

Image: Snapshot from near real-time hydrodynamic model of the Great Barrier Reef showing sea-surface salinity and surface currents.

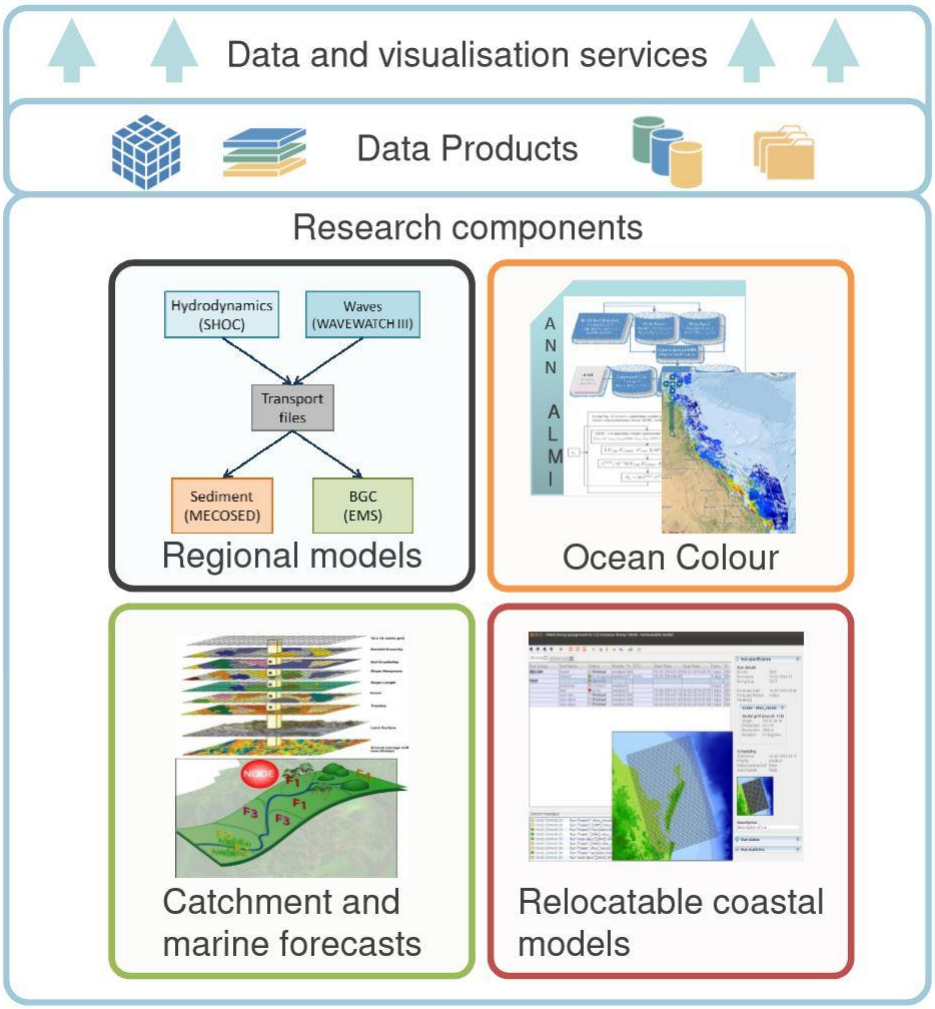

Image: Set-up of the eReefs system providing easy access to users

2.2.6 Australian Urban Research Infrastructure Network (AURIN) (Australia)

AURIN is a national collaboration delivering e-research infrastructure to empower better decisions for Australia's human settlements and their future development. AURIN is a powerful example of a VRE where different kinds of urban data from official sites can be added, processed and analyzed. AURIN commenced in 2010 and was a three year $20 million collaborative project with funding provided by the Australian Government under the National Collaborative Research Infrastructure Strategy (NCRIS) and associated programmes. The AURIN network nowadays comprises a diverse community of data providers, sub-project collaborators, technology partners and international connections, all brought together by a common vision of better information infrastructure driving Australia's urban future.

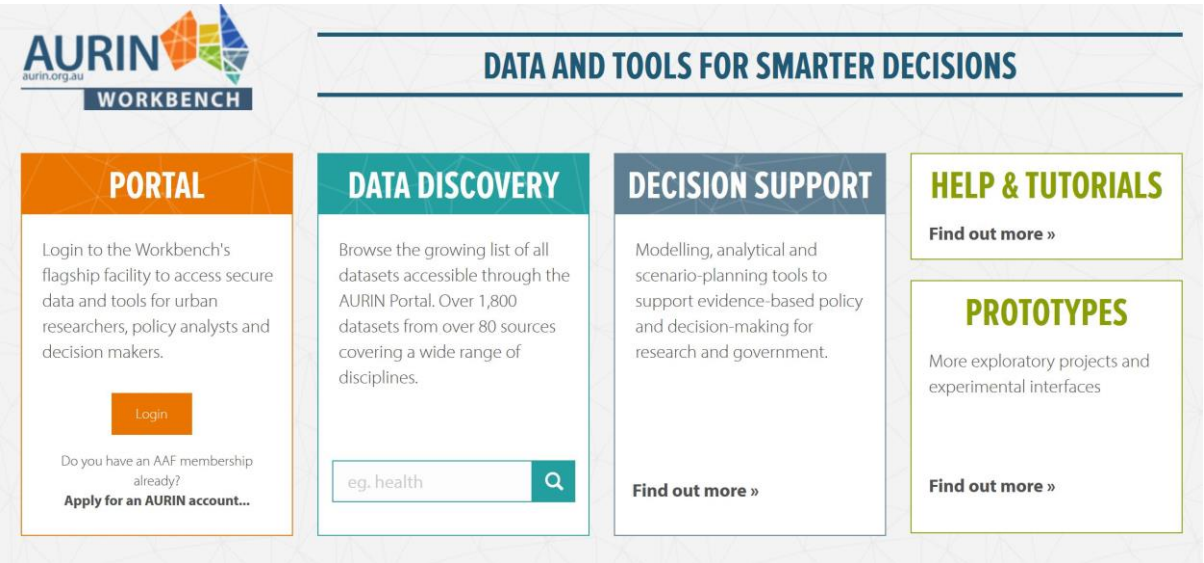

Image: Set-up of the AURIN Workbench

2.3 Marine biological data management

During the 2nd and 3rd ODIP II Workshops a number of topics were discussed concerning marine biological data management. This paragraph summarises a selection of these topics that might provide a basis for a dedicated ODIP II Prototype project for the marine biological data management domain.

2.3.1 OBIS-ENV-DATA: a global data sharing facility for sample and sensor-based data holding species occurrence and environmental measurements (global)

OBIS stands for Ocean Biogeographic Information System and is the largest database on species distribution in the world. Its content is increasing with circa 1000000 entries per year. OBIS is offering open-access to data to enhance capacity, research and international collaboration. At June 2009 OBIS became part of IOC-IODE and supports several international processes. The OBIS network consists of 600 institutions. Collectively, they have provided over 45 million observations of nearly 120 000 marine species. The data flow structure is based on three tiers of nodes. At tier I is the aggregate global database and is managed by the project office in Ostend (Belgium). Tier II nodes (like OBIS Australia and EurOBIS) are responsible for many of the quality control and other data management tasks. Tier II nodes are the key drivers who push data to OBIS data base. Tier III nodes are willing participants in adding data to the OBIS network, but may not have the expertise or resource base to meet all of the responsibilities of a tier II node. The addition of tier III nodes provides two added benefits to the network. First, expanded capacity for reaching out to the science community and second an opportunity for larger tier II nodes to mentor smaller or new member nodes. Only Tier II and global thematic taxonomic Nodes would feed the global dataset directly.

OBIS is connected to the GEOSS portal through the GEO DAB brokerage service which is also adopted in ODIP Prototype 1 and its successor ODIP II Prototype 1+ as mediator for the metadata exchange between SeaDataNet, AODN and US-NODC portals towards GEOSS. EurOBIS is the European component of OBIS and is managed by VLIZ. EurOBIS is also the core data management infrastructure behind the EMODnet Biology project and portal.

The OBIS data standard is Darwin Core. In practice biologists sample more data than just species occurrence. In many cases they also sample the environmental conditions at the sample site, such depth, temperature, nutrients and other parameters. And a lot of these marine environmental data are not reported to oceanographic data centres. Therefore at the XXIII session of the IOC Committee for IODE, March 2015, it was decided to undertake a pilot for expanding OBIS with environmental data that are collected by the biologists at the biological sampling sites. As part of the OBIS-ENV-DATA pilot project activities took place for extending the Darwin Core data schema, while the SEaDataNet Controlled Vocabularies have been adopted for describing the marine environmental parameters. Part of the extension is an Events component. An Event hierarchy makes it possible to record differences in sampling time, location, and depth while still grouping these samples together to the same biological station visit. A number of use cases have been worked out to consider how the extended Darwin Core scheme can be used in practice.

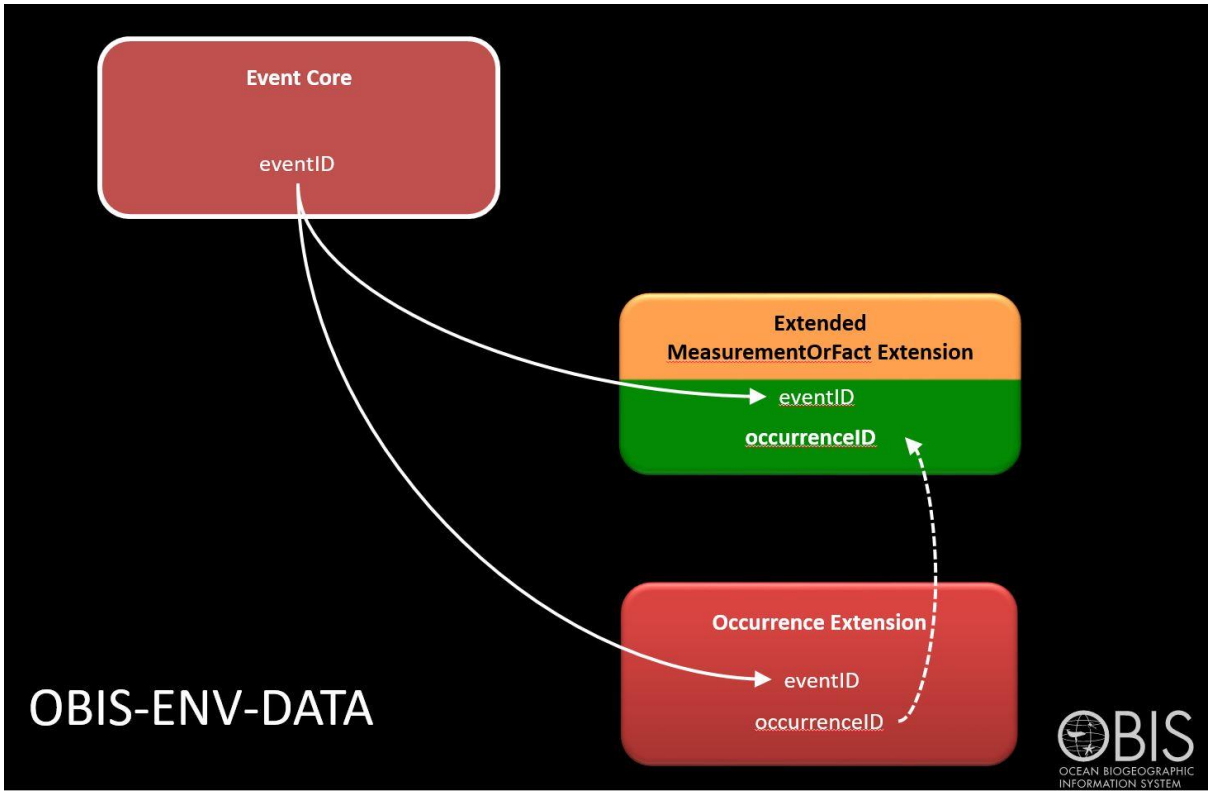

Image: Extended Darwin Core scheme

OBIS uses GeoServer to expose a number of tables or views as WMS or WFS services. Current work is on R packages to read the OBIS data structures.

2.3.2 WoRMS: the global authoritative list of names of all marine species (Europe)

Authoritative information on marine species needs to be easily available to allow for the rapid interpretation of the results of environmental surveys. The World Register of Marine Species (WoRMS) is a Controlled Vocabulary and it provides an authoritative and comprehensive list of names of marine organisms, including information on synonymy. While highest priority goes to valid names, other names in use are included so that this register can serve as a guide to interpret taxonomic literature. WoRMS includes much more information than only taxon names and their relationship such as unique and stable identifier for each taxon name, environment, distribution, specimen information, vernaculars, traits, etc. WoRMS is part of

the Aphia platform which is hosted at VLIZ. The Aphia platform is an infrastructure designed to capture taxonomic and related data and information, and includes an online editing environment. Aphia includes connections to more than 80 related global, regional and thematic species databases. It also allows the storage of non-marine data. The Ascidiacea World Database is one of the global species databases. WoRMS is working with almost 400 editors (both taxonomic and thematic), worldwide can use an online editing environment. There are on average 767 taxonomic edit actions per day, including bulk edits.

2.3.3 The creation of the e-infrastructure Lifewatch, supporting marine biological research (Europe)

Lifewatch is a European infrastructure project, included in the EU ESFRI roadmap, for which the preparatory phase was started in 2008. It is now at the construction phase. LifeWatch is a distributed Virtual Research Environment (VRE) for biodiversity, climatology & environmental impact studies. It consists of components for access, analysis and development. The Lifewatch Taxonomic Backbone consists of 5 major components of databases and data systems for: genomic data, taxonomic information, biogeographic data, trait data, and literature. For workflows use is made of: taverna, Python, R, and others.

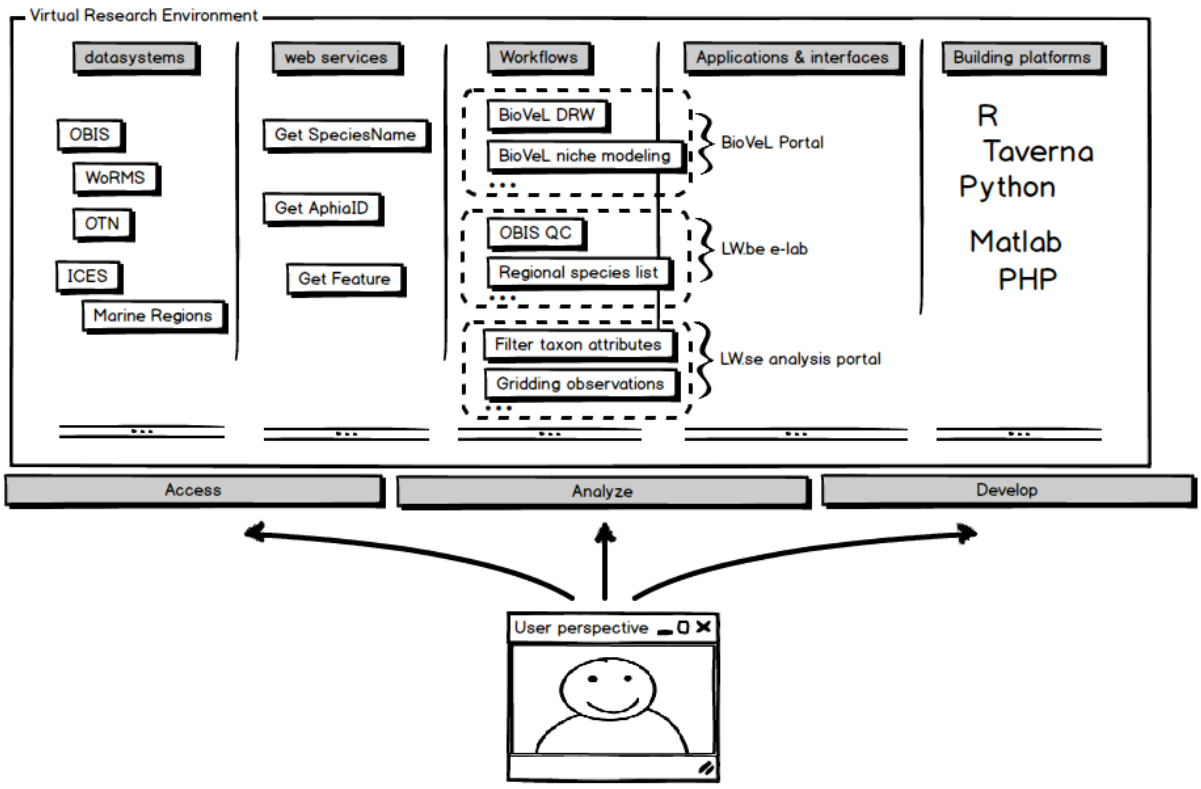

Image: Architecture of the LifeWatch VRE

In the framework of LifeWatch Belgium several web services are being developed to standardize, analyse and visualise user data, and to extract additional data from several internal and external data systems. The E-lab data service web interface was developed where users can select several web services at once in an easy, user friendly way.

The following image gives an example of the E-lab data service interface.

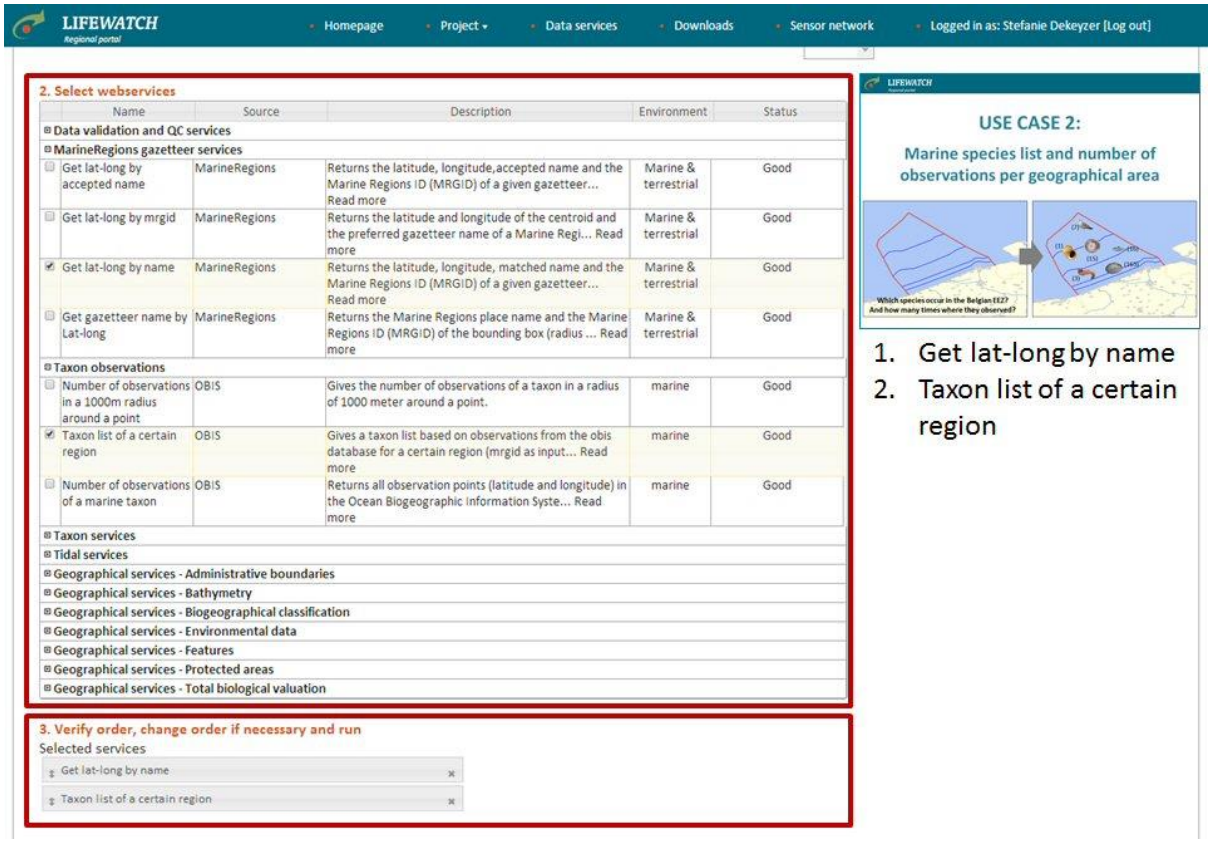

Image: User Interface of E-lab data service

2.3.4 The Global Ecological Marine Units (EMU) Project (USA)

The EMU project is one of the three ecosystem classification mapping initiatives commissioned by the Group on Earth Observations (GEO) in 2014 with the aim to develop a standardized, robust and practical global ecosystems classification and map for the planet's terrestrial, freshwater, and marine ecosystems, completely in 3D. The project is also related with the GEO Biodiversity Observation Network (GEO BON) and the GEO Ecosystems Initiative (GEO ECO). The basic idea is to map on global scale the physical parameters that served to structure the ecology. The EMUs are developed in a three step process: 1) create an empty, volumetric column-based mesh as a global, spatial reference standard and analytical framework, 2) populate the spatial framework with relevant marine physical environment data including water column variables and seabed topographic features, and 3) cluster the abiotic data into ecologically meaningful, 3D regions represented as volumetric polygons. The EMUs are subsequently analyzed against species distribution data to assess strength of relationship between distinct abiotic environments and species biogeography. The data framework will provide new opportunities for correlating physical, chemical, biological, and ecological variables in a 3D environment. The work is undertaken in a government / NGO / academic / private sector partnership which includes a large group of international marine experts in an advisory capacity. The global ecosystems data are intended to be useful in a variety of applications including climate change impacts assessments, ecosystem goods and services valuation assessments, conservation planning, resource management, and scientific research. The physical and chemical data are derived from NOAA's World Ocean Atlas (WOA). Current work includes efforts to integrate the OBIS data. The EMU 3D Point Mesh Framework has 52 million points. Each cluster is being

attributed with ecological and biological data. Each point is an average of an average of the prominent mean over 50 years. The temporal signal (monthly, seasonal) is not handled yet. There is not yet much progress in the integration of OBIS data because of the difference of spatial resolution of OBIS data compared with the WOA. A Vertical Profile App (livingatlas.arcgis.com/emu) allows exploration of each of the EMUs. Vertical diagrams of EMUs clusters at depth illustrates that there is a zonation but not a simple clear-cut boundary for water attributes. Another major point is that nutrients and oxygen distributions not only shape but are shaped by biological processes.

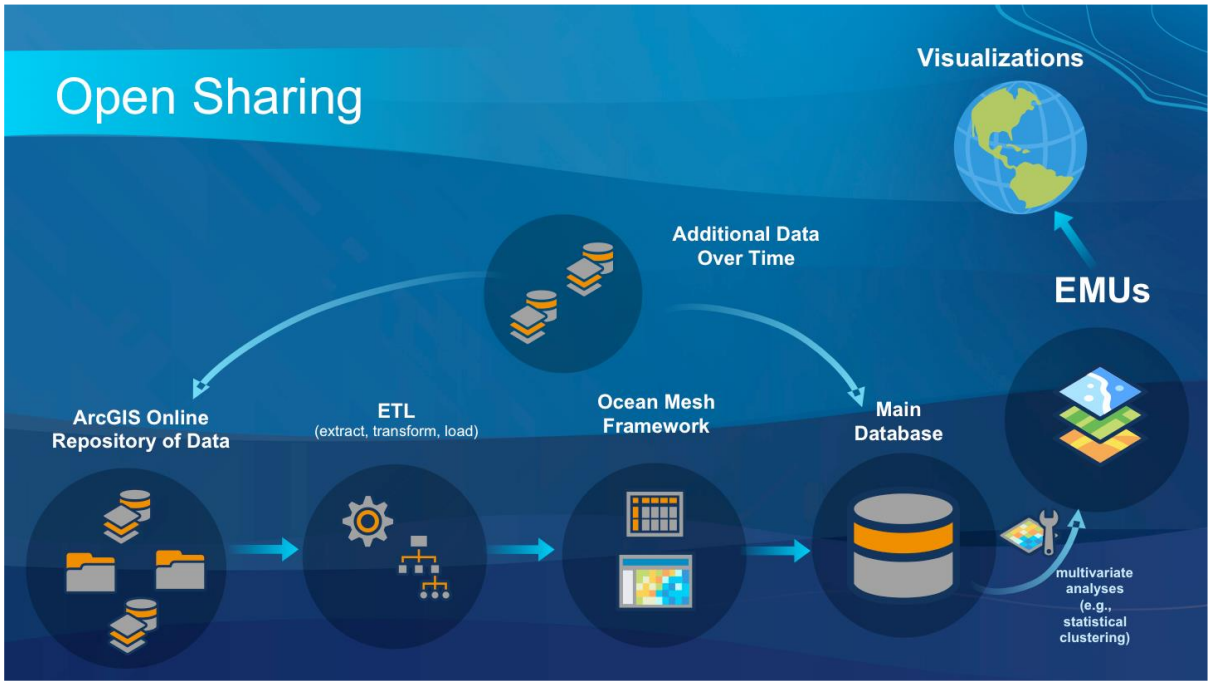

Image: EMU workflow

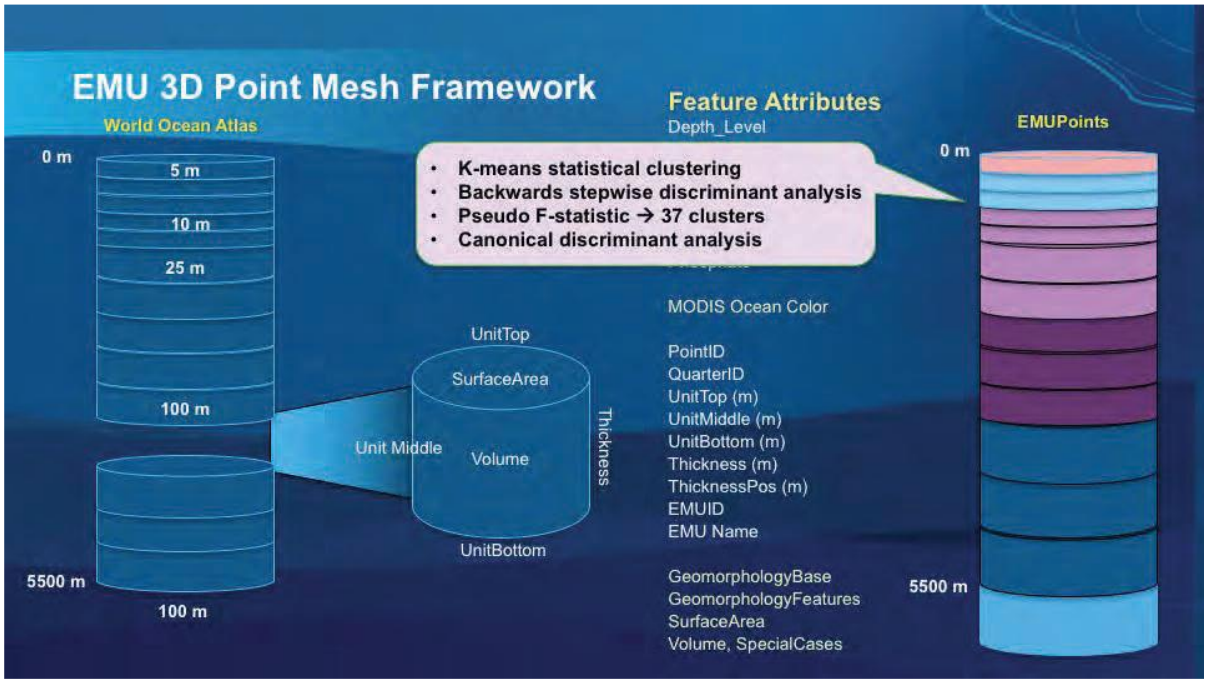

Image: EMU 3D Point Mesh Framework

3. Formulation of additional ODIP II Prototype 4 project: The Digital Playground

From the ODIP II sessions and the presented big data and virtual research environment projects / systems it can be observed that these all revolve around 'integration' of marine data: i) bringing together data from real-time and near real-time data services and from large archives and repositories; ii) undertaking analyses and various forms of processing, arranged in workflows; iii) publishing and visualizing data and data products; iv) reproducibility of calculations and analyses. Thereby we have to deal and find solutions for the increasing volume, velocity and variety of the data sets. To overcome barriers of increasing volume, programming codes and data sets can be moved to remote machines such as provided by cloud infrastructure. This poses technical and organizational challenges how to set-up and make use of such configurations. Increasing velocity gives technical challenges such as developing real-time quality control algorithms, overcoming performance issues because of large throughputs, and guaranteeing continuity of data streams in case of hick-ups in data streams from the field towards the receivers. Increasing variety poses interoperability challenges which can be tackled by applying standards for metadata and data formats, supported by controlled vocabularies, both for machine-to-machine services and human interfaces. This way automated workflows can be set-up, combining various analysis and process services, and multiple types of data and data products. Controlled Vocabularies is already one of the cross-cutting activities within ODIP II and its progress in the marine domain is regularly discussed.

Integration of marine data is a problem for which many projects are in need of solutions. This applies to projects with focus mostly on archived data such as the EMODnet projects within Europe as well as to various Ocean Observing Systems which form part of the Group on earth Observations System of Systems.

This has led to the formulation of the ODIP II Prototype 4 project: **the Digital Playground**.

Lead:

MARIS (Europe) + CSIRO (Australia)

Aims:
To explore, review, and formulate common solutions and best practices for setting up and configuring virtual research environments in the marine domain, dealing with a great variety of data types, processes, user classes, and both operational and delayed mode data services.

Approach:

The overall approach consists of reviewing relevant existing projects and initiatives through presentations and discussions at the ODIP II Workshops, followed by further analyses, looking for common developments as well as specific promising solutions. The analyses will focus and leverage on on-going projects of ODIP II partners, but will also include a number of 'external' projects, even outside the marine domain. Relevant topics are among others: service oriented architectures, building and managing workflows, exchange format standards between workflow components, interactive visualisation and analysis tools, collaborative tools, access to operational and delayed mode data systems, provenance and reproducibility of workflow analyses, and solutions for optimising performances. It is recognised that interactive visualisations are key to the success of virtual research environments or virtual laboratories. These visualisations should include search, composition and analysis tools to

allow users to truly interact with the data as we are moving from a paradigm in which users expect to download data before they can explore it to one in which they can use online tools to achieve these goals. Where possible, the ODIP II desk study review will be complemented by trials and demonstrations, leveraging on one or more on-going projects of ODIP II partners, such as SeaDataCloud (Europe) and/or the Australian Marine Sciences Cloud (Australia).

Activities:

- Presenting and analyzing existing virtual research environment projects and initiatives
- Extracting, drafting and discussing common standards and best practices for configuring virtual research environments, focusing on among others:
    - Service oriented architectures
    - discovery and retrieval of data from data archives and Sensor Web systems
    - processing and product generation using workflow management environment (e.g. Kepler or Taverna),
    - use of OGC standards and services (WMS, WFS, WCS, WPS) for chaining
    - interactive visualisations of data and data products
    - dedicated process engines for specific applications (e.g. DIVA, ODV)
    - publishing (including provenance metadata and DOIs) of created data products
    - data exchange formats including use of controlled vocabularies
    - interfaces for non-expert and expert users (i-notebooks)
    - possible trials and demonstrations for a use case such as Temperature & Salinity Climatology; the focus will be on methodology, not on completeness of data
- Reporting results of the ODIP II Prototype 4 as common standards and best practices of relevance for building virtual research environments and possible dissemination to the IODE Ocean Data Standards and Best Practices programme (ODSBP)

4. Formulation of additional ODIP II Prototype 5 project: Integration of data management for biological and physicochemical marine data

Within ODIP, prototypes are developed for the purposes of testing and evaluating potential solutions for solving the marine data management issues identified within different disciplines including physical oceanography, chemistry, geology and geophysics, and assessing the implications for their wider adoption. In ODIP II also attention is given to data management for marine biological data. This has been approached so far by including presentations and discussions at the ODIP II Workshops highlighting a number of leading marine biology data management systems, projects and initiatives.

Also of relevance is the integration of marine biological data management in the wider landscape of marine and ocean marine data management. This is highly relevant as there are many users and use cases that are interested in multidisciplinary applications combining marine environmental data sets from the various systems.

For that purpose the ODIP II Prototype 5 project has been formulated: **Integration of data management for biological and physicochemical marine data**

Lead:

VLIZ (Europe) + IODE (Global) + IMOS (Australia)

Aims:
To analyse the usability of the MEOP database ("Marine Mammals Exploring the Oceans Pole to Pole") within the context of the OBIS-ENV-DATA scheme and to assess if both data schemes can match in order to exchange information between the physical environment and the occurrence of a certain species between both data systems. This use case will also be relevant to examine in general differences between marine data schemes developed within different communities (biology and physics).

Approach:

Through **OBIS-ENV-DATA**, OBIS has successfully developed information technology solutions for combined data, as recommended in the OBIS-ENV-DATA proposal in 2015, as adopted by the IOC Committee on IODE. The combined data identified in the OBIS-ENV-DATA proposal addresses datasets provided to OBIS that include both biological and environmental data (hence "combined" data). OBIS's technology solution addresses not only combined biological and environmental data, it also incorporates details about sampling methods and effort, it expands OBIS's capacity for biological details, it enables OBIS to organize, aggregate, and link ocean observation events using "event hierarchy" and it implements identifiers to reference standard vocabulary for the parameters involved in biological, environmental, and sampling details. Also OBIS's solution maintains compatibility with Darwin Core, including use of the Darwin Core / GBIF capability known as "Event Core". OBIS can now refer to this complete combination of data features, including biological, environmental, sampling details and event hierarchy, as "OBIS Event Data".

The **MEOP consortium** (MEOP stands for "Marine Mammals Exploring the Oceans Pole to Pole") brings together several national programmes to produce a comprehensive quality-controlled database of oceanographic data obtained in Polar Regions from instrumented marine mammals. The MEOP database also should contain very valuable information on the

occurrence, behavior and migration of the tagged animals. MEOP currently proposes three data formats. They can be easily read in Ocean Data View, or using your favorite data processing software (e.g. Python, Matlab, IDL). Matlab tools and python tools to read and manipulate files in netCDF format are also available publicly. For a thorough scientific use of the data, or for oceanographic data centres, it is advised to use marine mammal netCDF format.

The overall approach will consist of comparing data sets and related data schemes from the 2 systems that combine occurrence and physical data. This should lead to further insight about compatibility and development of a demonstrator for integration.

Activities:

- Identify relevant datasets within MEOP and OBIS that contain occurrence and physical data;
- Cross match data schemes between data systems, identify commonalities & differences;
- Pilot dataset: integrate occurrence/physical data into relevant data systems and propose mechanism of deep linking;
- Reporting results of the ODIP II Prototype 5 and develop a plan for larger scale implementation by means of automation.

Appendix A: Terminology

Term	Definition
AODN	Australian Ocean Data Network
API	Application Programming Interface (API): a set of routine definitions, protocols, and tools for building software and applications
AURIN	Australian Urban Research Infrastructure Network
CDI	Common Data Index metadata schema and catalogue developed by the SeaDataNet project
CF	Climate and Forecast conventions: metadata conventions for the description of Earth sciences data, intended to promote the processing and sharing of data files http://cfconventions.org/
CoESRA	Collaborative Environment for Ecosystem Science Research and Analysis
CSR	Cruise Summary Reports is a directory of research cruises.
CSW	Catalog Service for the Web (CSW): OGC standard for exposing a catalogue of geospatial records in XML on the Internet
DataCite	Global non-profit organisation that provides persistent identifiers (DOIs) for research data to support improved citation https://www.datacite.org/
DIVA	Data-Interpolating Variational Analysis (DIVA) software
DOI	Digital Object Identifier (DOI): a unique persistent identifier for objects which takes the form of a unique alphanumeric string assigned by a registration agency

DQS	Data Quality Strategy
EDMO	European Directory of Marine Organisations
EMODnet	EU-funded intiative to develop and implement a web portal delivering marine data, data products and metadata from diverse sources within Europe in a uniform way. http://www.emodnet.eu/
EMU	Ecological Marine Unit
GEO	Group on Earth Observations: a voluntary partnership of governments and organizations supporting a coordinated approach to Earth observation and information for policy making
GEO-DAB	Brokering framework developed and implemented by GEO for interconnecting heterogeneous and autonomous data systems http://www.geodab.net/
GeoNetwork	An open source catalogue application for managing spatially referenced resources. It provides a metadata editing tool and search functions as well as providing embedded interactive web map viewer
GEOSS	Global Earth Observation System of Systems: international initiative linking together existing and planned observing systems around the world http://www.earthobservations.org/geoss.php
GitHub	Distributed revision control and source code web-based Git repository hosting service.
GML	Geography Markup Language (GML): XML grammar defined by the OGC to express geographical features
HDP	High Performance Data file format
ICES	International Council for the Exploration of the Sea http://www.ices.dk/

IMOS	Integrated Marine Observing System: Australian monitoring system; providing open access to marine research data http://imos.org.au/
INSPIRE	EU Directive (May 2007), establishing an infrastructure for spatial information in Europe to support Community environmental policies, and policies or activities which may have an impact on the environment.
IOC	Intergovernmental Oceanographic Commission of UNESCO (IOC/UNESCO).
IODE	International Oceanographic Data and Information Exchange" (IODE) of the "Intergovernmental Oceanographic Commission" (IOC) of UNESCO
IOOS	US Integrated Ocean Observing System https://ioos.noaa.gov/
ISO	International Organization for Standardization http://www.iso.org
jOAI	Java-based OAI software that supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), version 2.0 http://www.dlese.org/oai/
JSON	JavaScript Object Notation: an open-standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs. It is the most common data format used for asynchronous browser/server communication.
MarineID	Registration and authentication services for selected marine data services including SeaDataNet and EMODnet
MARVL	Marine Virtual Laboratory
MCP	Marine Community Profile: ISO19115 profile developed by Australian Ocean Data Centre Joint Facility (AODCJF) for marine data
MEOP	Marine Mammals Exploring the Oceans Pole to Pole

MIKADO	Java-based software tool, for creating XML metadata records for the SeaDataNet directories EDMED, CSR, EDMERP, CDI and EDIOS.
MNF	Marine National Facility is owned and operated by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) http://mnf.csiro.au/
MQTT	Message Queue Telemetry Transport is an ISO standard publish-subscribe-based "lightweight" messaging protocol for use on top of the TCP/IP protocol
NetCDF	Network Common Data Form (NetCDF): a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.
NCEI	NOAA's National Centers for Environmental Information https://www.ncei.noaa.gov/
NCI	National Computational Infrastructure (Australia)
NERDIP	National Environmental Research Data Interoperability Platform (Australia)
O&M	Observations and Measurements: OGC standard defining XML schemas for observations, and for features involved in sampling when making observations
OBIS	Ocean Biogeographic Information System
ODIP	Ocean Data Interoperability Platform http://www.odip.org
ODP	Ocean Data Portal: data discovery and access service, part of the IODE network http://www.oceandataportal.net/portal/

ODV	Ocean Data View: a software package for the interactive exploration, analysis and visualization of oceanographic and other geo-referenced profile, time-series, trajectory or sequence data
OGC	Open Geospatial Consortium: international voluntary consensus standards organization http://www.opengeospatial.org/
OIA-PMH	Open Archives Initiative Protocol for Metadata Harvesting https://www.openarchives.org/pmh/
OpenDAP	Open-source Project for a Network Data Access Protocol: a data transport architecture and protocol widely used by earth scientists https://www.opendap.org/
OpenSearch	Collection of technologies that allow publishing of search results in a format suitable for syndication and aggregation http://www.opensearch.org/Home
ORCID	Open Researcher and Contributor ID: a non-proprietary alphanumeric code to uniquely identify scientific and other academic authors and contributors http://orcid.org/
POGO	The Partnership for Observation of the Global Oceans: a forum created by the major oceanographic institutions around the world to promote global oceanography. http://www.ocean-partners.org/
R2R	Rolling Deck to Repository: a US project responsible for the cataloguing and delivery of data acquired by the US research fleet.
RDF	Resource Description Framework (RDF): family of W3C specifications for conceptual description or modeling of information that is implemented in web resources https://www.w3.org/RDF/
REST	REpresentational State Transfer (REST): an architectural style, and an approach to communications often used in the development of web services
SensorML	OGC standard providing models and an XML encoding for describing sensors and process lineage

SOA	Service Oriented Architecture
SOS	Sensor Observation Service: a web service to query real-time sensor data and sensor data time series. Part of the Sensor Web
SPARQL	SPARQL Protocol and RDF Query Language: a semantic query language for databases, able to retrieve and manipulate data stored in Resource Description Framework (RDF) format http://www.w3.org/TR/rdf-sparql-query/
SWE	Sensor Web Enablement: OGC standards enabling developers to make all types of sensors, transducers and sensor data repositories discoverable, accessible and useable via the web
THREDDS	Thematic Real-time Environmental Distributed Data Services, is a webserver, developed by UNIDATA, to facilitate provision of coherent access to metadata and data, both real-time and archived
US-NODC	US National Oceanographic Data Centre (now the NOAA National Centres for Environmental Information) https://www.nodc.noaa.gov/
VRC	Virtual Research Community
VRE	Virtual Research Environment
W3C	World Wide Web Consortium: main international standards organization for the World Wide Web http://www.w3.org/
WCS	Web Coverage Service Interface Standard: OGC standard defining Web-based retrieval of coverages i.e. digital geospatial information representing space/time-varying phenomena http://www.opengeospatial.org/standards/wcs

WFS	Web Feature Service: standards allowing requests for geographical features across the web using platform-independent calls
WMS	Web Map Service: standard protocol for serving geo-referenced map images over the Internet
WOA	NOAA's World Ocean Atlas
WoRMS	World Register of Marine Species
XML	Extensible Markup Language: a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable http://www.w3.org/XML/