Ocean Data Interoperability Platform II

Deliverable 2.5: Minutes of the 1st ODIP II Workshop

Work Package	**WP2**	**ODIP workshops**
Author (s)	Sissy Iona	HCMR
Author (s)	Helen Glaves	NERC
Author (s)		
Author (s)		
Authorized by	Helen Glaves	NERC
Reviewer		
Doc Id	ODIP II_WP2_D2.2	
Dissemination Level	PUBLIC	
Issue	1.0	
Date	23 February 2016	

Document History				
Version	**Author(s)**	**Status**	**Date**	**Comments**
0.1	Sissy Iona (HCMR)	DRAFT	23 February 2016	First draft
0.2	Sissy Iona (HCMR)	DRAFT	7 March 2016	Incorporating feedback from workshop participants
0.3	Helen Glaves	DRAFT	31 May 2016	Editing and review
1.0	Helen Glaves	FINAL	01 June 2016	Final edits and sign-off

Contents

Executive Summary

The 1st ODIP II Workshop was held at the IBIS Alésia Montparnasse hotel, Paris, France on 28 September to 1 October 2015. Local logistical support for the workshop was provided by IFREMER in consultation with the leader of WP2 (HCMR).

This workshop was the kick-off meeting for the ODIP II project which began on 1 April 2015. The second phase of the Ocean Data Interoperability Platform (ODIP) has a wider scope than the previous project with the inclusion of new partners and covering more disciplines.

The 1st ODIP II workshop focused on reviewing the results and possible follow-up actions for the three ODIP prototype development tasks that were developed in the previous ODIP project, providing updates on the cross-cutting topics including the introduction of two new themes on model workflows and big data, and capturing ideas for additional cross-cutting topics and new prototype development activities for ODIP II.

The topics addressed during the workshop were:

- **ODIP 1 prototype development task**: establishing interoperability between the SeaDataNet CDI, US NODC, and IMOS MCP data discovery and access services using a brokering technologies and moving towards integration with the global IODE-ODP and GEOSS portals
- **ODIP 2 prototype development task**: establishing interoperability between cruise summary reporting systems in Europe, the USA and Australia, with possible use made of GeoNetwork, for interacting with the POGO portal
- **ODIP 3 prototype development task**: developing a common marine profile of OGC standards for sensor web enablement (SWE).
- Vocabularies/Persistent identifiers
- Data publication and citation
- Model workflows and big data

More than 50 oceanographic data management experts from the three participating regions (Europe, USA and Australia) as well as representatives from the international IOC-IODE initiative took part in the workshop.

This deliverable, *D2.5 Minutes of the 1st ODIP II Workshop*, documents the organization, participation, proceedings and outcomes of the 1st ODIP II Workshop. All presentations made during the workshop are available from the IODE website with links to each one also embedded in the relevant section below.

1 Introduction

ODIP II: Extending the Ocean Data Interoperability Platform is the successor to the ODIP: Establishing an Ocean Data Interoperability Platform project. ODIP II continues to promote the development of a common global framework for marine data management by establishing interoperability between existing regional e-infrastructures in Europe, USA and Australia and also with global infrastructures such as GEOSS, IOC-IODE and POGO.

Building on the collaborative relationships developed during the first phase of the project, ODIP II will organize four international workshops to foster the adoption of common standards and develop prototypes to evaluate and test selected potential solutions for establishing improved interoperability across the selected regional and global marine data infrastructures.

The 1st ODIP II Workshop took place on 28 September to 1 October 2015 at the IBIS Alésia Montparnasse Hotel, Paris, France. It was organized by partner IFREMER in consultation with the leader of WP2 (HCMR). The aim of the 1st ODIP II workshop was to build on and further develop the outcomes of the first phase of the Ocean Data Interoperability Platform (ODIP) project as well as planning the additional activities described in the Description of Action (DoA) for ODIP II.

2 Participants

During the first phase of the ODIP project, and as part of its on-going communication strategy, an extensive mailing list of more than 100 experts representing the ODIP community is maintained and continues to be expanded. Using a similar approach to that taken in the first phase of the project, this mailing list together with the ODIP website was used to invite participants to attend the 1st ODIP II workshop. As a result, fifty attendees from nine countries took part in the workshop with 10 of them participating remotely via video conferencing.

The attendees at the 1st ODIP II workshop included representatives from the majority of the European, US and Australian regional data infrastructure projects and initiatives that are participating in the ODIP II project as well as a diverse range of relevant experts who were also invited to join this meeting.

Name	Affiliation
Robert ARKO (RA)	LDEO, USA
Christian AUTERMANN (CA)	52°North, Germany (remote participation)
Jean-Marie BECKERS (JMB)	ULG, Belgium
Sergey BELOV (SB)	RIHMI-WDC, Russian Federation
Justin J.H. BUCK (JB)	BODC, UK
Alberto BROSICH (AB)	OGS, Italy (remote participation)
Raquel CASAS (RC)	CSIC/UTM, Spain
Cyndy CHANDLER (CC)	WHOI, USA
Anne CHE-BOHNENSTENGEL (ACB)	BSH, Germany
Kinda DAHLAN (KD)	UCL, UK
Francisco S. DIAS (FD)	VLIZ, Belgium
Paolo DIVIACCO (PD)	OGS, Italy (remote participation)
Jocelyn ELYA (JE)	FSU COAPS, USA (remote participation)
Michele FICHAUT (MF)	IFREMER, France
Christiano FUGAZZA (CF)	IREA – CNR, Italy (remote participation)
Oscar GARCIA (OG)	CSIC/UTM, Spain
Helen GLAVES (HG)	BGS, UK
Jonathan HODGE (JH)	CSIRO, Australia
Sissy IONA (SI)	HCMR, Greece
Simon JIRKA (SJ)	52°North, Germany
Jonathan KOOL (JK)	Geoscience Australia, Australia
Alexandra KOKKINAKI (AK)	BODC, UK

Adam LEADBETTER (AL)	MI, Ireland
Thomas LOUBRIEU (TL)	IFREMER, France
Roy LOWRY (RL)	BODC, UK
Angelos LYKIARDOPOULOS (AL)	HCMR, Greece
Ana MACARIO (AM)	AWI, Germany
Sebastien MANCINI (SM)	IMOS, Australia
Youdjou NABIL (YN)	RBINS-BMDC, Belgium
Friedrich NAST (FN)	BSH, Germany
Elena PARTESCANO (EP)	OGS, Italy
Jay PEARLMAN (JP)	IEEE, USA
Francoise PEARLMAN (FP)	IEEE, USA
Leda PECCI (LP)	ENEA, Italy
Roger PROCTOR (RP)	UTAS, Australia (remote participation)
Lesley RICKARDS (LR)	BODC, UK
Dick SCHAAP (DS)	MARIS, Netherlands
Serge SCORY (SS)	RBINS-BMDC, Belgium
Adam SHEPHERD (AS)	WHOI, USA (remote participation)
Shawn SMITH (SS)	FSU COAPS, USA (remote participation)
Jean Marc SINQUIN (JMS)	IFREMER, France (remote participation)
Shane St CLAIR (SSC)	Axiom Data Science, USA
Rob THOMAS (RT)	BODC, UK
Charles TROUPIN (CT)	SOCIB, Spain
Mickaël TREGUER (MT)	IFREMER, France
Sebastien TREGUER (ST)	La Paillasse Ocean Project, France
Thomas VANDENBERGHE (TV)	RBINS-BMDC, Belgium
Rob VAN EDE (RvE)	TNO, Netherlands
Matteo VINCI (MV)	OGS, Italy (remote participation)
Lesley WYBORN (LW)	NCI, Australia

3 Workshop Agenda

The 1st ODIP II workshop aimed to build on and further develop the outcomes of the previous ODIP project and to also initiate the additional activities planned for the ODIP II project. The scope of ODIP II has been extended to include other disciplines and additional partners. The workshop agenda included sessions to introduce the project to the new partners and also highlight some of the additional themes and objectives outlined for ODIP II in the description of action (DoA).

As for previous workshops, the programme included a dedicated session for each of the existing prototype development tasks. These sessions provided a final progress report for each of these tasks as a wrap-up for the previous ODIP project activities, and also presented an opportunity to identify potential extensions for these prototype interoperability solutions which can be developed as part of ODIP II. The other sessions included in the agenda were used to introduce some of the new themes added for ODIP II and also to formulate further prototype development tasks.

The three recurring discussion topics which were identified and discussed during the previous ODIP project workshops were also included in the agenda for this meeting. These sessions provided an update on recent developments in these areas and were also used as an opportunity to identify further relevant cross-cutting topics that could be included in future workshops.

The workshop agenda was developed by the ODIP II coordinator in consultation both with the Steering Committee and also the session leaders. It was then circulated to all ODIP partners by e-mail for final approval prior to the start of the workshop. The final version of the agenda was then published on the public ODIP II website (http://www.odip.org).

Workshop Sessions

As with previous workshops, the programme for the 1st ODIP II workshop was made up of a series of sessions (see Table 1) addressing specific topics of interest for the ODIP project partners. Each session was led by a nominated chairperson who was responsible for developing the detailed agenda for their part of the workshop programme (see Annex A).

Session	*Title*	*Leader*
1	Introduction	*Helen Glaves*
2	ODIP Prototype 1	*Dick Schaap*
3	ODIP Prototype 2	*Anne Che-Bohnenstengel & Friedrich Nast*
4	ODIP Prototype 3	*Jonathan Hodge*
5	ODIP prototype development tasks: feedback on outcomes and possible next steps	*Helen Glaves*
6	Vocabularies/Persistent identifiers	*Roy Lowry*
7	Model workflows and big data	*Adam Leadbetter*

8	Data publication and citation	*Justin Buck*
9	Cross-cutting topics: break-outs	*TBA*
10	Cross-cutting topics break-out session reports	*Helen Glaves*
11	ODIP II: new development activities & cross cutting themes	*Dick Schaap*
12	Workshop wrap-up	*Helen Glaves*

Table 1 Workshop sessions including nominated chairperson

4 Workshop proceedings

Extensive minutes were taken during the workshop and these are documented below. Where appropriate actions have been recorded and a list of these actions is included to Annex B

All presentations made during the workshop are hosted by IODE (http://www.iode.org/index.php?option=com_oe&task=viewEventRecord&eventID=1737) and made available via the ODIP (http://www.odip.org) website under the "Workshops" menu option. Other reference documentation relating to the prototype development tasks can also be found on the ODIP web site.

4.1 DAY 1: SESSION 1 - Introduction

4.1.1 Welcome

The 1st ODIP II workshop was opened by Helen Glaves (ODIP coordinator) on Monday 28 September 2015, at the IBIS Alesia Montparnasse Hotel in Paris, France.

HG welcomed the participants, thanked the organizers and explained the logistics for the meeting. She introduced the agenda and the format of the meeting which would follow that used in previous workshops.

Participants were then given the opportunity to introduce themselves providing details of their affiliation, role and expected contribution to the ODIP II project.

4.1.2 ODIP II: Overview including aims and objective

HG began by giving a short introduction to ODIP II. This project is the second phase of the Ocean Data Interoperability Platform project which was recently funded by the EU. The successful proposal for ODIP II was submitted to the Horizons 2020 INFRASUPP-6-2014 call on 6th September 2014. As a result, the three-year follow-on ODIP II project officially started on 1 April 2015.

The basic concept of ODIP is to support multilateral cooperation on research data infrastructures in marine science. It is a collaborative project between Europe, the USA, Australia and related international initiatives such as IODE, GEOSS and POGO.

The key objectives are to:

- continue and extend the activities of the previous ODIP project;

- provide a coordination platform to facilitate the establishment of interoperability between regional data infrastructures in Europe, USA and Australia and also with global systems e.g. IODE Ocean Data Portal, GEOSS, POGO;
- formulate joint prototype interoperability solutions including the further development of the existing prototypes to fully operational systems to demonstrate this coordinated approach;
- establish common approaches for specific aspects of marine data management e.g. vocabularies, formats, sensor web enablement etc.;
- extend the scope of the project activities to include other domains e.g. marine biology and additional partners.

ODIP II will facilitate organized dialogue between key organizations in Europe, USA and Australia involved with the management of marine data through a series of workshops. It will also seek to engage organizations and data infrastructures dealing with marine data in other regions e.g. Canada, Asia etc.

ODIP Project Structure

The project management structure is similar to that for the previous ODIP project with NERC as Coordinator and MARIS acting as Technical Coordinator.

The European consortium has been enlarged from 10 partners in the previous project to include 19 partners from 9 countries in ODIP II. However, an unsuccessful proposal to the National Science Foundation by the R2R project has led to uncertainty about the level of participation by some partners from the USA. Four US partners (SIO, WHOI, LDEO, FSU) representing the R2R project will continue to participate in ODIP II for the foreseeable future whilst efforts are made to secure other sources of funding. As discussed at the 4th and final ODIP workshop funding is also still a major challenge for partners from Australia due to a change in priorities for research.

It was also noted that an expression of interest in joining the ODIP II project has been received from Ocean Networks Canada (ONC).

ODIP II includes a Steering Committee which acts as a strategic management board for the project. The membership of this committee is not static and may be modified as ODIP II evolves to ensure that it remains representative of the entire project. It is made up of the Coordinators, WP leaders and other representatives from the participating regions.

It was highlighted that there are currently a small number of vacancies on the Steering Committee that need to be filled as soon as possible. It was also observed that the current membership does not adequately reflect the project consortium as a whole as there is no representation from the biological oceanography community. A suitable person will need to be nominated from this community.

ACTION 1.1: Coordinators to identify a Steering Committee member for NOAA

ACTION 1.2: Coordinators must seek a nomination for a representative of the biological oceanography to sit on the Steering Committee

ODIP II also includes a Partners Committee which is made up of representatives from all of the consortium partners in the three participating regions and also from the international initiatives contributing to the project i.e. IODE, POGO, GEOSS. The Partners Committee is required to meet at regular intervals and it has been agreed with the EU Project Officer that this meeting will form part of the ODIP II workshops which are planned to take place twice a year.

The meeting then discussed other potential contributions such as those from related EU-funded projects e.g. ENVRIplus, relevant organizations e.g. JCOMMOPS and other associations and initiatives such as the RDA Marine Data Harmonization Interest Group.

HG concluded by informing the European partners about the allocation of resources across the different work packages. It was highlighted that resources for Other Direct (OR) costs are relatively high on this project due to the need for a significant amount of travel to attend the workshops, some of which are outside Europe.

Partners were reminded that allocation of resources must be prioritized to ensure that EU-funded partners attend all workshops planned in the DoA. Any requests to use ODIP II resources to attend conferences, other meetings etc. will only be approved by the Coordinator on the condition that this does not compromise a partner’s ability to attend the ODIP II workshops.

HG concluded with some information concerning the project initiation including the mechanisms for making pre-financing payments to partners.

4.1.3 ODIP II: development of potential activities

Dick Schaap (MARIS), ODIP II Technical Coordinator welcomed everybody to the second phase of the ODIP project. DS highlighted the fact that there are now many projects either just starting or planned for the future that are related to ODIP and make use of its approach and outcomes. ODIP is a coordination platform that brings together expertise, ideas and developments which can be aligned and implemented across other projects to support interoperability. At a more technical level the approach of the ODIP projects is to develop interoperability between existing regional marine e-infrastructures in order to create a global framework for marine data management.

To support the expansion of the existing prototype development tasks and also the formulation of new additional prototype interoperability solutions, ODIP II brings together expert developers and the managers of leading regional and global infrastructures to present and discuss various relevant topics and technologies.

As well as this content based approach there are also many IT challenges that will need to be addressed, for example, implementation of new OGC and ISO standards. The advent of the ‘Internet of Things’ will also bring significant changes and opportunities for using network services but for now the challenge is how to deal with access to the flow of data and metadata from new ‘plug-and-play technologies, for example, from the observation sensors using sensor web enablement (SWE).

ODIP II will also try to interconnect the existing data systems which have been developed using a bottom-up approach, and attempt to combine the data from these different sources to turn them into information and knowledge using various technologies. DS encouraged partners to bring forward new topics and themes that could potentially be included in the ODIP II project activities either for incorporation in the prototype development tasks or as additional cross-cutting themes to be addressed during future workshops.

A number of prototype interoperability solutions will be formulated and taken forward into development, largely by leveraging on the activities of current regional projects and initiatives such as SeaDataNet, EMODnet (Europe), IMOS and AODN (Australia), R2R, US NODC, UNIDATA and US IOOS (USA), and in consultation and direct cooperation with global initiatives such as IODE-ODP, GEOSS and POGO. ODIP II will function partly as a "think-tank" with agreed solutions carried forward by the related data infrastructures for further development, testing and, if successful, wider implementation and operation.

DS continued by providing an overview of the extensive list of potential topics for consideration in ODIP II which might also give rise to other additional ideas during the workshop. These ideas for additional topics to be addressed by ODIP II will then be used to formulate new prototype development tasks and cross-cutting themes. In addition, the three prototype development tasks initiated during the previous ODIP project will also be continued and extended within ODIP II.

Examples of topics to be addressed in ODIP II are:

1) Standardization of NetCDF which will be undertaken in cooperation with UNIDATA (the originator of this format and a contributor to ODIP II).
2) Extending the use of the data statistical analysis UIG/DIVA software in applications other than the SeaDataNet as users are becoming increasingly interested in data analysis products rather than the data itself.
3) Usage of controlled vocabularies has become a necessity in all platforms and domains and the ODIP activities on this cross-cutting activity will be continued.
4) Standardization of geographic marine names, including ocean basins, seas, seamounts, sandbanks and other sea features.
5) Harvesting of data from several sources, automated aggregation with duplicate elimination, gridding with on-line visualization tools, quality control issues and prototyping of aggregations of marine resources.
6) Data publication and citation as a mean to encourage researchers to open and publish their data
7) Use of WPS for processing near real-time data streams
8) Clouds technologies and finding the way to connect them will give new opportunities for horizontal data processing. In the recently submitted EU proposal on data ingestion systems, a lot of ideas and material were used from the previous ODIP workshops to formulate the new proposal.
9) Provenance of data from different sources for version control or use in environmental, management, or policy issues is becoming increasingly important and common standards for capture of provenance information will be explored.
10) Interoperability between operational marine observations systems - the multidisciplinary interoperability / System of Systems approach used by initiatives such as GEOSS and the GCI, EarthCube with BCube etc. will be evaluated because performance issues will also need to be addressed.

DS concluded the session by inviting partners to contribute new ideas and propose additional activities in order to widen the range of the potential future topics that could be addressed as part of ODIP II.

4.1.4 Discussion

It was agreed that a list of relevant topics would be maintained throughout the workshop so partners could add new items, comments and ideas.

Communication with RDA on relevant topics must also be ensured to allow sharing of information and expertise between the two initiatives. Several key people from RDA interest groups were identified that could be directly involved in both ODIP II and RDA. JP noted that there are a number of EU projects dealing with sensors, SWE, etc. and proposed that ODIP II makes manufactures of these instrument and platforms aware of the requirements for standards and/or data formats. HG noted that representation of manufactures is currently missing in ODIP II as well as in RDA and this connection needs to be made. TL noted that the people from the "Ocean of Tomorrow" projects should also be invited to participate in relevant RDA activities.

ACTION 1.3: TL to inform Oceans of Tomorrow project of relevant RDA IG/WG

The meeting discussed the problem of engagement of the manufactures and how to make them more interested in the requirements for standards by the marine research community. It was acknowledged that their current business model dictates that they are only interested in the requirements of those customers that are potentially purchasing large amounts of equipment (commercial companies, defense contracts etc.). There is a need to identify an area of common interest for the equipment manufacturers and marine researchers.

Another issue, and a significant challenge for ODIP II, is that the ocean community is not yet in a position to propose a unified approach to SWE because different groups and also the private sector use different standards. DS informed the group that a dedicated workshop addressing sensor web enablement for oceanography was being organized by the Eurofleets project which would take place during the Oceanology International 2016 conference in London, U.K. Partners from several projects and initiatives will discuss how to develop common marine profiles of the OGC SWE standards and sensor manufactures will be invited to share their views on their adoption.

4.2 SESSION 2 - ODIP Prototype Development Task 1: plenary

4.2.1 ODIP 1: aims, activities and progress

Dick Schaap, ODIP Technical Coordinator provided an overview of progress on the ODIP 1 prototype development task. This task aims to establish interoperability between the key regional data discovery and access services in Europe (SeaDataNet), the USA (US NODC) and Australia (AODN) and also with the leading global portals (GEOSS portal and the IODE - Ocean Data Portal), using the GEO-DAB brokerage service.

The GEO-DAB broker service harvests XML metadata output from the regional data systems and converts it to the generic model which is then used to populate the GEOSS portal. The GEO-DAB broker which is managed by partner CNR, has been developed through several projects in both Europe and the USA.

The agreed approach has been to use the broker service to deliver metadata to the GEOSS and ODP portals at the collection level and not at the individual granule level which represents millions of datasets. (Collections are aggregations of metadata records based on specific criteria e.g. originator, vocabulary term etc.). Access to the individual datasets at the granule level is achieved by navigation to regional data systems.

The initial plan was to start by integrating the SeaDataNet data discovery and access services with the global services using the broker, and then use the same approach for the US NODC (USA) and AODN (Australia) systems. The SeaDataNet discovery and access service incorporates a distributed network of more than 100 data centres that use OGC-ISO and INSPIRE compliant standards. There are currently 1.8 million metadata records available for individual datasets at the ‘granule’ level which have been aggregated into 400 collections using parameters such as data type, data providers, discipline and geometry type (point, track, area).

A REST web service has been set-up (IP – IP protected) which allows dynamic harvesting of the XML metadata records from the regional data discovery and access service by the broker service. These records are then used to populate the GEOSS and ODP portals.

DS highlighted the fact that, due to a misunderstanding with CNR, SeaDataNet has been listed as the domain in the current OAI-PMH interface so that it appears that SeaDataNet is harvesting from the regional systems rather than it actually being one of the contributing services. An OGC Catalogue Service for the Web (CSW) is also available but there are still some issues with this that need to be resolved.

DS continued by explaining the process model for implementation of the GEO-DAB broker between the SeaDataNet, and the GEOSS and ODP services. The GEOSS portal dynamically harvests the XML metadata for the SeaDataNet collections from the CSW service and imports it into the GEOSS portal, while the ODP portal harvests from the OAI-PMH service using jOAI. As a result, the SeaDataNet collections are now included and maintained in both the GEOSS and ODP portals. Triggers have been put in place so any update in the source system results in the change being automatically propagated throughout the chain to ensure sure that all content remains up to date. DS then showed how users can use the global portals to discover these collections and, through dedicated URLs, drill down to the SeaDataNet portal for further details at the granule level including submission of requests for access to the data.

A similar approach has been implemented for the US NODC which provides services both at the granule and at collections level (approximately 28,000 collections entries). The collection definition, although different from SeaDataNet, is still fit for the purposes of ODIP II. In the case of US-NODC, a collection can be data from an individual scientific project while in SeaDataNet, a collection is an aggregation of data sets from one data center collected from many projects e.g. thousands of geological samples held by one institute for a whole marine region. There is an extensive range of mechanisms supporting user access to the data: OPeNDAP, Hyrax, THREDDS, Live Access Server, ftp and http links, which are provided as links from the metadata collections. Comparable links exist at the granules level; the only difference is that instead of URLs (for collections) there are data links (for granules). Within the collections there are URLs for metadata description (as XML) and these links redirect to the data themselves. As with the SeaDataNet service, the broker harvests the collections from the US-NODC (28 000 entries), converts them to the generic XML schema, and then populates the GEOSS and ODP portals.

The final phase of development addresses the integration of the AODN service which makes use of GeoNetwork technologies. AODN is also delivering collections but these are aggregated using different criteria and are much smaller in number. It was also highlighted that AODN uses the Dublin core metadata profile because the ISO 19139 does not provide links to the data. The AODN collections are provided as CSW, OAI-PMH and OpenSearch endpoints for discovery. At present there are around 110 collections available but this will increase with time. DS illustrated this point by showing the AODN 1,2,3, web interface for discovery and access to the IMOS data catalogues. The AODN brokerage page which propagates to the global portals was demonstrated to illustrate how it can also be discovered in the GEOSS and ODP systems in a similar fashion to US-NODC and SeaDataNet. (It was also noted that the data labelled as ODIP in the Ocean Data Portal (ODP) needs to be revised to show the regional data system names).

The approach for the ODIP 1 prototype development task is essentially the same for each of the three regional data systems: metadata for the collections is harvested from the regional system it is then exposed in the global portals where users can discover these collections and navigate back to the regional portals for more information and in many cases direct access to the individual datasets (the 'granule').

DS concluded by confirming that the objectives for the ODIP 1 prototype development task had been achieved with a few minor issues remaining to be addressed. These are:

a) check the exact CSW URLs with CNR and also correct the inconsistencies with how the different regional data systems are displayed in the available web services

b) check that the numbers of collections delivered by the regional portals are the same as those actually being harvested and exposed in the global GEOSS and ODP services

c) check that the maps on the global portals are correctly representing the spatial information for the collections from the regional systems

d) report the achievements of the ODIP 1 prototype development task in deliverable *D3.4 Results and conclusions from prototype analyses*.

ACTION 1.4: CNR to correct current issues with CSW and OAI-PMH services

ACTION 1.5: Checks to be made to ensure numbers of collections delivered by the regional systems equate to the numbers being harvested and delivered to the global portals

ACTION 1.6: Checks to be made on the accuracy of the spatial information shown in the global portals for the regional collections

ACTION 1.7: Partner MARIS to complete deliverable D3.4 Results and conclusions from prototype analyses

DS then invited questions or suggestions for further development of the ODIP 1 prototype interoperability solution. JP asked for clarification regarding whether the system is only harvesting metadata or also providing access to the data. DS indicated that the broker is currently only used for discovery metadata but the user can drill down from the GEOSS or ODP portals to the granule metadata and then the data itself via the individual regional systems.

JH asked if there are any plans to incorporate data service endpoints such as OpenSearch or WMS. DS indicated that these services are currently provided for the individual entries but have not been taken up to provide interoperability at the global level. However, there is scope for an OpenSearch or WMS portal to be built onto the regional level services. The current work is only at the discovery level using the GEO-DAB broker to do the necessary integration between the regional services, but there is scope to develop this prototype solution in this direction in the future.

The discussion then continued on some general issues relating to data interoperability such as semantics and user registration. TL noted that the issue of user registration was identified in the impact analysis for the ODIP 1 prototype solution during the previous workshop. For example, to move from metadata to data interoperability necessitates management of user identification to accommodate the data access restrictions. DS commented that this type of issue was the reason that the prototype solution began with metadata interoperability. Furthermore, data are not always directly usable because they are in different formats, there is a lack of semantic interoperability, and in some cases users are not satisfied when they have direct access to data from different data centers because they simply cannot use them.

The use of a brokering solution helps to address these issues as it customizes the data for the user's requirements.

The discussion on what should be done to further develop ODIP Prototype 1 continued later in the workshop.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord &docID=16144

4.2.2 ODIP 1: report on impact assessment

TL leader of WP4, introduced the outputs of the impact analysis from the first phase of the ODIP project. A cost/benefit analysis for the impact of implementing each prototype interoperability solution was compiled. The document was sent to all partners prior to this workshop for updating and reviewing. It included all the invaluable information compiled from the discussions during the prototypes sessions at previous workshops concerning the potential positive impacts, the cost implications and changes that need to be implemented at regional level for the adoption of the prototypes.

Following the 3rd ODIP workshop (Australia, August 2014) demonstration use cases and performance indicators for the implementation of the prototypes were identified. Potential enhancements for the prototype solutions during the second phase of the project were also documented which could include: 1) identification of demonstration use cases for the eventual collection of success stories; 2) evaluate impact using defined indicators; 3) definition of simple targets for each of the prototypes to easily measure their efficiency; and 4) develop a roadmap as a record of the ODIP 1 development activities throughout the project.

It was also evident that the readiness of the prototypes is quite heterogeneous. The ODIP 1 prototype solution is near completion with the merged metadata descriptions available in the portals. However, this is not the case for the ODIP 3 prototype which is still in the research/innovation phase.

Assessing the readiness level of the ODIP prototypes and measuring the potential impact despite some of the solutions not yet being operational has proved to be a valuable exercise for the project as it provided an opportunity to discuss the concept and schema of the Framework for Ocean Observing (http://www.ioc-goos.org/index.php?option=com_content&view=article&id=363&Itemid=100006&lang=en . This is the approach used by TL as part of the impact assessment process and is also well aligned with the overall aims and objectives of ODIP/ODIP II.

The Framework for Ocean Observing identifies the first level of 'readiness' as the concept – this is already available in the case of the ODIP prototypes. The resulting impact is that ODIP is valuable for research, innovation and the pooling of technological expertise. The second level of readiness are the pilot trans-regional demonstrators (the prototype interoperability solutions) which can be achieved by technology pooling (sharing software, standard profiles). The final target of this activity within the Framework for Ocean Observing is to determine an implementation scenario which would lead to the deployment of an operational trans-regional infrastructure but this is not part of the objectives for the ODIP II project.

For purposes of the impact assessment the following cost/benefit impact classification scheme is useful:

1) concept and regional pilot phase - minimal cost implications while the benefits are the pooling of research and innovation.

2) trans-regional pilot phase - potential costs for the regional systems associated with the need for maintenance and operation of new infrastructures, software interfaces etc. as part of the transition but the benefit would be sharing of the software, for example the use of GeoNetWork in the ODIP 2 prototype development task.
3) operational implementation phase - assessment is needed to determine what the end-user is expecting which may also lead to increased operational costs at data center level. However, the expected benefits are improved end user services and an overall lowering of operational costs at the regional and national data center level.

Two examples to illustrate this approach are:

1) The European EMODnet check point projects which assess the fitness for purpose of the data services for specific applications e.g. marine renewable energy studies, oil spill response etc. One expected benefit of the ODIP 1 prototype development task would be to define metadata that contains the quality information required to assess fitness for purpose of the data sets.
2) Enhancing the provenance information in observation metadata. ODIP 3 could propose a common implementation of SWE so that provenance information is homogeneously encoded at trans-regional level as part of a pilot implementation.

In summary, demonstration use cases should be identified which are involved in the definition of the prototypes and ultimately provide success stories that can be reported. In addition, simple and accurate targets should be identified to show the user benefits.

In the case of ODIP 1 the target is to populate the GEOSS and ODP with metadata records for the datasets available in the regional data services. The readiness status of this activity can be classified as “toward operational implementation”. The regional CSW services are available and connected to the GEOSS and ODP. A demonstration use case to support the Committee for Conservation of the Antarctic Marine Living Resources (CCAMLR) to establish a marine protected area (MPA) in the seas of the Southern Ocean has been drafted. Exemplars for performance indicators have also been drafted e.g. quantifying the number of datasets from the three regional data systems (SeaDataNet, US-NODC and AODN) added to the ODP and GEOSS portal as a result of the ODIP project activities.

TL concluded the results of impact analysis of the ODIP 1 prototype development task by presenting potential further enhancements for ODIP II, some of which are cross-cutting topics relevant to all of the prototype development tasks e.g. reference services (vocabularies, further population of EDMO). The conclusion was that maintenance and upgrade of NVS and EDMO was needed at the European level and mapping with NVS and EDMO was needed at USA and Australian level.

Another conclusion of the impact assessment was that the federation of identity would be a topic of interest. In Europe there is a collaboration with EduGain (which is used in GEO) and MarineID (developed for SeaDataNet).

Standard profiles for datasets description were also identified as a potential area of impact, with on-going work within several RDA working groups. The IODE/ODSBP process might also provide another potential opportunity for collaboration, especially on the issue of obsolescence management (deprecation and supersede) and the granularity of the datasets.

Final impact that was identified was the functionality of the brokerage service which is operated by CNR. The possibility of extending the metadata interoperability demonstrated in ODIP 1 to semantic interoperability with proper connection of metadata formats with vocabularies was discussed at the previous workshop.

TL invited partners to comment and make suggestions for further use cases. JB commented that there are scientists who do not know that they can find datasets from multiple sources in the ODP, or that ODIP is supporting this delivery of metadata from the regional data infrastructures. DS suggested that this should be the responsibility of the GEOSS and ODP initiatives and not ODIP. These data portals are propagating delivery of the datasets to create a 'one-stop shop' for the end user. The role of the ODIP project is to facilitate this integration of the regional data discovery systems into these global systems.

The GEOSS and ODP initiatives should do more to inform the community that the data are available but it was noted that it is also difficult for users to find what they are looking for. TL commented that cross-disciplinary users can benefit from the work done by ODIP work but for those working in one specific region or discipline there is still a tendency to use the thematic portals. It was noted that the CCAMLR use case is an example of cross-disciplinary usage. YN commented that data workflows linked with data services would help, and not only with human discovery of data resources. DS agreed that making resources discoverable and bringing them together is already happening but it does not always help a large number of users. Further services are needed to allow users to find data in common formats, processed data, and more generic products in addition to the data. YN asked if the minimum metadata are supplied by the data providers. RL confirmed that the ISO 19139 standard is used which has more mandatory information content than the Dublin Core profile which is also widely used.

TL stressed the need to find use cases in order to illustrate all the issues that have been mentioned. DS concluded that more steps are needed to make data more easily discoverable and readily accessible because users often cannot find the resources they need when using the large scale portals such as GEOSS. Users need services on top of these portals to ingest and process the millions of data records so as to make them more useable.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16145

4.2.3 Discussion

DS began this session with a discussion about the future direction of the ODIP 1 prototype development task in ODIP II. To date this prototype has only focused on metadata, but there is a need to make greater use of the data brokerage and services in order to improve data accessibility and to build product services on top of the existing systems.

The first step is to make the ODIP I prototype interoperability solution fully operational with dynamic propagation of updates to the other services when new entries are added to the local system. The next step is to start exploring the data brokerage and associated services to determine how they can be used to assist users and reduce their work load using automated systems to harvest and aggregate data.

The group then discussed the issue of metadata. JH commented that it would be interesting to look at bringing the document focused metadata approach and linked data approach closer together in order to evaluate linking provenance management systems with metadata to determine if the provenance metadata could replace the lineage elements in metadata records. DS noted that from the data managers' perspective the users require more metadata but the data providers cannot always supply it because it is not available especially in the case of historical records. Where the ISO 19139 metadata standard is used there is often a compromise made on the mandatory fields. However, in the case of SWE this compromise can be avoided because a significant proportion of the mandatory metadata fields have been set from the beginning.

DS commented that joining the metadata that are automatically produced by systems like Argo with the data or putting together metadata from many different sources is not only a technical issue but also one of governance. TL noted that this is a similar situation to data ingestion systems e.g. how to streamline inputs from providers. RvE noted that there was the potential for bringing prototype development tasks ODIP 1 and ODIP 3 together to create something bigger.

DS added that a follow-up activity for the ODIP 1 prototype development task 1 could be the harmonization of vocabularies. To date ODIP 1 has only dealt with the metadata and made use of the vocabularies in their existing form. There is as yet no harmonization of the vocabularies at the GEOSS or ODP level. This could be a small but important step to make although still only at the metadata level.

RvE asked if we know how many of the metadata elements under consideration are covered by INSPIRE. DS replied that INSPIRE is a European directive for harmonization of geospatial data, making them discoverable and accessible. Discussions with the research team working on INSIRE made it clear that population of INSPIRE should be done at the national level and not just through SeaDataNet. There is also an INSPIRE Marine Pilot project which has been launched to help improve the understanding of INSPIRE in the management of Marine Strategy Framework Directive (MSFD)-related spatial information, and to provide guidance and tools that facilitate the mentioned obligations. The INSPIRE Marine Pilot has adopted the SeaDataNet vocabularies (P01 and P02 for the marine domain). The EMONDET project is also working with the INSPIRE team with the aim of becoming fully INSPIRE compliant. The only remaining issues are the data models that are currently being defined by INSPIRE because additional work will need to be done for EMODnet/SeaDataNet to become compliant with these new models.

4.3 SESSION 3 - ODIP Prototype Development Task 2: plenary

4.3.1 ODIP 2: aims, activities and progress

The session was opened by Friedrich Nast (BSH), who has taken over as the leader of the ODIP 2 prototype development task from Bob Arko who led this task during the first phase of the ODIP project.

FN began by introducing the ODIP 2 prototype development task. He highlighted the fact that the strength of this prototype solution is that it is very focused on the one topic of cruise summary reports (CSRs). The discovery and access of the CSRs is already possible in all the three participating regions: in Europe through SeaDataNet, in the USA through R2R and in Australia through the MNF system. The aim is to integrate the three regional systems with the global POGO system which may seem relatively trivial but in reality that are a number of issues that need to be overcome to achieve this objective.

FN described the evolution of the cruise summary reporting (CSR) system. The system began with the paper ROSCOP (Report of Observations/Samples collected by Oceanographic Programmes) forms which were submitted to International Council for the Exploration of the Sea (ICES) who maintained the cruise database. This was superseded by a very inflexible digital MS Word version of the form.

During the EU-funded SeaSearch project an CSR on-line system was built using web technology. This was followed by the XML and GML versions. The most recent developments are moving towards automatic generation of CSRs. The German MANIDA project at BSH is helping to achieve this objective. At the European level the EU-funded Eurofleets project is developing a mechanism for automatic generation of CSRs from the ship's systems. However, there are still some problems with the event logging that have to be solved but significant progress has already been made.

The next step in the ODIP 2 prototype development task is to harvest CSRs from different originators into the regional systems which are then exposed in the global portal (POGO). Users discover the CSRs and access the associated information using the global POGO portal. The most useful next step would be to link the CSRs to the data through the metadata. There will be further discussions on the possible next steps for this prototype development task during the impact assessment and dedicated ODIP 2 discussion sessions.

4.4 ODIP 2 development task: progress and results

Bob Arko (LDEO), gave an overview of the results of the ODIP 2 prototype development task from the U.S. perspective. BA reported that there has been significant progress since the 3rd ODIP workshop with 130 new CSRs from U.S. vessels having been exposed in the POGO portal. To achieve this outcome R2R selected two specific research vessels (RV Falkor and RV Kilo Moana) and, with the assistance of BSH, were able to populate the CSR database. An additional benefit of this activity was the mapping of the R2R vocabulary terms to the EDMO codes and SeaDataNet vocabularies for ports (C38) and devices (L05).

BA showed an example of a CSR for a recent cruise of RV Kilo Moana which included extended cruise metadata. The next steps for the US contribution to the ODIP 2 prototype development tasks include:

1) Publishing the remaining older cruises (around 4600)
2) routinely publish new cruises either quarterly or annually (around 400 to 500 cruises each year as part of the R2R mandate and depending on funding)
3) improve cruise records by: a) populating Sea Areas using the C16 vocabulary; P02/P03 Discovery Parameters, P08 Disciplines; b) including detailed cruise abstracts; and c) ensuring all NSF funded investigators, not just Chiefs and Co-Chiefs, have an ORCID persistent identifier in an effort to link scientists to their outputs e.g. publications, datasets etc.
4) upgrading the GeoNetwork portal using the May 2015 release

BA concluded that the US contribution to ODIP 2 prototype development task had made good progress with workflows now in place for routinely publishing R2R CSRs to the POGO portal.

LW asked how R2R encourages its researchers to provide sufficient information in their ORCID profile for it to be useful. BA replied that the only thing they do is confirm that each R2R cruise has an ORCID attached to it. It is the responsibility of the scientists to provide sufficient information.

It was noted that this issue is of specific interest for the ODIP II project. BA commented that over the next three years R2R will focus on associating the thousands of scientists in their catalogue to the cruises using ORCIDs and DOIs. AM asked if R2R would be an issuing authority for ORCIDs for those scientists who do not yet have one. BA replied that R2R does not mint identifiers. The researcher only requires a link to the ORCID registration system and some guidance on the minimum information that should be provided. HG commented that her own organization (British Geological Survey) is now recommending that staff register for an ORCID and providing instructions on the minimum information that should be included in their profiles such as their name and affiliation (organization).

The rest of the discussion regarding persistent identifiers was postponed until the relevant session later in the workshop.

RL commented that the C16 vocabulary that BA mentioned would be used to populate Seas Area in the R2R CSRs is relatively poor compared to C19. He suggested that R2R should consider using the C19 vocabulary for populating Sea Areas. RL noted that the SeaDataNet CSR already uses it and that the C16 entries are incorporated into the C19 vocabulary.

DS also observed that the Australians have done a large amount of mapping between their organizations and European Directory of Marine Organizations (EDMO). SM noted that there is a lot of work being done on vocabularies by the MNF but not much on mapping the cruise reports.

Friedrich Nast asked Jonathan Hodge to act as the Australian representative for ODIP 2.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16152

4.4.1 CSR harvesting: update on progress

Anne Che-Bohnenstengel (BSH) reported on the status of the CSR harvesting. She first explained that up until the end of 2014 it was only possible to submit CSR using either an online Content Management System (CMS) or by sending XML records via email or FTP. Since the beginning of 2015, there has been a weekly CSW harvesting of the CSRs from the connected data centres. The requirements for harvesting are: a) creation of CSRs in ISO 19139 format either using MIKADO or another in-house software solution, and b) implementation of an OGC CSW service at the data centre.

The workflow for harvesting of CSRs is as follows: once the XML records have been placed on the local CSW server, the data centre contacts BSH who then test the harvesting of the CSRs. If this test is successful, the CSR records are the harvested by BSH. The next step is the quality control which consists of automatic checks on mandatory fields and vocabularies and, if these basic requirements are fulfilled, the content of the records will undergo further manual/visual checks. Where there are any inconsistencies in the records the data centre will be requested to make the necessary corrections.

Once the CSR is valid, the next step is to insert the record into the master database which is used as the central directory for publishing the CSRs. Each record has a unique BSH identifier as well as a local identifier that is defined by the originating data centre which is also unique when combined with its EDMO code.

A comparison between existing and newly harvested records is carried out. If the record already exists it will be updated, otherwise it will be inserted as a new entry into the central database. The records in the central CSR inventory are then automatically published on the SeaDataNet and POGO websites. ACB showed an example of the POGO website with the recently published CSRs from the USA and a single Australian record. She also presented the harvesting statistics (new and updated records) since the beginning of 2015 for several data centres (IEO, OGS, HCMR, IFREMER) with the highest number being 3419 updates of records from IFREMER since the beginning of 2015. ACB explained that the number of updates refers to every change made to existing records which may include multiple changes to a single entry.

Current work concentrates on monitoring the harvesting of records from the connected data centres. Data centres can also access their own records after harvesting to check if the content is correct (e.g. track charts, etc.). All harvested records are also available on the

BSH GeoNetwork website (http://seadata.bsh.de/geonetwork-sdn/srv/ger/find) and can be downloaded in ISO19139 format.

The next steps for the ODIP 2 prototype development task include:

a) connecting more partners to the CSR harvesting system. The next candidates are the Belgian Marine Data Centre (BMDC), BODC (UK), Marine Institute (Ireland), and possibly R2R (USA) and an Australian institution.

b) improve quality control procedures for the harvested records

AM asked if all the new CSRs have track line geometry. ACB replied that all CSRs submitted in the new format can provide the cruise track in GML. RL commented that BODC is planning to improve its system and put geometry onto CDIs but not into the CSRs as yet.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16153

4.4.2 Partnership for Observation of the Global Oceans (POGO)

Lesley Rickards (BODC) presented the Partnership for Observation of the Global Oceans (POGO) which is a consortium of major oceanographic institutes from around the world. The Directors of these institutions and other senior officials meet annually to discuss issues of mutual concern or interest and where necessary plan appropriate action. POGO is supported by a subscription fee for members and grants from charitable foundations. Its goal is to promote the completion of a sustained, integrated, global system of ocean observations for the benefit of society. There are 33 member organizations in 19 countries (with some notable exceptions including Canada, New Zealand, and most African countries). POGO is a high level coordinating activity that aims to avoid duplicating the efforts of other initiatives.

The POGO partnership exists to promote ocean observations and improve scientific knowledge. It also aims to interpret scientific results for policy makers, enhance public awareness of oceanic issues and provide capacity building through training and technology transfer.

To optimize flexibility and provide links to the research community POGO has three core pillars which are: 1) promoting ocean observations, 2) capacity building and 3) influencing policy. LR noted that promoting ocean observations also requires enhanced availability of the data collected. This needs to be on a global scale and as quickly as possible which is where ODIP II can contribute.

LR also outlined the advocacy role that POGO plays in supporting various observing systems by linking with other partners (GOOS, SCOR, GODAE, etc.) and taking a leadership role for interactions with global systems such as GEO. POGO has also formed the Oceans United which brings together many of the international organisations concerned with oceans to speak with a common voice (http://www.oceans-united.org/).

LR continued by introducing the Blue Planet initiative. POGO is a participating organization in GEO and in 2011 led the creation of the Task “Oceans and Society: Blue Planet”, for inclusion in the 2012-2015 GEO Work Plan. POGO continues to be the lead organisation and point of contact for Blue Planet. It has submitted a proposal for inclusion of Blue Planet in the next (2016) GEO Work Programme.

Following the 2015 Blue Planet Symposium in Cairns a new Blue Planet vision and mission were adopted with the aim of ensuring that society recognizes the importance of the ocean and is committed to its future stewardship for the benefit of society.

Blue Planet brings together many ocean, coastal and inland water observation organisations and programmes with the aim of adding value to existing work rather than duplicating it. LR explained the importance of the flow of information from sustained ocean observations through data collection, data and information management and models into products and services. These are then used in a variety of applications such as climate forecasting, tsunami warnings, habitat monitoring etc. that are of direct benefit to society.

The Blue Planet initiative comprises a number of different components e.g. C1 Developing capacity and societal awareness, C2 Sustained ocean observations, C3 Data access and visualization etc. It is recognized that ODIP II could potentially take a lead role in component C3. JH who is directly involved with Blue Planet and the representative of C3, noted that data access and visualization is something new for ODIP II. He also emphasized that Blue Planet does not want to recreate or duplicate effort but build on existing activities and bring groups together for the benefit of the society. HG noted that ODIP was presented by Roger Proctor (IMOS) at the Blue Planet Symposium which took place in Cairns during May 2015.

LR continued by describing the POGO Cruise Information Database for managing cruise summary reports that has been operational since 2007. Initially it was funded by the Census of Marine Life and NOAA but this is no longer the case. There are now funds from the EU-funded Eurofleets project for the Cruise Programme for Europe but with improved efficiency this could be expanded to the rest of the world. The concept behind the Cruise Programme Database is to save costs through increased efficiency and for people to work more closely together, for example, if another organization is planning the deployment of a research vessel in your area of interest instead of running your own cruise you join their ship. The Cruise Programme Database also aims to provide details of legacy cruises through direct links with the CSR database maintained by BSH. The main challenge is obtaining the legacy information to populate the database. In Europe this is less of a problem due to project funding but this is not necessarily the case elsewhere. Geographical information is the most difficult to get into the system as is non-public data such as that for some US research vessels. There are currently 2965 cruise programmes from 20 countries and 60 ocean going vessels (more the 60 meters long) from 2007 onwards in the database. In addition to the CSR database maintained by BSH there is also a link to the Cruise Vessels Database run by EurOCEAN. LR stressed that the earlier information is inserted into the system and the greater the number of countries that join, then the more useful the Cruise Programme Database would become.

LR concluded her presentation by presenting another aspect of POGO that could potentially be an area of collaboration with ODIP II. POGO is interesting in helping to improve/facilitate access to existing data repositories, particularly for time-series data. AWI is leading this effort and creating a WebGIS that includes time-series data locations, metadata and links to web repositories. This includes a range of layers and sub-layers to visualize and sort stations, for example, by geography, parameters measured, length of the time series, etc.

LR commented that the key areas for collaboration between POGO and ODIP II are the existing work on CSRs, future data related activities such as model data workflows, big data, etc. and links with the Blue Planet initiative (which was confirmed by JH)

JP took the opportunity to inform the meeting about the Oceanic Engineering Society Symposium that takes place in Monterey, California, USA during September 2016. DS also noted that Blue Planet is a GEOSS initiative and there should therefore be a future EU call for proposals relating to this initiative.

FN commented that Blue Planet looks like a blue print for IODE and its future. He also added that in Germany they use Cruise Planning as a data tracking system. Chief scientists use CSRs to report both planned and completed research cruises but this is not the case in other parts of the world where scientists are not so keen to reveal where they are going due to the risk of piracy in some areas.

TL enquired about the relationship between POGO and GO-SHIP. LR replied that GO-SHIP is for particular sectors of ocean research and is a subset of POGO.

SM enquired about the tool used to populate the Cruise Information database. DS replied that this is a simple exercise using the SeaDataNet MIKADO tool.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16154

4.4.3 ODIP 2 report on impacts assessment

Following the same methodology as that used for the ODIP Prototype 1 during the morning session, TL reviewed the impact analysis results for the ODIP 2 prototype. The target is to populate the POGO portal with CSRs from the regional systems (i.e. those from the USA, Australia and Europe). The European system is already connected to POGO through BSH, but work still needs to be done to connect the US and Australian systems. Kim Finney has identified and drafted a demonstration use case for the Southern Ocean Observing System (SOOS) to improve access to information on research and monitoring that is already taking place around the Antarctic coastline and in the Southern Ocean.

(The presentation by SM documented in section 4.12 illustrates what is available in the SOOS Field Project Portals in more detail.)

For the purposes of assessing the level of success of the ODIP 2 prototype development task the number of cruises from each regional system available in POGO as a result of the ODIP projects have been used as performance indicators. At the end of the first phase of ODIP there were 7250 European, 1229 US and 10 Australian cruises in POGO. It should be noted that this is a decrease on the number previously reported for Europe because some ICES and BODC records in the old format had not previously been checked for duplicates but these have since been removed.

The implications of adopting the ODIP 2 solution for the regional data systems, identified during the workshops and reported in the related deliverable *D4.2 Final strategic analysis report* from the previous ODIP project are:

a) the controlled vocabularies referenced by the standards and profiles that are used in the transmission of the CSRs, particularly ISO19139, should become more mature through wider use of gmx:anchor linked data and a suitable version of GML. Furthermore, the ISO19155-1 metadata standard should be upgraded to ISO19115-2 to assist with the addition of digital object identifiers (DOIs) to CSR records. Either the next phase of SeaDataNet or POGO should be able to address this change of format.

b) an upgrade of the POGO interface at the regional level is also desirable as part of the ODIP II project. In the USA two interfaces have to be maintained, one for the national services and one for POGO. In the case of Australia there is a need to federate additional institutions including CSIRO for the purposes of cruise summary reporting.

c) there is also a general requirement for POGO to include vessels less than 60m long in the database. LR indicated that POGO is already doing this for European vessels. RL commented that it is easy to include these smaller ships, where there is an agreement with POGO, by simply adding the list of ship names to the POGO lists. BSH can then easily add the CSRs for these smaller ships to the POGO portal.

To implement this change would only require an email with the list of ships to be sent to BODC. DS also suggested that the Eurofleets research vessel database could be included. This would also enlarge the scope of POGO as this database includes additional information such as ship capabilities, mass, etc.

BA noted that changing from the ISO19115-1 standard potentially has implications for the ODIP 2 prototype solution. RL commented that ISO 19115 is a content standard not a schema, ISO19115-1 is a list of fields and element names. He also pointed out that ISO 19139 is an implementation of ISO19115 as an XML schema based on GML, an example of which has already been implemented by SeaDataNet. The meeting then discussed these changes including the extensions and next generation of the relevant ISO standards.

A critical question for ODIP II is where this evolution of the standards will end. To date usage and provenance information have been described using only discovery metadata and not usage (technical) metadata. The cruise summary report (CSR) schema does not include this information but it is covered by O&M, SensorML and the new versions of 19115. It was noted that ISO is being upgraded and this will have implications for the whole series of standards. It will also require the CSR to be upgraded as it makes use of the relevant ISO standards. The main objective is not to create different standards but to follow common pathways to avoid interoperability problems in the future.

TL concluded this session by inviting the ODIP II partners to highlight other potentially useful demonstration use cases.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16155

4.4.4 Discussion

FN summarized the discussions about the future of CSRs so far which includes additional partners, federated systems and more input from the USA, especially NOAA. He invited partners to provide more ideas about the future of CSRs and the role of the ODIP II project in formalizing these advances. FN commented that in the future CSRs will include both details of the cruise and descriptions of how the measurements and chemical analyses were carried out.

BA suggested linking the cruise to the data at the granule level (e.g. using the CDI or other granule). If the data are included in the discovery system, the CSR could be the link to the data. It was noted that IFREMER has already started on work on this and other SeaDataNet partners will follow this approach in the future. R2R is also using the cruise as an intermediate step for researchers to discover the associated data. Conversely, the European CDI provides details of the associated cruise.

FN asked how close the systems operated by the ODIP partners are to automatically generating and adding new information to the CSRs. The goal for the future is that the chief scientists will not have to fill in any CSR forms. However, as discussed at the last SeaDataNet plenary meeting, collaboration with the manufacturers of shipboard systems is needed in order to automatically generate CSRs during a cruise. BA commented that this is already done for the environmental sensors in R2R. Partners in Spain and Italy (OGS) are also doing this through the EU-funded Eurofleets project.

DS commented that this discussion pointed towards a potential task addressing different aspects of how to make CSRs more efficient such as how to include more data, ships and operators and also mechanisms for accessing this data. DS also highlighted the role of the POGO and Blue Planet initiatives in transforming data into knowledge and information. DS suggested that rather than following the previous approach a use case could be developed which demonstrates how the various available tools and data can be used to generate information and knowledge for a specific area of the ocean. LW noted that there are many activities in the coastal zone that do not generate CSRs and would therefore not be included in this scenario. There is a need to incorporate measurements from platforms other than cruises and to define what is being observed and/or measured.

TL highlighted the fact that information about small experiments at sea without CSRs does not reach the data centres and work is needed to address this that issue. RL commented that standards developed for CSRs are being extended to other platforms. For example, ICES started with ships but now covers a wide range of platforms. He suggested a metadata standard is needed to describe the deployment of a data collection platform. FN mentioned that the Indian Ocean Experiment 2015-2016 could be a use case for incorporating the use of many different platforms.

FN closed the session by informing the meeting that it has been agreed that the Canadians will adopt the SeaDataNet infrastructure in their system. As a result, a large number of additional cruises will be added to the CSR system.

4.5 SESSION 4 - ODIP Prototype Development Task 3: plenary

4.5.1 ODIP 3: aims, activities and progress

Jonathan Hodge (CSIRO), leader of the ODIP 3 prototype development task, introduced this activity and described the latest outputs. This development task has evolved into a number of different experiments on sensor observation systems (SOS) and OGC compliant sensor web enablement (SWE) services, for example, OGC services for performance time series etc. by different organizations. A key element of this discussion was the way forward for this task in the ODIP II project.

Sense OCEAN project

Alexandra Kokkinaki (BODC) presented the developments in the FP7 SenseOCEAN project for retrieving biogeochemical data from ocean sensors in a standardized format (http://www.senseocean.eu/).

AK stated that autonomous ocean observation is massively increasing the number of sensors in the ocean. The best practices for data management need to evolve to ensure that key metadata and technical data from these novel sensors are not lost, and the data are efficiently processed, archived and delivered in a seamless way. In order achieve this there is a need for interoperability and a pre-requisite for this is the ability to apply standards from the sensor through to data delivery.

A problem is that sensors are attached to legacy platforms that cannot transmit OGC SWE formats such as SensorML. A solution could be for the sensor to transmit a unique identifier (UUID) that references a NERC linked-data (RDF, SPARQL) server which provides a SensorML or JSONLD description for the sensor and the standards, e.g. NetCDF, for the file format which describes how it should be queried, visualized and also the standardized ontologies and languages used. To test this approach, BODC asked the equipment manufacturers to provide samples of the sensor data which they then evaluated and modelled.

AK described how this modelling of the sensor data re-used existing expertise from the linked-data and semantic web community including the SSN (Semantic Sensor Network) ontology which has been developed by the W3C Semantic Sensor Networks Incubator group (the SSN-XG) in which Simon Cox is a member. The SNN group has worked on an OWL ontology to describe the capabilities and properties of sensors, the act of sensing and the resulting observations. BODC has extended the existing SSN ontology for the sensor descriptions as part of the SenseOCEAN project.

The example given by AK is that of a wind sensor. It uses the Library for Quantity Kinds and Units (http://www.w3.org/2005/Incubator/ssn/ssnx/qu/qu), the Dublin Core ontology, and the Good Relations ontology to describe serial numbers, manufacturers, makes and models. Use was also made of the BODC P01 Parameter Usage Vocabulary, P06 BODC Data Storage units and C75 vocabulary for Organizations.

The ontology design was based on the data using the System class, the SensingDevice class, the MeasurementCapability, the OperatingRange and the Sensing classes. Subclasses were created beneath the System Class for each type of Sensing Device that was identified from the sample data. AK continued by explaining the model, the URI design and the RDF content. The final step in the process is to publish to a SPARQL endpoint, RESTful interface, ELDA, or Mash-up application.

Tasks for the future are: RESTFul Publication; metadata publication in VoID, PROV; effective discovery (CKAN); alignment with PROV-O ontology, links with O&M; produce sensor descriptions in SensorML/JSON LD; create persistent identifiers (pURL).

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16156

Sensor Web Enablement integration

Thomas Loubrieu (IFREMER) presented the SWE activities in SeaDataNet. The main tasks were to provide a graphical editor for observation systems and to demonstrate it using a 52°North application. The editor is a flexible web application that includes a drawing tool, Draw My Sensor, that allows the data providers/scientists who are making the field observations to describe their observation system. The tool includes preloaded information about the sensor models which is also extensible. The sensor model descriptions are extracted in SensorML format from the EMSO sensor model directory (also known as the FixO[3] yellow pages, http://www.the .com/). Descriptions of the sensor models are provided as “sml:typeOf” information within the sensor instance description.

The user can drag and drop sensor or platform model icons to create instances and link them together to describe more complex systems. Links are oriented and indicate the input/output relationship. They also show the type of connection which may be wired or not (e.g. acoustic). Some sensor properties can be edited such as name, description, identifiers and properties, outputs parameters, location, contact, and events which are included as free text information. Vocabularies are referenced using URIs (linked data). The system also provides functionality for auto-completion of some fields. The output from the tool can be exported either as SensorML or as a report in .pdf or .doc format. TL suggested that an interesting next step would be to work with BODC to look at exporting in JSONLD. The demo version of the Draw My Sensor tool is available online at: http://snanny.ifremer.fr//webgraphiceditorDemo/.

TL continued by giving an overview of the demonstration that was first presented during the SeaDataNet plenary meeting that illustrated the deployment of two 52°North servers at OGS and IFREMER. These services implement the SOS GetObservation operator and provide the response in O&M format using the 52°North SOS server. A test with the SensorML output format was not performed. URIs were added to the CDI records that point to the 52°North Sensor Web REST-API.

TL described the issues and benefits for the integration of the 52°North SOS service into the SeaDataNet CDI service. He described how the REST-API URLs are transformed to a human readable format in the CDI portal so that the user can access the time-series visualizations.

Going forward there is a requirement to:

- improve the user experience,
- extend the demonstrator to include vertical profile and trajectory data, sensor and system descriptions
- move back to the core SWE standards

Much of this work is already being conducted in a number of on-going initiatives including: Eurofleets2, JERICO-NEXT, NeXOS, FixO^3, ATLANTOS, EMSO-DEV; SeaDataNet etc.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16158

FME interface for populating SOS

Rob van Ede (TNO) presented a transactional FME process for populating SOS servers with grain size distribution data from the TNO database. Using the transactional SOS is not trivial or easy because there is not much software available, much of it is still experimental, and most requires extensive configuration.

FME is a data loading and manipulation tool (ETL tool) for translating data between several different formats and doing geographical manipulation across them. It supports more than 300 formats which is useful where data transformations are needed. FME also allows manipulation and restructuring of content, assists in identifying and rectifying quality issues and no coding is required. However, SOS is not supported in the standard version of FME. RvE suggested that this might be something for future versions.

RvE then showed FME including the flexible input/output and translation tools, and also explained the methodology that he has developed. The first step was to map the data to the NVS vocabularies. JSON requests were then generated from the Oracle database and posted to the SOS. RvE demonstrated that for each JSON request posted there was a corresponding response. He showed an example of a visual response and an overview of the dataflow.

RvE concluded by describing the next activities which include cleaning up the workbench and publishing it as a (free) custom transformer (supporting the InsertSensor and InsertObservation operations) in the FME store. Future work includes implementation of more requests and data retrieval because a lot of work is currently required to visualize the SOS requests with GIS visualization packages.

AL commented that there are R packages where the visualization of SOS requests is very straightforward and can be used with the 52°North SOS implementation.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16159

IOOS SOS Activities

Shane St Clair (Axiom Data Science) presented the work done on SOS by IOOS in the USA over the last three years. IOOS is a US federal/regional partnership for ocean data monitoring in the USA that aims to enhance, organize, analyze, and provide access to ocean data and tools. IOOS is the federal parent organization with 11 regional associations (RAs) for specialized issues in coastal areas around the USA.

Prior to 2012 the OOSTethys project made an initial attempt to adopt OGC SOS standards. There was some progress made but adoption was scattered with various implementations that had differences in behaviour/responses. In 2012 a meeting took place between all of the regional associations to develop a formal IOOS SOS application profile and software implementations. During this meeting it was decided to: develop templates for SOS responses, use CF 1.6 standard's sampling geometries (time series, profile, trajectory, etc.), use defined semantic vocabularies (CF parameters, IOOS agencies, etc.), include the notion of nested assets (network/platform/sensor), develop SOS software implementations depending on the use case, and develop sensor harvesting and testing software tools.

As a result of this meeting, a SOS application profile was developed between 2012 and 2014. Progress was slow following the initial meeting due to complex requirements (nested assets, feature types, etc.) and certain problems were only discovered after implementations were developed. As a result, version v1.0 of the application profile was eventually finalized in 2014. This profile included response templates for GetCapabilities and DescribeSensor for network and station asset types, and GetObservation for time series and time series profiles. It also used the standardized vocabularies (CF via MMI, IOOS, etc.). The profile included OGC CITE style test descriptions and the WSDD (web service description document). All documentation is available on GitHub at: http://ioos.github.io/sos-guidelines/.

SSC continued by outlining the Axiom Data Science SOS efforts that led to the IOOS i52n-sos service which is a software implementation of IOOS SOS application profile v1.0. It is a web-based Java application with a database backend that uses the 52°North SOS 4.x as the core. It supports multiple database management systems (PostgreSQL/PostGIS, SQL Server, Oracle, HSQL) and SOS versions 1.0 and 2.0. The i52n-sos service also includes an installer and administrator web interface, and support for multiple bindings (KVP, XML, SOAP, REST, JSON, etc.) with pluggable bindings, encoders, etc. On top of this core there are custom IOOS modules including custom response encodings, a test data generator, and a NetCDF encoder for generating CF feature type NetCDF files (now ported to upstream 52°North SOS). It is aligned with 52°North's rewrite for 4.x and there is a lot of collaboration/contribution for hierarchical assets/procedures, performance improvements, and simple transactional operation security. This implementation is well suited for harvesting/serving active sensor streams, and is available at: http://ioos.github.io/i52°North-sos.

Another IOOS application developed by RPS ASA (Applied Science Associates) is ncSOS. It is a UNIDATA THREDDS plugin to serve NetCDF sensor data via IOOS SOS v1.0. It supports the core SOS 1.0 KVP operations (GetCapabilities, DescribeSensor, GetObservation). However, due to THREDDS internal functionality, there is one SOS server per NetCDF file (station or sensor). This implementation is well suited for sensor data in NetCDF files (often archives or post-processed), especially large time-series.

In addition to these two SOS implementations, a suite of additional tools has also been developed by IOOS including:

a) the sos-injector: a Java library to insert data into SOS via OGC transactional operations;

b) the sensor-web-harvester: a Scala application to harvest data from web sources and inject into SOS servers using the sos-injector (many sources in the US including NDBC, CO-OPS, USGS, etc.);

c) the sos-injector-db: to inject data into SOS from an existing database,

d) the ioos-sos-compliance test: OGC CITE (team engine) tests for IOOS SOS v1.0 implementations

e) compliance-checker: a Python tool to check datasets (NetCDF, SOS) against standards (CF, ACDD, IOOS, etc.).

The current status of IOOS SOS efforts is that all 11 regional associations serve sensor data through one or both of the IOOS SOS v1.0 implementations (i52N-sos and/or ncSOS). Implementations provide equivalent behaviour/responses where standards dictate. SOS servers are registered in the IOOS catalog (http://catalog.ioos.us/).

Beyond SOS there is also the Sensor Scalability Experiment (http://axiomdatascience.com/maps/ioos) that aims to harvest "all" available sensor data and build a system to handle large data volumes. It will use hex binning to show trends at high zoom levels and hold around 90 million observations over 14 days in memory. The system will provide high performance statistical binning and analysis. The AOOS demo is available at: http://goo.gl/pFpRAR.

SSC concluded his presentation with the lessons learned from the IOOS experiences. These are: 1) prioritize client library development because a lack of these inhibits adoption; 2) provide a simple data format option (many people want CSV, which can still co-exist with SOS); 3) use a small group to propose initial drafts of standards because this will speed up the process. The wider community can then adjust or reject) the proposal; and 4) develop pragmatically (release software with basic functionality early, add new features prioritized by user demand and do not wait for complete implementation because software benefits most from being used!)

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16160

SensorCloud

Jonathan Hodge (CSIRO) gave a short update on the SensorCloud system which is a times series data aggregation, ingestion and delivery system. It uses a Java Message Service middleware API built on top of the configured data sources and signals. It is not a SOS implementation as such, but it is guided by some OGC standards such as O&M and used a simplistic mechanism for describing sensors (StarFL).

SensorCloud entries come from multiple sources with different data types and originating from various moving platforms (e.g. sensors on animals, ships, etc.). It allows data collection from sensors without the requirement for a full SensorML document. It also holds information on deployments and can store tracks using the identifiers and serial numbers of specific instruments which is useful for QA/QC.

There have been some developments for the location element with the concept of RelativeLocation: a platform can have a location while the individual sensors can have relative locations (e.g. the anemometer is mounted 10m above the platform).

The sensor calibration information is provided either from conformance testing or through field sensor calibration which is recorded as a calibration event. There are some changes in the structure of streamed data to supplement the actual data streams. New data types are: GeoLocation (e.g. GPS sensors); Scalar value (numerical value, e.g. Temperature); Vector value (array of values, e.g. spectrometer, depth profile); Sequence value (high resolution, e.g. audio for specific use case such as automated species detection of frogs etc.); Image value.

SenorCloud uses MongoDB, a free and open-source cross-platform database solution, for data storage. It provides: document storage, flexibility in data types, GridFS, distributed storage for files (images), an aggregation framework (distributed computation of aggregations) and geospatial indexes and queries (for mobile sensors). It is also easy to horizontally scale and there is experience of using it within the CSIRO development team.

Using Mongo DB v1 with 1.1 billion observations showed that scalar data is small in volume but the index size is relatively large in comparison. It also became evident that the nodes use a lot of RAM and the performance degrades as the data/ram ratio increases. MongoDB Version 2 offers: aggregated storage, multiple observations in a single 'document' and smaller indexes for high frequency (faster than hourly) data streams. MongoDB aggregation pipeline unwinds results, and provides improved sharding to balance data streams across nodes in a cluster.

JH continues by explaining some of the technical aspects of the data ingestion system and the Sensor Messaging Gateway (SMG). The data sources for the system can be a configurable generic polled file import (CSV, TSV, fixed width, FTP, HTTP); the existing library of data sources (Campbell Scientific; Libelium; PACP (DPF/CSIRO); AgIsp (DPF/CSIRO); ROS (Robotic platforms); BoM/SILO) or custom data sources (Java, Python). JH then provided an interesting example from the data perspective: a web JavaScript streaming application using a STOMP interface in a RabbitMQ system which serves oyster heart beats in real time (20 points/sec (Hertz).

All SensorCloud metadata and data streams belong to one or more groups. The user 'roles' define the permissions for a specific group. For example, John works on a project called Aquaculture that has its own private sensors. The researchers on that project need access to this private sensor data. John has the role 'aqua_researcher' which allows him read access to all data in the 'public' group and also data in the 'private_aqua' group.

The system now can handle model workflows and provides raw sensor data by references. The model services provide a framework to 'wrap' existing models as web services and describe the model inputs and outputs in a REST API. It currently supports Keplar, R, Python and Java. The gridded data services use THREDDS for data management of netCDF data sets, the provision of a catalogue and for web services (WMS, WCS, OpenDAP, NetCDFSubsetService, and HTTP).

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16161

Alfred-Wegener-Institute – ODIP 3 prototype development task

Ana Macario (AWI) presented the AWI activities relevant to the ODIP 3 prototype development task. She explained that the Computing and Data Centre group has traditionally developed different information systems for data acquisition and especially on board of the RV Polarstern. Only recently has there been a systematic effort started to adopt OGC standards relevant to ODIP 3 as a means for supporting the automated data flow from sensors to PANGAEA. These are mostly focused on devices and sensors on board the two main German research vessels RV Polarstern and RV Heincke as well as the land-station Neumayer. They are also trying to cover more exotic platforms such as sea-bottom crawlers, drones, etc. In terms of sensor characterization, they have developed a web client for describing the platforms, devices and sensors. The SeaDataNet SensorML profile has been adopted and extended with AWI-specific metadata for the needs of the Institute. There are currently about 100 ship-mounted sensors and around 500 sensors from other platforms that give approximately 10 million ($10x10^6$) measurements per year.

All sensor metadata is being stored in a near-real time database (PostgreSQL) which can produce SensorML 2.0. A series of REST-based access interfaces have also been developed to make the information of interest to users available for dissemination. In the case of monitoring equipment, it is important to keep the range of each sensor stored in the database. The AWI SensorML profile therefore includes range values for each sensor which is used in automated QA/QC procedures (e.g. measurements out of expected range are flagged) and in monitoring dashboards.

AM then showed an example of a monitoring dashboard to illustrate the practical use of SensorML. In this case it is used to trigger an alert when the value for temperature falls within a specified range. Access to near real time data is currently via web services (JSON encoding, tab-delimited) with track line geometry and atomic dissemination for selected sensors. Support for O&M standard is planned for the future.

The activities planned for ODIP II include:

a) adoption of NVS vocabularies for parameters (not a trivial task for PANGAEA) and EDMO manufacturer vocabularies (not yet available),

b) share the AWI SensorML 2.0 profiles in a GitHub repository,

c) implement O&M encoding for selected sensors.

These activities will also be related to those in other initiatives such as Eurofleets II, $FixO^3$, AtlantOS etc. There will also be a close cooperation with 52°North to support the installation of the SOS 4.x core server including the Sensor Web REST-API extension.

Of particular interest to AWI is the integration of their legacy data infrastructure and information systems. A series of “connectors” need to be developed due to the complexity of the data.

AWI also have an interest in using the O&M encodings and SOS not only for near real-time data, but also for disseminating their legacy data via PANGAEA. This will be a challenge and by no means a trivial task when they expect to receive around 10 million near real-time measurements per year and approximately 10 billion measurements from PANGAEA.

JH commented that there is clearly a lot of relevant activity within several projects that could contribute to the ODIP 3 prototype. There is therefore a need to discuss how to proceed with this task in ODIP II.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16162

4.5.2 ODIP 3 impacts assessment report

TL (IFREMER) reminded everyone that the aim of ODIP 3 is to bring together SWE resources from the three regions into a single portal. Using the readiness levels as defined in the *Framework for Ocean Observing* (2012) referenced by TL earlier in the workshop, this prototype is currently at the concept stage with regional pilots currently underway. There are many results from the different implementations including systems based on 52°North, ncSOS, oceanotron, sensorCloud etc.

Three suitable standards were identified for the conceptual approach to describing both the observations and the observing systems. The implementation of a RESTful JSON approach was considered to be the most efficient solution and would facilitate the development of web clients that could be used on top of these services. A demonstration use case combining RESTful JSON services from different servers (e.g. 52ºNorth, ncSOS, sensorCloud, oceanotron) into a single decision support system has been drafted but has not yet been implemented.

The performance indicators for the ODIP 3 prototype are dependent on the type of SOS service being considered. There are potentially two different approaches, the first is to have one ncSOS service per file (or per single granule) and the second, which is currently being developed by 52°North, oceanotron and sensorCloud, is to have SOS services at the collection level for data sets.

The implications identified for the ODIP 3 prototype are the need for standardization of the profiles for the observing systems and the observations themselves, and specifically the RESTful API profiles. Work needs to be done at the regional level on the implementation of these standards, but there seems to have already been significant progress on this since the previous workshop.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16163

4.5.3 Discussion

The discussion session focused on the way forward for the ODIP 3 prototype development task including potential areas of expansion within the ODIP II project.

JH questioned the potential implications of the various on-going SWE activities for the ODIP 3 prototype and, in particular, how to bring them together. The options are to either continue the informal approach currently adopted where the relevant groups meet and present their activities in an effort to learn from each other, or try to construct something more formal with targeted efforts. The challenge for the Australian partners is funding for such an activity. One potential option might be to develop an Asia-Pacific-USA project or similar ODIP II focused activity and then try to identify funding to support it.

DS suggested that a report on the SWE activities undertaken during the first ODIP project should be produced that describes the significant progress that has already been made on developments in this area by the regional systems.

It was noted that the initial ambitions of ODIP in the area were much greater at the start of the project but the objectives are now much clearer as SWE/SOS systems are evolving to become more mature. The addition of new partners in ODIP II that are directly involved with the development of these technologies, for example 52°North, has also helped to clarify the aims of the project with regards to sensor web enablement.

It was agreed that the aim should now be to define a suitable demonstration use case to show how the various SOS services potentially interoperate including any areas of significant overlap. One option might also be to make a formal approach for funding of this use case to an initiative such as Blue Planet.

HG supported the idea of a use case as this would be a good outcome for the ODIP II project. Several potential use cases were then discussed that included an assessment of different SOS services, definition of sensor descriptions or working with manufactures to encourage “buy-in” to the work being done on SWE in ODIP II.

Another issue that was discussed was the leadership of the ODIP 3 prototype development task in ODIP II. This activity was led by the Australian partners in the previous phase of the project but funding issues have necessitated that responsibility for this task passes to

52ºNorth. HG expressed concern that all three existing prototype development tasks are now being led by the European ODIP II partners. The rational for the leadership of the three tasks was that there was a balance across the participating regions. It was agreed that this balance needed to be redressed as the project evolves.

4.6 DAY 2: SESSION 5 – ODIP prototype development tasks: feedback on outcomes and possible next steps

4.6.1 ODIP prototype development tasks

Each group responsible for the development of the prototype solutions provided feedback on the final outcomes from their respective sessions and outlined potential further developments in ODIP II.

ODIP 1

DS noted that there is still some work to be done in ODIP 1 including: a) defining separate name spaces for the three regional data systems (SeaDataNet, US-NODC and AODN), and b) checking the number of records that are being harvested from the individual services against those being exposed in the global systems.

DS then summarized the issues that came out of the previous day's discussions regarding possible next steps for the continuation of the ODIP 1 prototype development task which were:

- Making the ODIP 1 solution fully operational as it is currently more of a demonstrator. The solution needs to be more dynamic so that any changes at the source (the regional data systems) are propagated to the other regional and global systems.
- The three regional data systems are currently using their own vocabularies which are harvested and pushed to the GEOSS and ODP in their existing state. An important task is to check the harmonization of these vocabularies especially in the brokerage. One possibility is the addition of semantic brokerage with ontologies put in place between the three systems, but this necessitates cooperation between those involved with the development and maintenance of the vocabularies, and the broker technologies.
- Explore data brokerage as there is currently only brokerage at the metadata level. Checks on the progress of existing projects addressing this topic should be made including identifying potential opportunities for collaboration between ODIP II and these initiatives. The ODIP II project can identify the actions that need to be taken but the nature of the project requires some of the work to be undertaken by the other initiatives.
- Develop a better understanding of user requirements by evaluating existing use cases and making an assessment of the impact made by the current consortium including the potential benefits it can provide to the wider community. It is already clear that users are increasingly more interested in aggregated data services and added-value services rather than basic data discovery and access.
- The regional discovery services currently provide access to data but they are still autonomous systems. There is a need to provide users with more information about the data being provided e.g. provenance.

- Metadata from the regional data systems is now being harvested automatically but manual intervention is still required to check if the harvesting has been successful. In many cases the results have been found to be unsatisfactory.
- The federated search mechanism is another possible approach that could be considered. Rather than bringing the metadata from each regional data system into a single discovery service results are returned directly from the three services. This approach is completely different to that used by GEOSS and the broker but this alternative approach will be looked at as a possible solution.

The workshop participants were then invited to provide further suggestions for the next steps in the ODIP 1 prototype development task.

It was observed that data quality is an issue when combining data from different systems. The H2020 AtlantOS project could be used as a use case where the data will be used both for both scientific and operational purposes (feeding into Copernicus, US and Canadian models). It will also make an assessment of the added-value of integrating the various resources and, in particular, in response to the needs of the user.

The issue of semantic interoperability and how it could be approached (for example in case of parameters) should be considered: each region uses different vocabulary lists, for example, SeaDataNet uses P02 (NVS), AODN uses a mix of BODC and Australian vocabularies, and US-NODC also uses a different one. To achieve semantic interoperability, the metadata from the regional systems must carry URLs for the parameters being used. These URLs resolve to a RDF document and then, through translation from one list to another, will allow searches across multiple systems to be carried out without the need for a common vocabulary. The client will select the appropriate list and semantics will point to the data.

Additional vocabularies such as those needed for biological oceanography e.g.GBIF, WoRMS etc. can also be included in the model by mapping to the SeaDataNet P01 BODC Parameter Usage vocabulary and extending the data model. Once the required vocabularies have been identified, the next step is to mark them up using the core metadata documentation.

The Ocean Data Portal (ODP) is also moving towards the same approach of using URLs but the ODP technical group will need assistance for the content mapping which can be provided by the ODIP community.

GEOSS is also in the process of addressing the semantics and could focus on coastal marine applications. However, GEOSS does not do any content harmonization for the broker. In coastal areas the biology community is quite different to the water community and GEOSS could potentially focus on this aspect. This cross-domain homogenization would be a big step for ODIPII. It was suggested that biogeochemistry in estuaries could be a possible use case, but it is not yet clear if ODIP II should extend outside the marine domain or even if this would be feasible.

Discussion during the workshop then focused on user needs and specifically searching for data not only at collection level as currently delivered by the global portals but also at the granule level. It was acknowledged that this is still difficult to realize when working with federated systems and metadata collection (and not point) searches.

Creating a link between the European Directory of Marine Environmental Data (EDMED) and CDI references could make EDMED a tool for discovery at the collection level. The user could define a polygon of interest which returns a list of the datasets identified in EDMED that resolve to individual CDI records. The CDI records include a geometry which can be used for filtering on an area of interest. DS concluded the discussion by suggesting that ODIP II should not create unrealistic expectations for what is achievable but that semantics is one aspect that the project can address.

ODIP 2

FN and ACB summarized the ODIP 2 discussions from the previous day. The main observation made is that the Cruise Summary Report (CSR) is an integral part of the POGO infrastructure. It has a key role to play in multidisciplinary research especially if the CSR provides an overview of the achievements of the cruise. From the data management perspective, the CSR can also be used as a powerful tool for data tracking.

Other key points were:

- A description of the CSR harvesting including minimum requirements can be found at: http://www.seadatanet.org/Standards-Software/Metadata-formats/CSR
- Manual for the CSW has been documented in the SeaDataNet deliverable SDN2_D92_WP9_CSW_harvesting.pdf
- A GeoNetwork instance has been created at: : ftp.ifremer.fr/ifremer/sismer/donnees/SeaDataNet_Software/CSW/geonetwork-sdn.war

Other more detailed feedback included requirements for:

- Connecting CSRs with the CDIs e.g. using CSR as a discovery tool for accessing data
- Links to underway sensor data
- Extension of CSR schema/standards to include O&M or to introduce next generation of standards. (It was noted that this is a large task due to way that parameters are currently handled within the CSRs)
- Extending CSRs to include other platforms, e.g. small boats, gliders, mammals, etc.
- Include more ships less than 60m in POGO
- Blue Planet Mission
    - transferring data into knowledge: use CSR in selected products to illustrate potential benefits
    - Find use case to evaluate how the tools work together

Next steps for the European partners:

- Upgrade of the POGO interface
- Linking CSRs to CDIs
- Partners in Spain and Italy will continue to work on automatic CSR generation and potentially make these tools applicable for other ships. This approach could also be documented and implemented as a standard for the next generation of research vessels. This would also potentially help with interactions between researchers and the manufacturers of the ships systems.
- Introduction of ORCIDs for researchers
- Introduce DOIs for CSRs

Next steps for US partners:

- Maintaining the CSR interface and continued population that will add a large number of additional cruise summary reports to the system.
- Adding Ocean Area (C19) parameters and more detailed abstracts
- CSW harvesting
- Submit CSRs for NOAA ships. This is an outstanding issue and contact with NOAA is needed.

Next steps for Australian partners include:

- Federation of additional CSRs from other Institutes
- CSW harvesting

Presentation available at:
http://www.ioc-unesco.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16166

FN opened the floor for further discussion on the topic of the ODIP 3 prototype development task.

It was noted that the standard format of Argo float data has header information that includes a significant proportion of the platform metadata (codes) required for the CSR, and with the proper cooperation between Argo, JCOMMOPS, Copernicus etc. cruise summary reports could be automatically created in real-time mode.

All the observing programmes use different flavours of netCDF-CF but it is unclear if there is a general agreement to extend CSRs to other platforms. The gliders community have previously requested platform codes from the CSR group at BSH. The participants at the workshop then discussed if CSRs should be reserved for research vessels and not include other platforms or is the community more concerned with using the CSRs to locate datasets. A major concern is that the cruise databases would be flooded with information from floats and gliders. To address this issue the French NODC uses the European Directory of the initial Ocean-observing Systems (EDIOS) to describe Argo floats instead of cruise summary reports.

However, a cruise can include many platforms and the data of interest can be located using a suitable filter. For example, BODC has a metadata system that is platform agnostic. They maintain platform deployment metadata that are sent to BSH who then filter it according to the platform type to determine which information is to be included in the CSR database.

It was observed that the most common request received by BSH relating to the CSR data base is still for data from a specific cruise.

The discussion concluded with a number of actions being identified that relate to user requirements for CSRs and the role that can potentially be played by the ODIP II project.

ACTION 1.8: BSH to define a generic use case to capture user requirements for cruise summary reports (CSRs)

ACTION 1.9: Develop a specific use case to drill down into the detail of these user requirements including how the user wants to find the resources they need. This use case should also include the expected role of the ODIP II project from the user perspective.

ACTION 1.10: Put a WMS-WFS on top of the CSR services to expose them to other systems including for the purposes of implementing a linked data approach.

ODIP 3

Jonathan Hodge (CSIRO) noted that a decision needed to be made on how the ODIP 3 prototype development task will progress in ODIP II. There are two potential approaches:

- Option 1: continue as before and basically function in a similar manner to that of an RDA interest group by reporting the relevant activities of the ODIP II partners and exchanging ideas in an attempt to align relevant SWE activities.
- Option 2: identify a topic that could be turned into an actual more structured prototype development task

It was agreed that there are challenges both in coordinating the different contributors to ODIP in this area of interest and also for creating funding opportunities. DS suggested that the best solution was to continue with the current approach and for the partners to try to explore potential sources of funding for a joint activity which includes initiatives such as GEOSS and relevant RDA interest groups.

JH indicated that there are currently no Australian activities related to SOS development or even projects of common interest with ODIP that could move towards it. The Australian partners could contribute to identification of commonalities between sensor outputs which include definition of O&M specifications that can be applied to different systems. However, there needs to be clarification of how to structure and coordinate such an activity. HG confirmed that the format of activities in the ODIP II project can potentially be quite flexible. DS suggested that the current activities in this field are continued whilst also seeking “low-hanging fruits” by identifying the commonalities and differences between existing SWE/SOS systems to determine if there is a worthwhile activity for the ODIP II community to undertake.

As there is currently no active SOS work in Australia RP suggested that another region should take over the coordination of this activity. As there is an acknowledged need to have a balance of responsibilities for the prototype development tasks across the three contributing regions in the ODIP II project, the possibility of re-scoping the ODIP 3 prototype development task or creating new tasks where the Australian partners could contribute e.g. sensor data, model workflows and big data were discussed. It was pointed out that there is interest in SOS and SensorML in the other regions. SSC suggested focusing on SensorML and exploring suitable interoperability formats.

Following this discussion it was agreed that the ODIP 3 prototype development task will be re-formulated and instead of focusing solely on SOS, it will have a broader focus on interoperability. Furthermore, leadership of this task will transfer to partner 52°North.

AK proposed using ontologies to map to SensorML and to expose the sensor description in RDF. The group then discussed this proposal regarding how to get the required information from sensors. It was agreed that AK will produce a summary of what currently exists and a proposal of what could be done next that can then be turned into a use case.

ACTION 1.11: AK to produce a summary document that outlines current approaches to using ontologies for mapping to SensorML which also includes suggestions for possible next steps

4.7 SESSION 6 – Vocabularies: plenary

4.7.1 NVS Developments

'One-armed bandit semantic model'

Roy Lowry (BODC) presented the recent NVS developments. The L22 instrument vocabulary has been extended to support harmonization with IMOS and R2R. As a result 89 additional concepts have been added to L22 since the last ODIP workshop. This work is largely completed with IMOS already finished and R2R only having a few minor mapping issues to be resolved.

There have also been 757 concepts added to the P01 BODC Parameter Usage Vocabulary since April 2015. In addition, 13350 P01 concepts have been marked-up with a Chemical Abstract Service (CAS) registry number, which is a unique identifier assigned to every chemical substance described in the open scientific literature.

The semantic model for P01 has been converted to a 9-wheel 'One-Armed Bandit' that currently covers chemical substances found in biota. RL explained the semantic model fields exposure to the set of the 9 one arm 'Bandit' wheels. Each measured phenomenon e.g. each P01 term, is not only concentration (e.g. a measurement) but it could be: Measurement + Substance + Measurement Matrix Relationship + Matrix + Matrix Subcomponent + Biological entity (Taxon/ITIS/WoRMS, Organism Name, Organism Specifics) + Technique.

RL indicated that the development of the Chemical Substance 'wheel' had required the most work. After checking the P01 concepts, it was determined that a good definition of a chemical substance is that it is an element, isotope, compound or mixture. For example, Chlorophyll-c is a mixture of the subtypes c1, 2 and c3, while Chlorophyll-a is a pure compound.

An ideal scenario for defining a substance wheel (Simon Cox proposal) was to find this type of resource outside the NVS. It also needed to fulfil certain criteria:

1) It has to be a comprehensive, authoritative collection of substances;
2) Each substance has a URI
3) Each URI resolves to RDF documents

RL pointed out that this was not a realistic scenario and currently unattainable. Two potential candidates were the Chemical Abstracts Service (CAS) Registry Number and the Chemical Entities of Biological Interest (ChEBI). However, both have issues that prevent them being universally adopted. In the case of CAS there are a small number of duplicates and some ambiguity. It has poor coverage of mixtures because the chemical manufacturing industry do not deal with mixing catalogues. CAS also has poor coverage of compounds of research interest that are not manufactured commercially. The issues with ChEBI include patchy coverage of some substance types such as large organic molecules, isotopes, mixtures because ChEBI tends to avoid large organic molecules, and some isotopes and mixtures (Chl c) are also absent. In addition, ChEBI has a confusing range of entities, for example, Cadmium atom, Elemental cadmium, Cadmium molecular entity, have different identifiers. Furthermore, ChEBI has a huge number of replicated IDs, for example, there are instances where three different URIs resolve to the same thing. It also contains multiple URIs that are not helpful for semantic interoperability.

The decision was therefore taken to create a Chemical Substance Wheel on the NERC vocabulary Server (NVS) as part of the BODC parameter semantic model chemical substances (S27) vocabulary. The objective being to guarantee coverage for every chemical substance in P01 (about 1200). It will also allow operational (trigger-driven) mapping to external resources such as ChEBI, CAS and eReefs to be maintained so that, for example, when a new P01 parameter is created and the substance is recorded in ChEBI, then the trigger automatically pushes the URLs into the RDF document.

The population work is underway with 191 out of 400 concepts currently included with a target of around 1000-1200 concepts. The downside of this approach is that every substance has yet another URI but this was unavoidable.

RL continued by observing that the MARIS vocabulary client has done sterling service in supporting SeaDataNet and many other projects. However, long usage has shown it to have some limitations that include: a) limited search behaviour control that can cause hit floods; b) no management of deprecated concepts; c) no ability to locate vocabularies by searching concepts; d) it is dependent on BODC Oracle back office for cache refresh; e) it does not cover all (190+) vocabularies in NVS; f) display not optimised for mobile devices. As a result of these recognized limitations of the existing service a new search client has been developed by BODC (see NVS search client below)

TL commented that there is also the INCHI identifier for chemicals entities, and enquired how this was related to those that had already been described during the presentation. RL confirmed that ChEBI also includes the INCHI identifiers. He also commented that there are several other identifiers, each having its own coverage and that it is technically feasible to extend the mapping to these if people find it useful.

DS asked if there has been any contact with Simon Cox on this topic. RL confirmed that he is not directly involved in this activity. During the previous workshop it was hoped that ChEBI would do the integration into the substance wheel but as this did not happen so it has been necessary to make progress elsewhere. However, RL suggested that if ChEBI or anyone else can provide full coverage in the future it can be plugged into the RDF to replace S27.

CC asked if Simon Cox is aware of the duplicate entries. RL replied that they have been in contact but not specifically to do with the duplicates. Their interactions have been more about the inconsistencies and broken URLs in eReefs which have now been fixed.

Presentation available at:
http://www.ioc-unesco.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16173

NVS search client

Alexandra Kokkinaki (BODC) presented the NVS2 Search Service that has been developed for searching the NVS2 Vocabulary Server (https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search/) , especially the collection URLs e.g. http://vocab.nerc.ac.uk/collection/ including inside each individual collection. The tool has been developed to help users easily locate the codes and the related vocabularies for the terms they are interested in.

The technical architecture is built on a SPARQL endpoint that includes a triple store that contains all the collections and the concepts. Two types of search are possible; 1) locating controlled vocabularies that contain a specified search term e.g. show all vocabularies that contain the term “Mytilus%edulis”; and 2) locating a search term within a particular vocabulary e.g. find the term “pH” in the P01 vocabulary. The service has been developed to accommodate both basic and advanced users.

AK concluded the presentation with a live demonstration of the Vocabulary Search Client (https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search) showing examples of simple and advanced searches, including how to narrow search criteria to individual catalogues etc. and how to sort and download the results.

It was highlighted that NVS Vocabulary Search Service is now fully operational and can also be released via the SeaDataNet web pages.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16174

NVS Linked Data demonstration

Roy Lowry (BODC) outlined his thoughts on the concepts of linked data which can be considered to be a series of links to different resources providing information that are all driven by RDF (the fundamental standard for linked data).

For example, http://vocab.nerc.ac.uk/collection/P01/current/VLZJ0092/ is the URL to a single concept within the P01 vocabulary. It looks like html but it is RDF with 5 “wheels” exposed (S06/observed phenomenon, S27/substance, S02/relationship between the substance and the matrix, S26/matrix, S25/biota) as URIs. Clicking on the S27 link provides links to other P01 ‘total iron’ concepts, and also links to ChEBI, CAS and eReefs. Similarly, clicking on the S25 link provides links to other P01 'Asteroidea' concepts, and links to WoRMS and LSID RDF. This is currently a demonstrator containing the ‘total iron in biota’ P01 concepts but work is underway to make this a fully operational system. This requires population of the S27 and S05 ‘wheel’ vocabularies, and migration of concepts to the ‘clean’ chemical substance semantic model. The target for completion (15-20,000 concepts) is the end of 2015.

RL concluded by outlining the changes to the BODC vocabulary team due to his forthcoming retirement on November 1, 2015. RL was congratulated on his contribution to the field of vocabularies and the marine community in general by everyone attending the workshop.

Use of Controlled Vocabularies by US partners

Cyndy Chandler (WHOI) began by providing a reminder of the US activities and in particular those related to vocabularies. This work is being conducted under the Rolling Deck 2 Repository (http://www.rvdata.us) and BCO-DMO (http://bco-dmo.org) projects with participation by four key institutions (LDEO, FSU, SIO, WHOI). It was noted that, as well as the vocabularies work, both R2R and BCO-DMO are also actively promoting the use of ORCIDs by their scientists.

The R2R project has a different scope to that of BCO-DMO and uses different vocabularies but both are using the same format (NVS: http://vocab.nerc.ac.uk/collection/###/current/) and reference cruise activities in the same way. Both use URIs in RDF.

Vocabularies used both by R2R and BCO-DMO for cruises:

- ICES Platform codes (for vessels) (NVS C17)
- SeaVoX Device Catalogue (L22)
- SeaVoX Platform Categories (L06)
- SeaDataNet Device Categories (L05)
- Climate and Forecast Standard Names (P07)

- Country codes: ISO 3166-1 alpha-3
- European Directory of Marine Organisations (EDMO)
- ORCIDs for person (when available)

NVS vocabularies used by R2R

- SeaDataNet measurand qualifier flags (L20)
- SeaDataNet Ports Gazetteer (C38)

Vocabularies used by BCO-DMO

- MEDATLAS Parameter Usage Vocabulary (P09)
- Climate and Forecast Standard Names (P07)
- SeaDataNet Agreed Parameter Groups (P03)
- SeaDataNet Parameter Discovery Vocab. (P02)
- BODC Parameter Usage Vocabulary (P01)

R2R is also undertaking a programme of matching and mapping between vocabularies currently in use. These vocabularies are:

- SeaVoX Sea Areas (C19)
- BODC Parameter Usage Vocabulary (P01)
- SeaDataNet Parameter Discovery Vocab. (P02)
- SeaDataNet Disciplines (P08)

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16176

4.7.2 Report on AODN and ANDS vocabulary developments

Sebastien Mancini (IMOS) gave an update on the vocabularies work being done as part of the Vocabulary and Creation Management (VOCRAM) project which started in September 2014. The eMarine Information Infrastructure (eMII) is using the VOCRAM tools to build vocabulary services which can also potentially have implications for ODIP II.

Work started 3 years ago to improve the AODN 1-2-3 portal functionality by adopting and mandating the use of some controlled vocabularies. eMII and IMOS adopted the MCP 2.0 metadata schema and built a small number of AODN vocabs to support content population. These vocabularies also re-used existing terms managed by other systems wherever possible.

Until recently creation and administration of these vocabularies was restricted to internal management by eMII. However, this is not a useful model for encouraging community participation and eMII approached ANDS for a national-in-scope project (VOCRAM).

ANDS already had some infrastructure that could be used to support vocabulary services but they were not properly integrated and the creation/editing functionality was missing. VOCRAM sourced existing vocabulary editing software (Pool Party©) to bind all of the components together to form an integrated, user-friendly vocabulary services tool suite.

SM continued by describing the Pool Party© software which is a web-based commercial tool but is provided to all research Institutes in Australia on a reduced cost academic licence (5000 AUS$/per year). Its functionalities include vocabulary management, creating new concepts, and a wiki viewer. SM also showed the associated SPARQL query endpoint and how to publish SISSVoc, etc. via the ANDS portal. It was noted that very few issues were found during the data integration process. ANDS also built an additional tool on top of the application in order to do some data cleaning before it is exported. The European EDMO directory can also be used to populate the organisations. DS commented that there are around 50 Australian EDMO entries included in USA cruises that need to be checked for duplication.

SM concluded by providing details of the ANDS Publishing Portal and the vocabularies published which include: Parameters (167 terms), Instruments (236 terms), UoM (62 terms), Platforms (324 terms) and Organisations (366 terms) which has links to EDMO entries. Most of these vocabularies also have at least one published classification scheme.

Work will be done with AODN community representatives to expand the content of the existing vocabularies and to add new ones where required. It has also been recognized that there is a need to pilot an appropriate community moderation and governance model.

As a result of this work the AODN Community can now (September 2015) discover and download (versioned) available vocabularies through the ANDS Portal, link directly to published concept definitions through URL references, and use the ANDS tool suite to create and publish their own organisation-level vocabularies.

eMII will be seeking collaborations to make sure that the organisation-level vocabularies created in the community that potentially overlap with the AODN vocabularies have appropriate mappings. However, the ideal scenario is the adoption of the AODN basic vocabularies. There is also now a better opportunity for looking at (automated) metadata mapping between the MCP 2.0 and the SeaDataNet CDI.

The session ended with a discussion on how AODN could apply a “spinning wheels” approach similar to that used by BODC that would allow scientists to do their vocabulary mapping outside the Pool Party tool.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16177

4.7.3 Report on RDA VSIG activities

Rob Thomas (BODC), attended the 6th Research Data Alliance (RDA) Plenary meeting which also took place in Paris during the previous week. RT provided a report from the Vocabulary Services Interest Group which is co-chaired by Adam Shepherd (BCO-DMO), Simon Cox (CSIRO), Stephan Zednik (Rensselear Polytechnic Institute) and Adam Leadbetter (MI).

Around 50 people attended the meeting including ODIP representation by BCO-DMO, CSIRO and BODC. The session began with three earth science/marine centric presentations to set the scene. Participants in the meeting were from diverse range of domains but had a common interest in vocabularies and associated best practices.

Potential action items for the Vocabulary Services Interest Group are:

- Establish a Wiki to list:
  - vocabularies
  - services and tooling
  - vocabularies services:
    - catalogue known practices which can be evaluated for identifying best practices
    - collect use cases for vocabulary services

- evaluate use cases against the RESTful API approach for fitness of purpose

The take away messages from the session were:

- A number of people are interested in regular virtual group meetings (outside of the RDA plenary).
  - There is general interest in identifying best practices for vocabulary services and driving towards this objective should help to identify any problems. Solving of these problems can then become the focus of the VSIG.

Additional material relating to the Vocabulary Services Interest Group:

- Link to meeting notes: http://bit.ly/1Fwa5MY
- Slides: http://bit.ly/1Fwbovk
- Presentations (Google Folder): http://bit.ly/1PxabUW
- RDA Interest Group: https://rd-alliance.org/groups/vocabulary-services-interest-group.html

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16178

4.8 DAY 3: SESSION 7 – Model workflows and big data: plenary

4.8.1 Model workflows and big data

Introduction - what is Big Data (not just volume, but other aspects too)

Adam Leadbetter (MI) began by introducing the session including the contributions that will be made by the various speakers.

The term “big data” encompasses variety of different things including the data itself, its storage and the associated technologies. The most common definitions of big data focus on the five Vs of big data: volume (the amount of data), variability and veracity of the data (quality and fitness for purpose in terms of data analytics), variety (different types of data), and velocity (speed data streams move around different systems).

AL concluded by explaining that the focus of this session is to look at these five aspects of Big Data and identify potential applications and use cases in the marine domain. LW provided favourite definition for big data which is:

“Big Data’ is really about having more data today than I had yesterday, such that I need to find and apply different ways and means of processing it to meet my funding deadlines.”

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16184

Australian perspective – what has already been achieved and more

Lesley Wyborn (NCI) presented the Australian perspective on the control of Big Data and those aspects related to volume. She explained that Big Data is a relative term where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making.

The problem is the scale of the storage, not moving the data around. When combined and integrated, the NCI collections are too large to move because bandwidth limits the capacity to move them easily and the data transfers are too slow and expensive. LW noted that even if the NCI data can be moved into the public domain, few can organizations can afford to store 10 PB of data on a spinning disk.

A change of focus was therefore needed which involves moving users and/or processing to the data and having online applications to process the data in-situ. This called for a new form of system design where storage and computation are co-located, and systems are programmed and operated to allow users to interactively invoke different forms of analysis in-situ over integrated large-scale data collections.

The new paradigm in data access is that we are moving from Data "My-ning" where users access the data they require and do their own processing on a local machine, to a new, more complex Data Mining approach. This new paradigm requires the user to be enabled to discover, access and process in real-time only those parts of multiple files and/or databases that are needed. Furthermore, they should be able to access these services online and drive them from any device including an iPad or a smart phone. LW noted that the impact of moving to this new paradigm for the marine and oceanographic community is that some degrees of freedom will be lost because common storage formats need to be used for the large data collections.

LW continued by describing the development of the high-performance computing (HPC) infrastructure in Australia. The Australian Government invested 375 million AUS$ to build a Big Data research infrastructure to make their national publicly funded data available through the Research Data Services initiative (RDS; https://www.rds.edu.au/project-overview). RDS supports over 40 petabytes of multidisciplinary data at nine nodes around Australia. Of this total, 1319 TB is marine/oceans data (not including the marine satellite data used for sea surface temperature monitoring).

The National Computational Infrastructure (NCI) is one of these nodes. It has established a powerful and comprehensive in-situ peta-scale computational environment to enable both high performance computing and data-intensive science across a wide spectrum of national environmental and earth science data collections. This comprehensive platform, called the National Environmental Research Data Interoperability Platform (NERDIP), includes the data, data management, data catalogues and data services needed to enable access by a variety of communities for multiple use cases.

More than 10PB of data have been co-located at NCI and comprise major national and international data collections from social to space data. These collections are called the National Environmental Research Data Collection (NERDC) and extend from the Earth's core up to astronomy. They represent one of the largest collections of Earth and environmental data in the world at a single site. The data is largely sourced from NCI's partners, major research communities, and collaborating overseas organizations (Evans et al., 2015). When combined, they offer unparalleled opportunities for geoscience researchers to undertake innovative data-intensive science at scales and resolutions never before attempted, as well as enabling participation in new collaborations in interdisciplinary research.

The NCI National Environmental Research Data Interoperability Platform (NERDIP) is a unified data platform that is being built to enable the same data to be used for multiple use cases both within, and beyond the oceans and marine community. To achieve this, formats need to be self-describing (netCDF) and all attributes need to conform to international standards for vocabularies and ontologies. NERDIP can loosely couple to multiple tools, virtual laboratories (VREs) and data portals.

LW presented two examples: the Virtual Geophysics Laboratory (VGL) and the Marine Virtual Laboratory (MARVL). High Performance access to data is facilitated through OpenDAP, OGC and other services, and fast programmatically-searchable catalogues. However, compared with other `Big Data' science disciplines (climate, oceans, weather, astronomy), current geoscience data management practices and data access methods need significant work to be able to scale-up and thus to take advantage of the changes in the global computing landscape.

Although the geosciences have many `Big Data' collections that could be incorporated within NERDIP, they typically comprise heterogeneous files that are distributed over multiple sites and sectors, and it is taking considerable time to aggregate these into large High Performance Datasets (HPD) that are structured to facilitate uptake in HPC environments. Once incorporated into NERDIP, the next challenge is to ensure that researchers are ready both to use modern tools, and update their working practices so as to process these data effectively. This is an issue in part because the geoscience community has been slow to move to peak-class systems for data-intensive science and integrate with the rest of the Earth systems community (Blue Planet Symposium, nci.org.au).

Potential issues for ODIP II to address:

- Converting Terabyte scale ‘Big Data’ sets that comprise thousands of individual heterogeneous files (e.g., bathymetry data sets) into ‘High Performance Data’ (HPD) sets
- Merging the high resolution LiDAR data sets (in LAS formats) with shallow water bathymetry (in CARIS, ASCII, ESRI Grid, and NetCDF) to create high resolution coastal elevation data sets for accurate tsunami and storm surge modelling.
- An agreed CF convention for data relevant to marine and oceans data

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16179

EU perspective – Streaming data processing

Adam Leadbetter (MI) presented the velocity aspects of Big Data in terms of getting data back to the data centre in real-time using some of the same technology as selected big commercial companies. We know how to do batch processing but we do not really know how to do it in real-time or near real-time and at scale as yet.

AL questioned how to scale real-time data streaming. He then explained the UNIX philosophy of McIlroy, Pinson and Tague (1978) and how the Marine Institute (Ireland) has applied the UNIX methodology of programming and tools for real-time data feeds. The MI explored the stream composition within the context of the Galway Bay cable observatory project. AL continued by providing some background information about the observatory and explained its workflow. One of its components is a CTD on a serial port with a hardware Moxa switch to make the serial connection available to multiple machines. A docker container is on a shore station server with a serial2kafka application running. The shore station Kafka service holds the data for a fortnight with the queue replicated across the network back to the laboratory. The raw data is stored in a Cassandra database and made available through ERDDAP. Alternatively, some augmentation of the data may be done through stream processing in Storm and then returned to a Kafka queue which is exposed through WebSockets.

The Marine Institute is also looking at how to provide the engines to build these data flows through message queues or processing tools without the need to write code. They are currently using NiFi, a web flow based programming tool, that is based on a drag and drop interface with a workbench for configuration so very little coding is required. However, it may only be of use mainly for ingesting data as far as a message queue.

AL continued by outlining how data streams relate to the Internet of Things (IoT). The IoT is the network of physical objects or "things" embedded with electronics, software, sensors, and connectivity to enable these 'objects' to collect and exchange data. Typically, IoT is expected to offer advanced connectivity of devices, systems, and services that goes beyond machine-to-machine communications (M2M). The interconnection of these embedded devices (including smart objects), is expected to usher in automation in nearly all fields.

The use of JSON messaging in this context is being addressed at the OGC SWE Domain Working Group which takes place on 16th September 2015. There is also a proposal for an O&M encoding in JSON that includes URIs so that SOS can be easily built on top of it.

Further work includes: a) looking at the architectures for reprocessing historic data; b) incorporating Sensor Web Enablement (SWE) and O&M-JSON; c) deploying it on vessels and mobile, remote platforms; and d) further investigation of Apache NiFi.

A group discussion concluded that a space should be created on the ODIP II website to share tools and libraries. It can also be used for relevant resources and links to other external material.

ACTION 1.12: DS to create discussion space on ODIP II website

JH then gave some online examples of different types of model workflow environments and how they have been implemented in Australia: the CoESRA virtual experiment environment (http://www.tern.org.au/CoESRA-pg29647.html) for complex data analysis, the Australian Urban Research Infrastructure Network (AURIN) Project (https://portal.aurin.org.au/), and the eReefs Project for nested modelling of hydrodynamic (3D water flows, temperature, salinity), geochemical (nutrients, Chl), ecosystem and fisheries models with other local models along the coast to support better understanding of the ecosystem. The framework used is HTML and Java based for the website with a postGIS database, THREDDS server and NetCDF CF compliant data files. All of the data services use OGC WMS/WFS standards.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16180

Addressing Variety and Veracity with GeoLink: a US perspective

Cyndy Chandler (WHOI) addressed the variety and veracity challenges of Big Data referencing the work done in the USA by the R2R and BCO-DMO projects.

In the context of Big Data, the term variety means the need to integrate a vast array of different data types. There has been a distinct change from scientists working with data they have collected themselves to using different data types from a variety of distributed systems. This data has broad temporal and geographic ranges and scales. There are in-situ observations and measurements from hypothesis-driven research (new instruments), model results and data from laboratory experiments. Further complexity arises from the need to integrate social science data with the full range of natural science data as well as taking into account new data types (e.g. metabolomics).

Veracity is data quality and accuracy. There are two specific aspects to be considered: 1) Data integration which requires high quality metadata that can be trusted. It must describe the data resource of interest and allow the scientist to ensure it is fit for their purposes. Provenance information is a huge challenge especially for data coming from many different sources. Efficient federation requires accurate resource matching between repositories with complementary content. 2) The increasing requirement for information to be made available in machine-interpretable forms. Increasingly the customer for the data is a machine rather than a researcher.

CC provided the example of the NSF EarthCube GeoLink project (Semantics and Linked Data for the Geosciences; http://www.geolink.org/) which is addressing some aspects of how semantic web technologies can be used to meet the variety and veracity challenges of big data.

The GeoLink project aims to bring together experts from the geosciences, computer science, and library science to develop semantic web components for geoscience research data that support discovery and reuse of data and knowledge. Existing semantic web technologies are stacked and used instead of inventing new ones: effectively using the Web as the API.

The project team includes a number of environmental data providers representing a broad cross section of the geoscience research community. The domain focus so far has been on marine environments using fieldwork and expeditionary data. It includes both sensor and sample data from observing networks including the "long tail" of research data. These activities have also been informed by those in many other projects and communities including ODIP.

Two use cases are being used to drive developments in the GeoLink project. The first on ocean ecosystems is looking at the data required for assessing distribution of zoo plankton in the sub-polar North Atlantic Ocean. The second is data requirements for understanding seabed morphology.

The GeoLink work plan includes publication of the ontology design patterns (ODPs) that were used to harmonize the content of different repositories. These ODPs were used to identify the essential attributes and properties that describe the main concepts of the information, for example, a field expedition is called 'a cruise' but the actually definition of this term can vary across different organizations. Different kinds of information were also identified, not only data but publications from peer reviewed journals, conference presentations, abstracts, PhD thesis, funding awards etc. The ODPs were used as a filter to publish subsets of their content as Linked Data (using to W3C standards). An integrated knowledgebase has also been developed and populated which is being exercised against science use cases.

The key "Linked Data" principles were followed: 1) use URIs (network names) to identify things (using controlled vocabularies definitions); 2) use HTTP URIs, so these things can be referenced by both humans and machines; 3) describe these things using standard languages such as RDF and SPARQL; and 4) include URIs (i.e. links) to other related things.

R2R and BCO-DMO both have a definition for cruise, platform (e.g. vessel) and instrument which are also matched and linked to the NVS terms and ICES platform codes. CC commented that this simple concept is very effective when bringing together a large amount of information. In his 2013 paper AL highlighted the importance of the NVS in such an approach by stating that "Terms from the NERC Vocabulary Server (NVS) are important for federating content from distributed systems" (Leadbetter et al., 2013). The vocabulary terms were mapped and linked to R2R and BCO-DMO content using the URIs from the NVS, for example, the R/V OCEANUS is defined as: http://vocab.nerc.ac.uk/collection/C17/current/32OC/. The MBLWHOI library is now also using the NERC vocabularies and publishes URIs in their RDF. CC suggested that it would be interesting to use this mechanism to connect data and publications but that this would require open data access.

An example from the VERTIGO Project illustrated how a cruise description and track line held by R2R is connected with a multi-band sonar dataset at R2R/NOAA and a sediment trap flux dataset held by BCO-DMO/WHOAS through the unique cruise identifier. The dataset is also linked to a journal article on the VERTIGO project using a DOI and to the NSF Funding Award. The chief scientist was also encouraged to get an ORCID that provides a unique person identifier. CC suggested that if all of the different repositories are using at least one common persistent identifier the web will make many of the connections automatically.

CC continued by providing a second example using work being done to collect sea samples in the Bering Sea. A cruise with an R2R cruise identifier and an ICES platform code are connected with a sample analysis published by the USGS National Geochemical Database (NGDB). This sample also has an International Geological Sample Number (IGSN) for identifying physical samples which allows the analyses to be linked to the cruise. These examples illustrate the power of adding persistent identifiers to as many things as possible at the instance level.

CC concluded that Semantic Web Technologies offer some solutions for meeting Big Data “Variety and Veracity” challenges. The key elements of this approach are the use of controlled vocabularies and Linked Open Data (RDF/SPARQL). Use of standards are also important e.g. BCO-DMO supports ISO19139, FGDC, GCMD DIF, schema.org dataset extension, formal data publication with a DOI, and RDF with semantic mark-up including PROV, FOAF and more (such as SKOS, OWL).

The main challenges, and ones that ODIP II could potentially help to address, are: the lack of key vocabularies published online using OWL with URIs; a lack of gazetteer data (e.g. physiographic features) published online with URIs and proper geometries; a lack of universal person and organization identifiers (with sufficient metadata); and the need to manually map/match instances at least in the beginning.

HG noted that one of the drawbacks of using ORCIDs is that they only require a person to provide their name in order to register and get a unique personal identifier without any other information being mandated e.g. affiliation. CC confirmed that there is a need for persistent identifiers that are unambiguously connected to a single researcher and ORCID is the best option. Researchers must be encouraged to populate a minimum set of information.

AK asked if GeoLink/BCO-DMO publish their own ontologies if they do not find the original ones. CC confirmed that they use URIs internally to connect with their knowledge base but they do not make them publicly available. AK suggested that there was a benefit to be had from using ontologies from other sources as these can also be extended and published. LW commented that this is similar to how vocabularies are maintained and published in Australia.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16181

OUTILS HYDRORUN: MarsWeb

Thomas Loubrieu (IFREMER) presented the MarsWeb service that provides access to the HPC resources located in Brest to people that are spread along the coast and in other overseas territories who are monitoring the coastal environment.
MarsWeb is a web interface which enables scientists to run models with configuration for inputs for atmospheric conditions, hydrology, bio-geo-chemistry, and tides. Functionality includes the ability to monitor the model running on the HPC in Brest, to visualize the results on-line and produce advanced model outputs. TL provided some illustrations of the system and its architecture.
There then followed a demonstration of the Sensor Nanny web application developed by IFREMER that provides services for observatories to manage their ocean observations on the "cloud" (http://snanny.ifremer.fr/dashboard.html). There are currently more than 2 million indexed points. These can be shown as either a density map or the actual measurements can also be visualized. Sub-setting of the data sets is also possible, the example given shows merging of Argo profiles with navigation data from IFREMER research vessels. The system is currently a cluster of seven servers but the intention is for IFREMER to expand this capacity as part its plans to build a Big Data infrastructure.

4.8.2 Discussion

AL summarized the points identified for further discussion during this session:

- NetCDF CF for marine and oceans data
- Merging key data sets (LiDAR, bathymetry, terrestrial geomorphology)
- Creating high performance datasets
- Data and workflows at scale

Net CDF CF for marine and oceans data

The first topic to be addressed was that of CF standardization. netCDF is already an OGC standard that was developed by a group led by CNR (Stefano Nativi) and UNIDATA (Ben Domenico). It was agreed that any proposals for CF standards should come from this group.

There are several CF attribute lists covering different topics but there appears to be no interaction between related groups in OGC, for example, there is no apparent overlap between bathymetry and geophysics. It was observed that these groups should be more coordinate in order to move forward together. Some of these groups also extend the CF standard terms but not in a compliant manner. It was acknowledged that there is a huge degree of freedom in the CF standard. The guidelines are relatively loose and the set of conversions are branched with two different versions for gridded data (1.5 and 1.7) and a version (1.6) for point data. There are also issues with data being identified as CF compliant, and even passes the CF checker, when it is not. The problem here is that official CF checkers are only for gridded data and not point data.

Organizations in Australia have adopted the US-IOOS netCDF CF checker and extended it by adding an IMOS plug-in. The IOOS checker can be modified to accept multiple plug-ins to satisfy different requirements.

It was acknowledged that one global implementation of the detailed CF conventions is not possible because different communities have distinct requirements. For example, the SeaDataNet CF profiles cannot carry the additional information required for bathymetric data. RL has proposed a layered structure with the CF conventions underlying a layer with the specific community conventions required for all data types. This approach has been implemented in SeaDataNet by adding an attribute to the profile for the SDN parameter codes in addition to the standard names. Other layers can also be added on top for specific types of data, e.g. for a bathymetry profile.

A new prototype development task was identified as a result of this discussion. It was agreed that the ODIP II partners would make an inventory of the CF profiles and checkers that are currently being used and by whom. It should also identify the plug-ins available for each checker. This inventory can then be made part of an OGC standard. The prototype will be led by the Australian partners (Sebastien Mancini).

ACTION 1.13: Sebastien Mancini (IMOS) will initiate the new NetCDF CF prototype development task

It was observed that netCDF is unpopular with developers who prefer to store data and metadata in a database. However, netCDF is meant to be an exchange format in this context and not for the purposes of archiving.

This new prototype development task can also address the issue of feature types, specifically the number of instruments/platforms per file including the current best practice. The most common approach is to have one instrument per netCDF file, but there are cases e.g. OceanSITES where a single netCDF file includes multiple instruments which makes it difficult for data management purposes. (It should be noted that OceanSITES is the official netCDF CF profile for the in-situ Thematic Assemble Centre (TAC)).

To support the development of this new prototype and other project activities, it was agreed to set-up a wiki (or similar tool) on the ODIP II website which will be used to host and manage all the information and discussion relating to standards (not only netCDF CF) in use within the marine community. An appropriate governance structure including responsibility for content management will need to be explored.

ACTION 1.14: Dick Schaap, Adam Leadbetter & Jonathon Hodge to explore an appropriate governance structure for a ODIP II wiki

Merging key data sets

The discussion then turned to the issue of merging key data sets for the marine community e.g. LIDAR, bathymetry and coastline. LW asked if there are countries that have problems with an agreed coastline and are unable to merge LIDAR with bathymetry data. HG noted that this is an issue for UK which uses a line to define the edge of the land. This makes integration of LIDAR and bathymetry difficult and requires significant effort. DS noted that in the Netherlands there are many different coastlines identified according to the definition e.g. high/low water mark) and depending on the usage. The issue is not the definition of the coastline but the methodology used for merging of the data that affects the results.

EMODnet bathymetry uses the relevant OpenStreetMap layers as it does not have a defined coastline. A new EMODnet project called Coastal Mapping will deal not only with the methodology of merging the data but also with the definition of a common coastal mapping in Europe (Note: the UK is not part of this initiative).

High performance datasets

The group agreed that the issue of creating high-performance datasets, the topic of data and workflows at scale and all the related issues such as how to build workflows, standards, handling of big data packages, performance of formats and visualization tools etc., could potentially be an interesting topic for a new prototype development task. It could also become an ODIP best practice that other projects can implement. It was agreed that the Australian partners would potentially lead a new prototype development task in this area but this would need to be confirmed with other partners not present at the workshop.

ACTION 1.15: Australian ODIP II partners to discuss possibility of leading a prototype development task on the topic of high performance datasets and related topics.

DS gave an example from the previous phase of the EMODnet project which had the right methodology but the performance of the tools was not right. The generation of EMODnet products could be greatly improved using the best practices that are developed by ODIP II. The relevant bodies could then be convinced to use the resulting powerful new tools (and the new data) for impact assessment calculations as part of the MSFD and other regional conventions. This makes the authorities, decisions makers (as end users), and not just the scientists, part of the process and also owners of the outcomes.

4.9 SESSION 8 - Data publication and persistent identifiers

4.9.1 Plenary

Introduction

Justin Buck began by giving an outline of the presentations that will be included in the session.

Coalition for Publishing Data in the Earth & Space Sciences (COPDESS)

Helen Glaves (BGS) presented the COPDESS initiative which is bringing together the Earth Science journal publishers and the data repositories to promote putting the policy of open, available, and useful data into practice.

The main drivers for the COPDESS initiative are: 1) to ensure that open access and open data mandates are widely acknowledged and being addressed; 2) the growing number of repositories that all need to adhere to best practices for open access and open data; 3) data and its citation are increasingly being recognized as part of the scholarly record; and 4) developments in cyberinfrastructure and eScience increasingly require access to data in standardized formats.

HG explained the data publishers' perspective. Many of the journals have had data supplements for some time but they are difficult to deal with and costly. Furthermore, they are often in PDF format which means they are not searchable, poorly indexed and of variable quality. The publishers require authors to comply with their policies on data availability and attempt to police this wherever possible. Another challenge for the publishers is the lack of guidance on individual community standards. The journal publishers also want to use and promote repositories and in particular those that have undergone some form of certification process but the repositories and journals are not well integrated with a few exceptions. The publishers are also concerned about repository stability, funding and longevity.

The data repository perspective is different with their main functions being data management, including quality control, and supporting discovery of the data. Their connection with the publishers is limited and often on an ad hoc or case-by-case but the repositories want better integration with the publishers.

COPDESS was founded in October 2014 in an attempt to address some of these issues. It is a permanent international coalition of publishers, data facilities and other consortia on Earth and space science data publication. The structure of COPDESS is still to be fully agreed and will be more clearly defined over a series of upcoming meetings. A draft Statement of Commitment was drafted and released on 15 January 2015. It states that that repositories and, to a lesser degree, the journals will adhere to best practice regarding data sharing and archiving, and how these two bodies of expertise will interact.

There are a number of actions required to support the aims of the COPDESS initiative. An online directory of Earth and space science data repositories that can be used by journals and authors needs to be built. This will assist in promulgating metadata information and data standards. There is also a need to develop common workflows within the repositories that support the peer review process, and within the editorial management systems to ease transfer of data to the repositories. Other activities supported by COPDESS include: promoting the referencing of data sets using the Force-11 data citation principles; promoting and implementing links to data sets in publications and corresponding links to journals in data facilities via persistent identifiers. Data sets should ideally be referenced using registered DOI's and COPDESS also promotes the use of other relevant community persistent identifiers for samples (IGSN), researchers (ORCID), and funders and grants (FundRef).

Forthcoming activities include publication of the COPDESS Directory of Repositories which will be released during August 2015. A workshop will take place in Europe (funded by the Sloan Foundation and the NSF) on 20 and 21 October 2015 in Oxford, UK. The main foci of this meeting are organizing training on the use of the directory, alignment of the journal publishers and integration with editorial managers.

Following the presentation several questions and issues were raised regarding repository certification and how this can be achieved. HG commented that there are several mechanisms including ICSU-World Data System (WDS https://www.icsu-wds.org/services/certification) and Data Seal of Approval (DSA http://www.datasealofapproval.org/en/). It was observed that that the biggest challenge for COPDESS will be the structure of the directory of repositories.

There was also the question of whether a repository should keep the streams of raw sensor data or just the processed data. The consensus was that, depending on the data type (satellite, water samples, etc.) and the level of confidence in the instrumentation, there is often no point in keeping the raw data. There are now intelligent sensors that process the data and there is no point in keeping the original signal, although it might be meaningful to keep the provenance metadata for the standards applied as part of the processing.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16187

RDA Marine Data Harmonisation IG and Data Citation WG: joint session

Helen Glaves (BGS) presented the outcomes from a joint session of the RDA Marine Data Harmonisation Interest group and the Data Citation Working Group that was held as part of the 7th RDA Plenary meeting which took place during the preceding week (also in Paris, France).

The Marine Data Harmonization IG was set–up to function in parallel with ODIP and bring together all those people with an interest in marine research and/or involved in marine data stewardship. The MDH IG aims to include all domains working in the marine environment and therefore has a wider scope than that of ODIP which was focused on physical oceanography in the previous project. The objectives of the MDH IG are similar to those of ODIP by aiming to promote a common global framework for marine data management through the use of common standards and best practices. This group also seeks to inform the activities of other RDA IG/WGs with relevant input and feedback from the marine domain including providing suitable use cases and disseminating the outcomes of other relevant RDA WGs/IGs to the marine data management community. The group is also working closely with other related initiatives such as Belmont Forum, IOC-IODE, GEOSS etc.

Within RDA there are a number of Working and Interest Groups with relevance to the marine community. The MDH IG is currently working with the Metadata IG/WGs to identify suitable use cases from the marine domain for the implementation of the outcomes of these groups. The MDH IG is also developing a plan for the evaluation and potential adoption of the outcomes of the Data Citation WG for the citation of dynamic data by the marine community. The aim is to identify a small number of suitable use cases/pilot studies and if these are successful develop a strategy for wider adoption of these solutions. The MDH IG in partnership with ODIP II will develop a proposal for an RDA Collaboration project to further evaluate the adoption of the outcomes of the Data Citation WG within the marine domain.

HG continued by outlining the proposals from the Data Citation WG for the citation of dynamic data. Citation of static datasets is relatively easy and already well understood where the whole dataset is being used. Mechanisms such as providing a URL for the datasets or assigning a persistent identifier to a dataset in a repository are used.

The reality is that citing data is more complex. There are complicated issues around the granularity of the data to be identified/cited. Databases collect enormous amounts of data over time and researchers use specific subsets of this data which need to be precisely identified. Among the current practices used is the storage of a copy of the data subset that has been used in the study but there are problems of scalability in using this method. Another approach has been to cite the entire dataset and provide textual description of subset but this is imprecise and leads to ambiguity. Attempts have also been made to store lists of record identifiers in the subsets but again this approach suffers with issues of scalability and is unsuitable where the entire record is not selected. There is an increasing need to be able to precisely identify and cite the subset of (dynamic) data used in a study.

Citation of data has previously required a dataset to be static and fixed with no changes being made (i.e. no error corrections, no new data being added). However, research data is often dynamic with new data being added and corrections being made. These changes are sometimes highly dynamic and made at irregular intervals.

Some of current approaches to citing dynamic data are: 1) to identify the entire data stream without any versioning; 2) use of “accessed at” date to indicate when the dataset was used; 3) use of “artificial” versioning by identifying batches of data as “snapshots” at a given time e.g. daily, weekly, etc.; or 4) aggregating all changes into releases of the dataset but this can lead to delays in corrections being exposed to the user. The ideal scenario is to be able to precisely cite the data as it existed at a certain point in time.

The solution proposed by the RDA Data Citation WG is to precisely cite the data as it existed at a certain point in time without delaying release of new data/updates to existing data. To make dynamic data citable it needs to be time stamped and versioned. This mechanism also requires the query used to create the subset of the data to be retained rather than the actual subset of the data. The query is assigned a unique persistent identifier and additional information such as time stamping etc. are then stored.

The RDA DCWG has created a suite of 14 recommendations for the citation of dynamic data relating to the preparing the data and storing the queries, persistently identifying specific datasets, resolving these persistent identifiers and retrieving the data, and the required actions when changes are made to an existing data infrastructure.

It was observed that the proposed solution is unworkable in real-world scenarios. The databases are dynamic and the data change so the queries also need to be dynamic which creates performance issues. JB commented that the Argo experience indicates that what is needed is to keep “snap shots” of the data and not keep track of the changes. However, Argo has now found an alternative solution which is currently being explored. It was also highlighted that the DCWG solution is not workable in Australia because any governmental database archive more than 7 years old cannot be recreated. A frozen local copy of the database is one possible solution because it can be managed locally and storage is not a problem with the advent of cloud computing technologies.

It was concluded that the RDA WG Data Citation recommendations are not suitable for most scenarios in the marine domain identified by the ODIP II community because the existing systems and infrastructure have been built in such way that cannot meet their criteria. To implement their proposals which require significant modification to these systems and a large investment of resources. It was agreed that a more consistent dialogue should be initiated with this group via the Marine Data Harmonization IG and there also needs to be greater representation from the marine community outside Europe in this group.

More information about the RDA Data Citation Working Group and its 14 recommendations on data citation are available on the RDA website at: https://rd-alliance.org/group/data-citation-wg/outcomes/ data-citation-recommendation.html. A wiki is also available for input and comments by the community. CC indicated that BCO-DMO will upload some use cases for data sets that meet these criteria but examples of time series data will be more difficult.

SSC commented that there is a difference between identifying a specific point in time for a dynamic dataset and capturing static monthly snapshots. He proposed defining a use case to look at the difficulties of getting to an arbitrary point in time for a dataset rather than restoring versioned static copies from the database or monthly snapshots of the data. LW noted that as these extracts from datasets can be used for activities that have legal implications or uses they should be clearly labeled.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16188

Argo DOI progress

Justin Buck (BODC) began his presentation by explaining the global Argo array which is made up of more than 3,000 free-drifting profiling floats. Each Argo float measures the temperature and salinity of the upper 2000 m of the ocean. This allows continuous monitoring of the temperature, salinity, and velocity of the upper ocean, with all data being relayed and made publicly available within hours after collection. This means that the data set is continually growing. At the same time the scientists and operators are also carrying out quality control on the data and this may lead to changes in the time series that already exist. As a result, there is a time series going back to 1998 which continually changes and expands. There are currently more than 2000 publications going back 15 years that cite Argo data but none of these citations can unambiguously reference the dataset at the point in time when the analysis was carried out which is a significant problem for reproducibility etc.

ARGO Global Data Assembly Centers (GDACs) were created 15 years ago to provide the current versions of the data. However, from the outset the US NODC decided that they needed snapshots of the entire data set every week. As a result, the Argo data cannot be reproduced at any given point in time but only at the granularity of approximately a week.

IFREMER currently assigns a separate DOI to every single snapshot but the goal of the Argo steering team and the scientific community is to have a single DOI for the whole dataset as this makes tracking and citation easier and more transparent.

The current proposal is to have a single DOI with a time identifier:

- Using the URI for the archive of Argo snapshots, followed by a "?" or a "#", followed by a query string identifier for the snapshot:
  - For example, http://dx.doi.org/10.7289/[Argo_accession_DOI]?[time_slice _information]
    - Using # references a client/browser side snapshot resolving service via a specific javascript for the accession
    - Using ? references a server side snapshot resolving service which is the preferred option but is not currently supported by DataCite.
- Where 7289 is the NOAA or IFREMER DOI prefix code
  - http://dx.doi.org/10.7289/argo_doi_identifier?result_time=2005-01-11T16:22:25.00

JB then showed an example of an Argo snapshot in the Sextant metadata catalogue and explained how to cite these data. In addition, the catalogue provides information to data managers on what metadata are available and how to mint a DOI but there is a debate on whether the citation should also be shown. Many peer reviewers want the data citations to go into the references section of the journal articles but in many cases it appears in the acknowledgments section which makes it harder to trace them, for example, a Google® search of an Elsevier paper will not find individual DOIs.

The recommendation from RDA for the single DOI with a time stamp (with a # or ? between), was not to include the date as part of the identifier. This could be a good idea because resolving time down to the second maybe somewhat spurious.

It was agreed that JB should put in a proposal to the current call from the RDA for collaboration projects. If successful it would provide a small amount of funding to explore more fully the issues of citing dynamic data from the Argo array.

ACTION 1.16: JB to consult with IFREMER and US NODC on developing a proposal for the RDA Collaboration Projects call.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16190

Australian Perspective on Dynamic Data Citation & IGSN

Lesley Wyborn (NCI) presented the Australian perspective on the dynamic citation of dynamic data. LW explained the concept of Dynamic-Dynamic Data which is data that is dynamically changing whilst being accessed by dynamic queries (those that are different and often unique each time). This situation is becoming more common with the increasing use of web services.

The issue of new data being dynamically added to existing data sets has been raised with the RDA Data Citation Group and at least two use cases were identified to illustrate these points:

- Use case 1: new data are regularly and systematically appended to an existing data set over time, e.g. outputs from a satellite or sensor continue to be added but no changes are made to the existing dataset.

- Use case 2: pre-existing data in a large data set is modified or updated. This use case is common where errors are found in pre-existing data, or new analytical and/or processing techniques are applied to selected attributes/components of the existing data set.

For use case 1 (appending): it was felt that RDA approach was rather database centric and did not apply to large volume raster arrays. This can easily be resolved by time stamping the source data and saving the query and the time of the query.

For use case 2 (subtle changes made to an existing large volume data set): large volume raster arrays can be over a Petabyte in volume and in multi-petabyte climate models, storing multiple time stamped snap shots of them is not feasible, primarily due to cost of the infrastructure.

A practical and workable solution includes:

- a controlled release process for the dataset, similar to that used for software, with the exact changes documented, so that if required (e.g., legal case), a data set can be recreated.
- a recommendation to use provenance workflow engines, that automatically capture the version of the data set that was used, the version of the software as well as the infrastructure to process the data, and the exact time the process was run. The provenance workflow itself will have a persistent identifier, as will all components of the workflow.

LW continued by presenting the Virtual Hazard Impact and Risk Laboratory (VHIRL) service as an example of how to preserve dynamic queries in a dataset using a provenance workflow engine. The VHIRL aims to advance natural hazard and risk modelling by accessing a collection of open community standards-based web services for both data selection and then processing of the selected data. It provides natural hazard researchers with access to an integrated environment that exploits eResearch tools and Cloud computing technology. The PROV service captures data service information (hosted on RDS), subset details for the data selected, code utilized along with "how" it is used (template/input files) and location (PID) of where input files/scripts are persisted. The finalized outputs are persisted with PIDs on RDS and captured in provenance information. After the job is completed and finalized, a provenance record is published to the provenance store. The PROV record endpoints could also be registered in ANDS Research Data Australia (RDA) alongside the output data.

LW continued by giving progress report on ANDS DOI minting:

- 41 institutions have signed minting agreements (34 are production ready, with 3 yet to mint their first DOI, a further 7 are in testing). Over half of all Australian universities are now minting DOIs for their published datasets (i.e., 24 out of 39 Universities). However, the total number of ANDS DOIs is still relatively low
- Manual minting of DOIs was introduced in December 2014 to augment machine-to-machine minting. A total of 9 institutions have taken up this option (only 1 is doing both machine-to-machine and manual)

Update from NCI

- M2M minting is ready to go but no production DOIs minted yet due to some political issues over who issues the DOI.
- currently working through business processes with their providers. Issues being addressed include:

- when and if NCI should mint a DOI for data to be made public (i.e. what if the provider has their own minting capability?);
- agreeing on the DataCite metadata - in particular the role of the provider institution and NCI.

Other nodes

- Each of the nodes has/will address the issue of DOIs, though some have/are likely to determine that this function is best managed by the provider institution and not the node.

Finally, LW outlined the Australian contribution to the International Geological Sample Number (IGSN) project. It is funded by Research Data Australia (RDS) for petascale data challenges which is seeking to bring in data on physical samples that can be used to calibrate the petascale, proxy data sets. There are three IGSN allocating agents in Australia Curtin University, CSIRO and GA. The project aims to better coordinate the implementation of IGSN in Australia, in particular how these agencies allocate IGSN identifiers. The project will register samples from pilot applications in each agency and these local catalogues will then be aggregated into an Australian portal which will ultimately be expanded for all geoscience specimens. The development of the portal will involve developing a common core metadata schema for the description of Australian geoscience specimens, and will also require formulation of an agreed governance model for registering samples.

JB enquired about OGC standards for a sample. LW replied that it is entirely based on the O&M content model which is designed for sampling features and the samples taken from these feature (e.g. a borehole and all the samples derived from it from it. BA noted that an IGSN can be assigned both for a physical specimen and the feature of interest (e.g. an IGSN can be allocated to an ocean site. This drill site can be revisited 10 years later and further deepened but reference is made to the original identifier). An IGSN can be used to search repositories and find what work has been done on an individual sample. CC mentioned that WHOI uses IGSNs to identify water samples from NISKIN bottles and not only rocks samples.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16191

IGSN: International Geo Sample Number

Bob Arko (LDEO) provided a short overview of the International Geo Sample Number (IGSN) system for unambiguous citation of physical samples. The rationale behind the persistent identifiers for physical specimens is to have discoverable, accessible and citable physical specimens allowing:

- Discovery and access for re-use and reproducibility - samples need virtual representations which have resolvable persistent identifiers (PiDs)
- Sample Citation - sample collectors need to get credit for the intellectual effort and resources they put into the collection (especially in the ocean), preparation, and curation of these samples
- Data Integration - sample data are highly dispersed because a single sample is often studied in many different labs and over long periods of time with data published in multiple articles. The usefulness of these data is substantially higher when combined

- Sample Management: allows tracking of samples and sub-samples

IGSN was introduced 10 years ago (http://www.igsn.org/) to provide a globally unique and persistent identifier for physical Earth science samples. The concept is to have a guaranteed unique identifier for every specimen which can resolve to a landing page in a similar way to a DOI. BA reported that IGSN are now beginning to appear in the literature.

A recent (September 2015) IGSN Working Group meeting recognized that there is core set of essential metadata elements for an individual sample that are needed to describe it and make it discoverable in a search engine. The curation information about all the things that happen to the sample through time are distinct from the core metadata. The core metadata for a sample will be turned into an updated ISO and OGC O&M schema (including some extra information such as sampling type, material type, sampling method, etc. that are used in DOI) and will be published following the DataCite metadata principles.

Using a specific example of an ODP core, BA demonstrated how IGSNs are used to track the provenance: first the sampling hole itself gets an IGSN (the parent), then all of the cores from the same hole get an IGSN which identified them as a sibling. Each section of a borehole is also allocated and IGSN which allows it to be identified as a subsample of an individual core (sample). This system also allows additional information to be associated with both the hole and the sample at a much later date and can also allowing tracking of sub-samples.

Initial efforts were focused on rocks and sediments but was then extended to include fluids and gases, and most recently biological samples. Funding agencies advised that in order for publishers to adopt IGSN it would need to be applicable to biological and geo samples.

IGSNs now appear in peer publications and when published online the sample's 'birth certificate' can be obtained through the IGSN links.

For ODIP, one potential use case is to link the IGSN with sampling data from research cruises. Geological samples (core, dredges, grabs, etc.) could be catalogued in the Index to Marine and Lacustrine Geological Samples using IGSNs. These identifiers could be linked to R2R Cruise IDs for U.S. vessels and sample inventories included in the cruise summary reports published to POGO.

Presentation available at:

http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16192

4.9.2 Discussion

A discussion followed around the need for well-organized local management systems with local samples identifiers that are linked with the global IGSNs. Linking IGSNs with sensor data was also discussed.

JB summarized the session discussion points and possibilities for ODIP II which were:

- Citing and versioning of large (Petabyte+) dynamic datasets
- The scalability of DOIs
- RDA collaboration proposal on dynamic data citation (BCO-DMO and BODC)
- The development of citation indices

4.10 SESSION 9: Cross–cutting topics: break-out sessions

During this part of the workshop programme the attendees were offered the opportunity to participate in three parallel break-out working groups on: vocabularies, data publication/citation and model workflows/big data. The parallel working groups were organized as two separate sessions allowing each attendee to participate in two of the three topic groups. A rapporteur was appointed for each break-out group who then provided a report of the key discussion points during the next session (see below).

4.11 DAY 4: SESSION 10 -Cross-cutting topics break-out session reports

4.11.1 Model workflows and big data

Adam Leadbetter summarized the main points from both 'Model workflows and Big Data' break-out groups. The discussion focused on the tools and platforms that are used for big data and model workflows. The key ones being:

- Kepler (workflows engine)
- Taverna (workflows engine)
- Zoo (WPS Wrapper)
- Model interfaces to connect inputs and outputs (OpenMI – OGC Standard)
- Lab Collector (a laboratory information management system for biogeochemical workflows)
- Cloud provisioning
    - Cloud first design vs. redeployment on the cloud

The group also identified some relevant use cases that could be developed further in the ODIP II project:

- Use Case 1:
    - Climatology
        - Using NCI platform
        - SeaDataNet harvested temperature and salinity data
        - US, Australian data
        - Choose a focus area
        - Build workflow for creating the climatology
        - Visualisation, performance, scalability: all issues to be addressed

It was noted by SM that the Australians have recently put together all IMOS temperature and salinity measurements along the shelf as point data. DS commented that for the use cases the methodology and not the product that is of most interest.

- Use Case 2:
    - Biochemistry mooring data
        - Laboratory analysed samples
        - Automation of workflows once data are analysed
        - Requirements
            - Vocabularies
            - Sensor descriptions
            - Calibration information

- Discover environmental information relating to taxonomic identifications
- Simple estimation of where else the species might occur

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16196

4.11.2 Vocabularies

Alexandra Kokkinaki (BODC) presented the outcomes of the vocabularies working group. A vocabularies "wish list" was compiled that focused on three areas of investigation:

1.Further development of mappings

- Implement unit conversions through rich predicates such as:
  - http://vocab.nerc.ac.uk/collection/P06/current/ULCM/
  - 1/100
  - http://vocab.nerc.ac.uk/collection/P06/current/ULAA/
- Map Marine Metadata Interoperability Ontology Registry and Repository (MMI orr) to P07
  - http://mmisw.org/orr/#http://mmisw.org/ont/cf/parameter
- P02 upgrade to GCMD 8
  - NVS2 currently uses version 6
  - mapping to GCMD version 8 URIs
  - investigate how GCMD currently works with URLs

2. Further development of content

- ODIP to expand C19 vocabulary
  - SeaVoX salt and fresh water body gazetteer
  - add more content to the geometry server
  - submission by ODIIP partners of content originating from the relevant authority
    - Preferred Label for the sea (e.g. Adriatic sea)
    - Spatial Coverage in GML or WKT

- Create a vocabulary with terms for fitness for purpose semantic annotation of datasets (from EMODnet Check Points)
- Access GEBCO undersea features as linked data, e.g. Australian local seas, by using URIs
- Overlay SKOS with OWL (show A01 example)
  - Based on wish list of vocabs (P01/P02, L22/L05)
- Add Semantically richer predicate set in NVS2
- Create a self-service governance to help users create their own P01 one arm bandit vocabularies

- Add richer predicates to P01 to map with P07

3. Further development of tooling

- Create ontologies or rules to hold the knowledge and its eccentricities e.g. describe the different components of vocabularies

4. Best practices

- Develop best practices for embedding vocabularies (parameters, units, instruments) in netCDF files
- Feed mappings back into MMI

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16197

4.11.3 Data citation / persistent identifiers

Justin Buck (BODC) reported that the breakout group focused on various aspects of citing dynamic data in an attempt to identify some suitable use cases for ODIP II.

Relevant use cases that are creating dynamic data that were identified during the session included:

- IMOS – versioning of multiple datasets
- Marine Institute – near real-time platform measuring chlorophyll data
- IFREMER/BODC - Argo float data
- BCO-DMO – file versioning
- IOOS – workbench data
- NCEI – Argo data

Of the use cases identified it was agreed that there are at last 4 or 5 good exemplars that could be explored further.

The break-out group also discussed some of the details of citing dynamic data. The model for citation of data is:
http://dx.doi.org/10.7289/[Argo_accession_DOI]?[time_slice _information]

The use of “#” or “?” was discussed:

- IFREMER has worked with CNRS to make the use of # possible. This also needs to be submitted to DataCite and CrossRef for wider adoption by the DOI issuing authorities.

The issue of opaque or transparent time information was also discussed:

- Advantages from user perspective with using transparent time but there is the potential for the citation of ambiguous data state
- Further clarification to be sought from DataCite and CrossRef on this issue

Outcomes of the RDA Data Citation working group on the citation of dynamic data:

- There is currently an opportunity for funding from the Research Data Alliance to develop pilot studies for testing the implementation of the recommendations for the citation of dynamic data from the RDA DCWG.
- A potential proposal from the ODIP community would need to address two themes:
  - Liaison with DataCite and CrossRef to address implementation issues
  - Develop and show the viability of prototypes
  - Decision was taken to put in a bid as an ODIP consortium rather than from individual data/observing entities. The opportunity for the US partners to obtain matching funding from RDA USA will also be explored

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16198

4.12 SESSION 11 - ODIP II: new development activities & cross cutting themes

SOOS Field Projects Portal

Sebastien Mancini (GA) presented the Southern Ocean Observing System (SOOS) Field Projects Portal. “SOOS is an international initiative that facilitates the collection and delivery of essential observations on the dynamics and changes of the Southern Ocean systems through the design, advocacy and implementation of cost-effective observing and data delivery systems”.

Two key tasks of SOOS are:

a) building tools to share the data once it has been collected (the GCMD metadata platform will be used)

b) building a field project planning portal to promote discussions before the data are collected.

The Field Project Planning Portal which will be developed is a tool with a publicly-editable spatial database to show what field campaigns are being done where and by whom over the coming years. Many oceanic and Antarctic research communities want to develop similar things (e.g. ICED, COMNAP, AFIN, ASPeCt, SCAR, SOCCOM, BCO-DMO, Argo etc.) but resources are scarce and users are unlikely to use multiple parallel tools. A collaborative modular process may be the best approach to solving a hard coding problem with limited resources. The Field Project Planning Portal will be based on this approach with users being provided with data entry forms that are tailored to their specific requirements. The user will use this simple entry form to provide details of the field survey including where, when, what the instruments that will be used, Principle Investigators, etc. This will system will allow researchers to potentially pool their resources and effort in a similar manner to a cruise planning system.

Ideally, the portal will have the following features:

- Publicly submittable/editable spatial information (points, lines, multi-point features)
- The capacity to attach multiple records to a single geographic feature

- Record information on ship name, dates, geographic region, planned experiments, PI contact details, berth availability, data URL (if available), requests for collaborators

It was observed that the majority of those people participating in SOOS are also involved in ODIP II. There is therefore the potential for ODIP II to assist with some aspects of developing the Field Project Planning Portal, particularly with the design of the infrastructure and the data input form; building a robust back-end and intuitive user-interface; testing and maintenance of the code, and web hosting.

DS and LR commented that both POGO and Eurofleets have addressed some of these issues. They will therefore provide the relevant people involved in SOOS with feedback.

ACTION 1.17: DS and LR to provide feedback to the SOOS development team (data@soos.aq) on the relevant experiences from POGO and EMODNET of developing portal technologies

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16199

4.12.1 Discussion

Helen Glaves (BGS) asked partners to provide suggestions for new cross-cutting topics that should be included in the agenda for the next ODIP II workshop including any additional issues that came out of the discussions during the current workshop that are not covered by the existing list of topics.

It was agreed that going forward the topic of data ingestion will no longer be considered as a separate topic in ODIP II unless the need arises at some later stage. Instead, it will be integrated into other topics such as SensorML, or as part of the discussion around standards for ingestion systems etc.

The scope of ODIP II was also discussed during this session. HG pointed out that the project has now been extended to include marine biology. There is therefore a need to expand existing ODIP activities and create new ones that are also relevant to this discipline. However, there was limited representation from the marine biology community at this workshop so it was agreed to discuss this topic in greater depth at the next one.

It was observed that it is only a matter of time before other communities such as climatology become interested in the participating in the work being done by the ODIP community. However, this is beyond the scope of the DoA of the current ODIP II project. Its mission is interoperability and common standards in the marine and ocean domain. DS commented that ODIP II is not providing an end-to-end system, in that it does not communicate with the end users and cannot cover every perspective. ODIP is a technical project that feeds into other projects that are directly addressing the needs of the end user. ODIP is also leveraging these existing projects and can have access to end users through them in order to understand their requirements and to identify success stories for the ODIP II impact assessment.

The role of the Research Data Alliance (RDA) was also discussed and how it is aligned to that of the ODIP II project and in particular the role played by the RDA Marine Data Harmonization IG in acting as an interface between the marine domain and the other more agnostic activities of the RDA IG/WGs. It was agreed that cooperation with RDA, Belmont Forum, OGC, RCN OceanObs and other groups ensures that ODIP II is not operating in isolation but instead is fully integrated into the relevant activities of other related initiatives.

Where there is not already representation from these initiatives within the ODIP II community they should be invited to participate in future workshops. For example, management of polar data is a potential new topic for ODIP II and this community could be invited to participate in the next workshop. This could also potentially expand the ODIP community into other regions such as Canada.

Another key discussion point was how to share the knowledge that is produced by the ODIP II project. It was agreed that this should be carefully managed to avoid adding extra technical work load.

The group also discussed how to address the specific topics that were either included in the original ODIP II proposal or subsequently identified during the workshops. Each topic needs to be addressed either as a cross-cutting theme or included as part of the new ODIP II prototype development tasks. It was also noted that the development of a prototype solution is more constrained and has specific targets whereas the cross-cutting topics are more flexible with partners providing updates on their respective individual activities rather than coming together to work on a specific task as a group.

It was agreed that work will continue on the three existing prototype development tasks and that the new topics that have been identified during the workshop will be addressed as cross-cutting themes at least until the next workshop.

The partners from the USA and Australia USA indicated that they are willing to continue to be part of the ODIP II consortium despite on-going funding issues which they will aim to resolve. However, the level of contribution to the new and existing prototype development tasks is uncertain without support from their respective funding agencies. HG commented that the EU is aware of these issues and the potential implications for the ODIP II project. The EU Project Officer has indicated that there is a degree of flexibility in how the project is delivered as long as the Commission is kept informed of any significant deviations from the DoA.

It was suggested that the ODIP 2 prototype should become a fully operational system. This would allow R2R to use existing funding to enhance and routinely populate the CSR system (and therefore continue to part of the ODIP II project). There is also scope for the same approach to be used by some Australian partners (IMOS) who have implemented the NERC vocabularies (NVS).

It was also suggested that a way forward might be to combine a number of the separate regional uses cases into a single larger more robust use case which has more impact and wider applicability. It could still be called a new prototype development task but adopt a different approach.

4.13 SESSION 12 - Workshop wrap-up

4.13.1 Plans for next 8 months

Helen Glaves (NERC- BGS) outlined the project activities for the next 8 months until the next ODIP II workshop. These actions include those required to finalize the first phase of the ODIP project which ends on 1 October 2015.

ODIP project finalisation

There are seven ODIP deliverables still outstanding that must be completed and submitted immediately. The most important of these is *D3.4 Results and conclusions from prototype analyses* which requires input from partners by the end of next week (9 October 2015).

ACTION 1.18: MARIS to finalize and submit deliverable D3.4

The outstanding ODIP deliverables are:

- D1.12 Final report including cost statements (September 2015; M36)
- D3.4 Results and conclusions from prototype analyses (May 2015; M32)
- D4.2 Final strategic analysis report (September 2015; M36)
- D5.6 Promotional leaflets and posters (July 2015; M34)
- D5.7 Future ODIP exploitation plan (July 2015; M34)
- D5.8 Common ODIP standards submitted to the IODE Ocean Data Standards (ODS) process (September 2015; M36)

The leaders of the work packages must submit their reports to the coordinator by 12 October 2015 at the latest. Partners must submit their final cost statements, including justification for the use of resources, via the EU-ECAS system by 30 October 2015. All deliverables must to also be completed and submitted to the EU by 30 October 2015.

ACTION 1.19: All partners to submit final cost statements (FORM C) by 30 October 2015

ACTION 1.20: Work package leaders to submit their reports to the coordinator for inclusion in the final report by 30 October 2015

The final review of the ODIP project will be on 13 November 2015 in Brussels. Only the coordinators and the work package leaders are required to attend.

ODIP II: upcoming deadlines

HG highlighted the following deliverables which are due for completion and submission to the EU during the next 8 months:

- D1.1 6 monthly progress report (M7: October 2015)
- D1.2 Minutes of ODIP II steering committee (M6: September 2015)
- D1.5 Operational extranet (M3: June 2015)
- D2. ODIP II workshop 1 (M5: August 2015)
- D2.2 Minutes and actions of ODIP II workshop (M7: October 2015)
- D3.1 Definition of prototypes (M6: September 2015)
- D5.1 Dissemination and communication plan (M5: August 2015)
- D5.2 ODIP II website (M3: June 2015)
- D5.4 Promotional leaflets and posters (M5: August 2015)

The following dissemination opportunities have also been identified:

- AGU Fall Meeting 2014: 15 – 19 December 2015, San Francisco, USA
- AGU Ocean Sciences: 21 – 26 February 2016, New Orleans, USA

- EGU General Assembly: 17 – 22 April 2016, Vienna, Austria
- 7th RDA plenary: 29 February - 3 March 2016, Tokyo, Japan
- Others ??

HG concluded by informing the participants that the 2nd ODIP II workshop will be held in Boulder, Colorado, USA and will be hosted by UNIDATA/SIO. The provisional dates being considered for the workshop are either May or June 2016, and it will be organized back to back with the R2R annual meeting.

Presentation available at:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=16200

4.13.2 Closing remarks

HG closed the 1st ODIP II workshop by thanking Sissy Iona (leader of WP2) and the IFREMER staff, especially Thomas Loubrieu and Béatrice Milosavljevic, for organizing the workshop. She also thanked everyone who participated both in person and remotely. DS also thanked the participants for their contributions which led to a very productive workshop.

Annex A Workshop Programme

DAY 1: Monday, 28 September 2015

SESSION 1 - Introduction

08:45 –09:00 Registration

09:00 –09:10 Welcome & Workshop logistics, *Helen Glaves/Dick Schaap*

09:10 –09:20 Workshop aims and objectives, *Helen Glaves (ODIP project Coordinator)*

09:20 –09:35 *Introduction by partners (Name, Country, institution, main responsibility, expectations for this workshop, 30 seconds max)*

ODIP II Overview

09:35 – 09:55 ODIP II: overview of the project including aims and objectives, *Helen Glaves (Coordinator)*

09:55 –10:15 ODIP II: development of potential activities, *Dick Schaap (Technical coordinator)*

10:15 –10:35 Discussion
Partners are invited to propose additional activities (max 2 slides)
Led by Helen Glaves & Dick Schaap

10:35 –11:00 Break

SESSION 2 - ODIP Prototype Development Task 1: plenary

11:00 –11:40 ODIP 1: aims, activities and progress, *Dick Schaap (EU)*

11:40 – 12:00 ODIP 1: report on impact assessment, *Thomas Loubrieu*

12:00 – 12:30 Discussion, *Led by Dick Schaap*

12:30 –13:30 Lunch

SESSION 3 - ODIP Prototype Development Task 2: plenary

13:30 –14:20 ODIP 2: aims, activities and progress, *Led by : Anne Che-Bohnenstengel& Friedrich Nast*

- ODIP 2 development task: progress and results, *Bob Arko*
- CSR harvesting: update on progress, *Anne Che-Bohnenstengel*
- Partnership for Observation of the Global Oceans (POGO), *Lesley Rickards*

14:20 –14:40 ODIP 2 report on impacts assessment, *Thomas Loubrieu*

14:50 –15:20 Discussion, *Led by Anne Che-Bohnenstengel & Friedrich Nast*

15:20 –15:45 *Break*

SESSION 4 - ODIP Prototype Development Task 3: plenary

15:45 –16:25 ODIP 3: aims, activities and progress, *Led by Jonathan Hodge (CSIRO)*

16:25 – 16:45 ODIP 3 report on impacts assessment, *Thomas Loubrieu*

16:45 – 17:15 Discussion, *Led by Jonathan Hodge*

DAY 2: Tuesday, 29September 2015

SESSION 5 - ODIP prototype development tasks: feedback on outcomes and possible next steps

09:00 –10:30 ODIP prototype development projects,
Feedback from each group on final outcomes and potential further developments in ODIP II (30 minutes each)

- ODIP 1, *Dick Schaap*
- ODIP 2, *Anne Che-Bohnenstengel & Friedrich Nast*
- ODIP 3, *Jonathan Hodge*

10:30 –11:00 *Break*

ODIP prototype development tasks outcomes and possible next step: discussion

11:00 –12:45 Discussion: *Led by Dick Schaap*

12:45–13:45 *Lunch*

SESSION 6 – Vocabularies: plenary

13:45 –15:15 Vocabularies, *Led by Roy Lowry*

- NVS Developments, *Roy Lowry & Alexandra Kokkinaki*
  - 'One-armed bandit semantic model'
  - NVS search client
  - NVS Linked Data demonstration

- Report on AODN and ANDS vocabulary developments, *Sebastien Mancini*
- Report on RDA VSIG activities, *Rob Thomas*

15:15 –15:45 *Break*

Vocabularies: discussion

15:45–16:45 Discussion, *Led by Roy Lowry*

DAY 3: Wednesday, 30 September 2015

SESSION 7 – Model workflows and big data: plenary

09:00 –10:30 Model workflows and big data, *Led by Adam Leadbetter (EU), ?? (USA) & Lesley Wyborn (Australia)*

- Intro - what is Big Data (not just volume, but other aspects too), *Adam Leadbetter (MI)*
- Australian perspective – what has already been achieved and more, *Lesley Wyborn (NCI) & Jonathan Hodge*
- EU perspective – Streaming data processing, Adam Leadbetter
- Addressing Variety and Veracity with GeoLink: a US perspective, *CyndyChander (WHOI)/Bob Arko (LDEO)*

10:30 –11:00 *Break*

Model workflows and big data: discussion

11:00 –12:00 Discussion, *Led by Adam Leadbetter (EU) & Lesley Wyborn (Australia)*

12:00 –13:00 *Lunch*

SESSION 8 - Data publication and persistent identifiers

13:00 –14:30 Plenary, *Led by Justin Buck (EU)& Lesley Wyborn (Australia)*

14:30 – 15:30 Discussion, Led by Justin Buck

15:30 –16:00 *Break*

SESSION 9 – Cross –cutting topics: break-out sessions

Smaller informal group discussions addressing the cross-cutting topics currently being addressed in the ODIP II project. These discussion groups were run as two parallel 45 minute sessions.

16:00 –17:30 Cross-cutting topics: break-out session

- Vocabularies
- Data publication/citation
- Data workflows/big data

DAY 4: Thursday, 01 October 2015

SESSION 10 - Cross-cutting topics break-out session reports

Feedback on outcomes from workshop and proposed next actions

09:00 – 09:20 Model workflows and big data, *Adam Leadbetter*

09:20 – 09:40 Vocabularies, *Roy Lowry*

09:40 – 10:00 Data citation/Persistent identifiers, *Justin Buck*

10:00 –10:30 Break

SESSION 11 - ODIP II: new development activities & cross cutting themes

10:30 –12:00 Discussion, *Led by Helen Glaves/Dick Schaap*

SESSION 12 - Workshop wrap-up

12:00 – 12:15 Plans for next 8 months (including final ODIP reporting, status of ODIP and ODIP II deliverables and next workshop), *Helen Glaves/Sissy Iona/Dick Schaap*

12:15 – 12:30 Closing remarks, *Helen Glaves/Dick Schaap*

ANNEX B Table of Actions

Action No.	Action	Responsible
1.1	Coordinators to identify a Steering Committee member for NOAA	Helen Glaves
1.2	Coordinators must seek a nomination for a representative of the biological oceanography to sit on the Steering Committee	Helen Glaves
1.3	TL to inform Oceans of Tomorrow project of relevant RDA IG/WG	Thomas Loubrieu
1.4	CNR to correct current issues with CSW and OAI-PMH services	Stefano Nativi
1.5	Checks to be made to ensure numbers of collections delivered by the regional systems equate to the numbers being harvested and delivered to the global portals	Dick Schaap
1.6	Checks to be made on the accuracy of the spatial information shown in the global portals for the regional collections	Dick Schaap
1.7	Partner MARIS to complete deliverable D3.4 Results and conclusions from prototype analyses	Dick Schaap
1.8	BSH to define a generic use case to capture user requirements for cruise summary reports (CSRs)	Friedrich Nast
1.9	Develop a specific use case to drill down into the detail of these user requirements including how the user wants to find the resources they need. This use case should also include the expected role of the ODIP II project from the user perspective.	All Partners
1.10	Put a WMS-WFS on top of the CSR services to expose them to other systems including for the purposes of implementing a linked data approach.	Friedrich Nast
1.11	Produce a summary document that outlines current approaches to using ontologies for mapping to	Alexandra
1.12	Discussion space to be created on ODIP II website	Dick Schaap
1.13	IMOS to initiate the new NetCDF CF prototype development task	Sebastien Mancini
1.14	Governance structure for a ODIP II wiki to be explored	Dick Schaap, Adam Leadbetter, Jonathon Hodge

1.15	Australian ODIP II partners to discuss possibility of leading a prototype development task on the topic of high performance datasets and related topics	Jonathan Hodge
1.16	BODC, IFREMER and US NODC to consult on developing a proposal for the RDA Collaboration Projects call	Justin Buck
1.17	Provide feedback to the SOOS development team (data@soos.aq) on the relevant experiences from POGO and EMODNET of developing portal technologies	Dick Schaap, Lesley Rickards
1.18	MARIS to finalize and submit deliverable D3.4	Dick Schaap
1.19	All partners to submit final cost statements (FORM C) by 30 October 2015	All Partners
1.20	Work package leaders to submit their reports to the coordinator for inclusion in the final report by 30 October 2015	Work Package Leaders

ANNEX C Terminology

Term	Definition
ANDS	Australian National Data Service
AODN	Australian Ocean Data Network
API	Application Programming Interface (API): a set of routine definitions, protocols, and tools for building software and applications
ARGO	Argo is a system for observing temperature, salinity, and currents in the Earth's oceans which has been operational since the early 2000s
BCO-DMO	Biological and Chemical Oceanography Data Management Office at the Woods Hole Oceanographic Institution http://www.bco-dmo.org/
BCube	A two-year project funded by the National Science Foundation (USA) exploring the use of brokering technologies to make it easier for scientists to discover, share and access data http://www.earthcube.org/group/bcube
CAS	Chemical Abstracts Service: a division of the American Chemical Society. It is a source of chemical information https://www.cas.org/
Cassandra (Apache)	Free and open-source distributed database management system developed by Apache
CCAMLR	Committee for Conservation of the Antarctic Marine Living resources https://www.ccamlr.org/
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/

CheBI	Chemical Entities of Biological Interest: a freely available dictionary of molecular entities focused on 'small' chemical compounds https://www.ebi.ac.uk/chebi/init.do
CKAN	Comprehensive Knowledge Archive Network: a web-based open source management system for the storage and distribution of open data http://ckan.org/
COOPEUS	EU-NSF funded project promoting open access and sharing of data and information produced by environmental research infrastructures
Copernicus	A European system for monitoring the Earth. Previously known as Global Monitoring for Environment and Security (GMES)
CrossRef	Official Digital Object Identifier Registration Agency of the International DOI Foundation http://www.crossref.org
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
CTD	Oceanography instrument used to determine the conductivity, temperature, and depth of the ocean
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 software for the spatial interpolation/gridding of data
DoA	Description of Action
Docker	Docker containers wrap a piece of software in a complete file system that contains everything needed to run: code, runtime, system tools, system libraries – anything that can be installed on a server. This guarantees that the software will always run the same, regardless of its environment.

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
EarthCube	NSF-funded initiative creating cyberinfrastructure to improve access, sharing, visualization, and analysis of all forms of geosciences data and related resources in the USA http://www.earthcube.org/
EDIOS	EuroGOOS initiative to provide an overview of the ocean measuring and monitoring systems operated by European countries. http://www.seadatanet.org/Metadata/EDIOS-Observing-systems
EDMED	European Directory of Marine Environmental Data
EDMO	European Directory of Marine Organisations
EduGain	Service interconnecting identity federations around the world enabling the trustworthy exchange of information related to identity, authentication and authorisation (AAI).
eMII	eMarine Information Infrastructure (IMOS facility name). As of 1 June 2016 it was renamed as the Australian Ocean Data Network (AODN)
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/
EMSO	European Multidisciplinary Seafloor and water-column Observatory: a large scale, distributed, marine Research Infrastructure (RI) http://www.emso-eu.org/
ERDAPP	NOAA's Environmental Research Division's Data Access Program: a data server providing access to subsets of scientific datasets in common file formats that also provides visualization tools.
eReefs	Australian collaborative project to develop n information system for monitoring the Great Barrier Reef and predicting future changes http://ereefs.org.au/ereefs

ETL	Extract, transform and, load database functions combined into a single software tool
FOAF	Friend of a friend: a machine-readable ontology describing persons, their activities and their relations to other people and objects http://www.foaf-project.org/
GBIF	Global Biodiversity Information Facility: an international organisation making scientific data on biodiversity available via the Internet using web services http://www.gbif.org/
GCI	GEOSS Common Infrastructure https://www.earthobservations.org/gci_gci.shtm
GCMD	NASA's Global Change Master Directory: descriptions of Earth Science data sets of relevance to global change research http://gcmd.nasa.gov/
GEBCO	General Bathymetric Chart of the Ocean: authoritative, publicly-available bathymetry data sets for the world's oceans http://www.gebco.net/
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
GO-SHIP	Global Ocean Ship-based Hydrographic Investigations Program: initiative to develop a globally coordinated network of sustained hydrographic sections as part of the global ocean/climate observing system http://www.go-ship.org/
GridFS	A scalable MongoDB file system for storing and retrieving large files
Hyrax	OPeNDAP 4 Data Server
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/
INCHI	International Chemical Identifier: a textual identifier for chemical substances, designed to provide a standard and human-readable way to encode molecular information http://www.inchi-trust.org/
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).
IOOS	US Integrated Ocean Observing System https://ioos.noaa.gov/
ISO	International Organization for Standardization http://www.iso.org
JCOMM	WMO-IOC Joint Technical Commission for Oceanography and Marine Meteorology: partnership of marine meteorological and oceanographic communities to respond to interdisciplinary requirements for met/ocean observations, data management and service products http://www.jcomm.info/

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.
JSON-LD	JavaScript Object Notation for Linked Data: a method of encoding Linked Data using JSON http://json-ld.org/
Kafka	An open-source message broker developed by the Apache Software Foundation.
LAS	A public file format for the interchange of 3-dimensional point cloud data between data users.
LSID	Life Science Identifier: I3C and OMG Life Sciences Research (LSR) Uniform Resource Name (URN) specification
MANiDA	Marine Network for Integrated Data Access (MaNIDA): e-infrastructure to support discovery and re-use of marine data from data providers in Germany http://manida.awi.de/
MarineID	Registration and authentication services for selected marine data services including SeaDataNet and EMODnet
MCP	Marine Community Profile: ISO19115 profile developed by Australian Ocean Data Centre Joint Facility (AODCJF) for marine data
MIKADO	Java-based software tool, for creating XML metadata records for the SeaDataNet directories EDMED, CSR, EDMERP, CDI and EDIOS.
MMI	Marine Metadata Interoperability Project https://marinemetadata.org/
MNF	Marine National Facility is owned and operated by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) http://mnf.csiro.au/

MPA	Marine Protected Area
MSFD	European Marine Strategy Framework Directive (2008)
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/
NiFi	Apache NiFi: an integrated data logistics platform for automating the movement of data between disparate systems https://nifi.apache.org/
NVS	NERC vocabulary server http://vocab.nerc.ac.uk/
O&M	Observations and Measurements: OGC standard defining XML schemas for observations, and for features involved in sampling when making observations
OceanSITES	Global system of long-term, deep water reference stations measuring a range of variables and monitoring the full depth of the ocean http://www.oceansites.org/
ODP	Ocean Data Portal: data discovery and access service, part of the IODE network http://www.oceandataportal.net/portal/
ODP	Ontology Design Pattern
ODSBP	IODE Ocean Data Standards and Best Practices Project http://www.oceandatastandards.org/
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/
OWL	Web Ontology Language: a family of knowledge representation languages for authoring ontologies https://www.w3.org/OWL/
PANGAEA	An Open Access information system aimed at archiving, publishing and distributing georeferenced data from Earth system research https://www.pangaea.de/
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/
PROV	W3C model to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web http://www.w3.org/TR/prov-overview/
PROV-O	PROV Ontology (PROV-O) defines the OWL2 Web Ontology Language encoding of the PROV Data Model http://www.w3.org/TR/prov-o/
R	Programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing https://www.r-project.org/
R2R	Rolling Deck to Repository: a US project responsible for the cataloguing and delivery of data acquired by the US research fleet.
RabbitMQ	Open source message broker software that implements the Advanced Message Queuing Protocol (AMQP) https://www.rabbitmq.com/

RDA	Research Data Alliance: international initiave to accelerate and facilitate research data sharing and exchange. https://rd-alliance.org/
RCN OceanObs	Research Coordination Network: initiative funded by the National Science Foundation (NSF)
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
SDN	SeaDataNet: EU-funded pan-European e-infrastructure for the management and delivery of marine and oceanographic data http://www.seadatanet.org/
SensorML	OGC standard providing models and an XML encoding for describing sensors and process lineage
SenseOCEAN	EU-funded FP7 project creating a highly integrated and multifunction in-situ marine biogeochemical sensor system http://www.senseocean.eu/
SiSSVOC	Spatial Information Services Stack Vocabulary Service: a Linked Data API for accessing published vocabularies http://www.sissvoc.info/
SKOS	Simple Knowledge Organization System: a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary http://www.w3.org/2004/02/skos/
SOOS	Southern Ocean Observing System: international initiative of the Scientific Committee on Oceanic Research (SCAR) and the Scientific Committee on Oceanic Research (SCOR). http://www.soos.aq/
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/
STOMP	Simple (or Streaming) Text Oriented Message Protocol: a simple text-based protocol, designed for working with message-oriented middleware (MOM) https://stomp.github.io/
Storm	Python programming library for object-relational mapping between one or more SQL databases and Python objects https://storm.canonical.com/
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
TAC	Thematic Assembly Centre for the MyOcean project. Responsible for collection and management of ocean observation data
THREDDS	Thematic Real-time Environmental Distributed Data Services: a web server that provides metadata and data access for scientific datasets, using OPeNDAP, OGC WMS and WCS, HTTP, and other remote data access protocols. http://www.unidata.ucar.edu/software/thredds/current/tds/
UNIDATA	A community of education and research institutions aiming to transform the geosciences community, research, and education by providing innovative data services and tools http://www.unidata.ucar.edu/
US-NODC	US National Oceanographic Data Centre (now the NOAA National Centres for Environmental Information) https://www.nodc.noaa.gov/
VOCRAM	Vocabulary Creation and Management project led by Australian National Data Service (ANDS)
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
WebEx	On-line web conferencing and collaboration tool
WebSockets	A protocol providing full-duplex communication channels over a single connection.
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
WoRMS	World Register of Marine Species: an authoritative and comprehensive list of names of marine organisms, including information on synonymy http://www.marinespecies.org/
WPS	Web Processing Service
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/