Grant agreement no: 654310

Ocean Data Interoperability Platform

Deliverable D2.2: Minutes of the First ODIP II Workshop

Workpackage	**WP2**	**ODIP workshops**
Author (s)	Sissy Iona	HCMR
Author (s)		
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)	EDIT	7 March 2016	DRAFT incorporating feedback from Bob Arko, Simon Jirka, Michele Fichaut, Anne Che-Bohnenstengel

Table of Contents

1 Executive Summary

The 1st ODIP II Workshop was held on 28 September - 1 October 2015, at the IBIS Alésia Montparnasse hotel, Paris, France, with logistic support from IFREMER. The programme was dedicated to kick-off of the ODIP II project involving more partners, reviewing the results and possible follow-up of the 3 ODIP Prototype activities, updates of the ODIP cross-cutting activities, introducing Model Workflows and Big Data as new subjects, and brainstorming on additional topics for ODIP II discussions and prototype developments.

The topics addressed were:

- ODIP Prototype Project 1: Establishing interoperability between SeaDataNet CDI, US NODC, and IMOS MCP Data Discovery and Access services, making use of a brokerage service, towards interacting with the IODE-ODP and GEOSS portals
- ODIP Prototype Project 2: Establishing deployment and interoperability between Cruise Summary reporting systems in Europe, US and Australia, making possible use of GeoNetwork, towards interacting with the POGO portal
- ODIP Prototype Project 3:Establishing a prototype for a Sensor Observation Service (SOS) and formulating common O&M and SensorML profiles for selected sensors (SWE), installed at vessels and in real-time monitoring systems
- Vocabularies/Persistent identifiers
- Data publication and citation
- Model workflows and big data

The Workshop was joined by 50 oceanographic data management experts from the 3 regions (Europe, USA and Australia) and IOC-IODE.

This deliverable reports on the organization, participation, proceedings and outcomes of the 1st ODIP II Workshop. The presentations are available from the IODE website.

The 2nd ODIP II Workshop is planned to take place at Boulder, Colorado, 2-5 May 2016.

2 Introduction

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

Building on the collaborative relationships developed during the first phase of the Project, the ODIP platform will organise four international workshops to foster the development of common standards and develop prototypes to evaluate and test selected potential standards and interoperability solutions for establishing improved interoperability between the regional infrastructures and towards global infrastructures.

The 1st ODIP II Workshop took place on 28 September - 1 October 2015, in Paris, France, organized with the help of IFREMER, Brest, France. The meeting took place at the IBIS Alésia Montparnasse hotel, Paris, France. The Workshop was dedicated to build on the outcomes of the first phase of the Project, further develop these and plan the future activities.

3 List of Participants

As in the first phase of the ODIP project and as part of the project strategy for wide communication, an extensive mailing list of more than 100 experts is maintained and is continually increasing, representing the ODIP project partners and their associated projects and initiatives. Following the same successful approach of the first phase of the Workshops organization, this list together with the ODIP website was used to invite participants for the 5th ODIP Workshop (and 1st Workshop of ODIP II project). Fifty (50) attendees from 9 countries took part in the 5th ODIP Workshop (10 of them participated remotely by "WebEx" video conferencing). They were:

Robert ARKO	LDEO, United States
Christian AUTERMANN	52°North, Germany (remote participation)
Jean-Marie BECKERS	ULG, Belgium
Sergey BELOV	RIHMI-WDC, Russian Federation
Justin J.H. BUCK	BODC, United Kingdom
Alberto BROSICH	OGS, Italy (remote participation)
Raquel CASAS	CSIC/UTM, Spain
Cyndy CHANDLER	WHOI, United States
Anne CHE-BOHNENSTENGEL	BSH, Germany
Kinda DAHLAN	UCL, United Kingdom
Francisco S. DIAS	VLIZ, Belgium
Paolo DIVIACCO	OGS, Italy (remote participation)
Jocelyn ELYA	FSU COAPS, United States (remote participation)
Michele FICHAUT	IFREMER, France
Christiano FUGAZZA	IREA – CNR, Italy (remote participation)
Oscar GARCIA	CSIC/UTM, Spain
Helen GLAVES	BGS, United Kingdom
Jonathan HODGE	CSIRO, Australia
Sissy IONA	HCMR, Greece
Simon JIRKA	52°North, Germany
Jonathan KOOL	Geoscience Australia, Australia
Alexandra KOKKINAKI	BODC, United Kingdom

Adam LEADBETTER	MI, Ireland
Thomas LOUBRIEU	IFREMER, France
Roy LOWRY	BODC, United Kingdom
Angelos LYKIARDOPOULOS	HCMR, Greece
Ana MACARIO	AWI, Germany
Sebastien MANCINI	Geoscience Australia, Australia
Youdjou NABIL	RBINS-BMDC, Belgium
Friedrich NAST	BSH, Germany
Elena PARTESCANO	OGS, Italy
Jay PEARLMAN	IEEE, United States
Francoise PEARLMAN	IEEE, United States
Leda PECCI	ENEA, Italy
Roger PROCTOR	UTAS, Australia (remote participation)
Lesley RICKARDS	BODC, United Kingdom
Dick SCHAAP	MARIS, Netherlands
Serge SCORY	RBINS-BMDC, Belgium
Adam SHEPHERD	WHOI, United States (remote participation)
Shawn SMITH	FSU COAPS, United States (remote participation)
Jean Marc SINQUIN	IFREMER, France (remote participation)
Shane St CLAIR	Axiom Data Science, United States
Rob THOMAS	BODC, United Kingdom
Charles TROUPIN	SOCIB, Spain
Mickaël TREGUER	IFREMER, France
Sebastien TREGUER	La Paillasse Ocean Project, France
Thomas VANDENBERGHE	RBINS-BMDC, Belgium
Rob VAN EDE	TNO, Netherlands
Matteo VINCI	OGS, Italy (remote participation)
Lesley WYBORN	NCI, Australia

The participants of the 1st ODIP II workshop represented a diverse range of expertise and good cross-section of the relevant EU, USA and Australian regional infrastructure projects and initiatives that are stakeholders of the ODIP II project.

4 Workshop Agenda

The first workshop of the ODIP II project will build on the outcomes of the previous ODIP project and will further develop these and the additional activities planned for the follow-on project. The scope of ODIP II has been extended to include other disciplines and new partners. The workshop agenda includes sessions to introduce the project to the new participants and also some of the additional themes and objectives outlined for ODIP II in the description of action (DoA).

As for previous workshops, the programme includes a dedicated session for each of the existing prototype development tasks. These sessions will provide a final progress report for each of these tasks including an opportunity to identify potential prototype extensions which will be developed as part of the on-going activities in ODIP II. The additional sessions included in the agenda will introduce some of the new themes added for the ODIP II project and will also be used to formulate further prototype development tasks for ODIP II.

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

While every effort has been made to have a coherent programme for the workshop it has been necessary to schedule some topics/discussions to accommodate those people who are participating in the workshop remotely from other time zones.

The overall workshop agenda was circulated to all ODIP partners by e-mail before the workshop and also published on the public ODIP website.

Workshop Sessions

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*

During the Workshop a further detailing through presentations took place which is given below.

Programme

	SESSION 1 - Introduction
	Monday, 28 September 2015
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*

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*

Wednesday, 30September 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

Attendees will have the opportunity to participate in smaller informal group discussions addressing the cross-cutting topics currently being addressed in the ODIP II project. These discussion groups will be 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

Thursday, 01October 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*

5 Workshop proceedings

All presentations are available at the ODIP website (www.odip.org) under the "Workshops" menu option. The presentations are hosted by IODE.

Reference documentation about the developments for the ODIP Prototype activities can be found at the ODIP web site.

Day 1 of the Workshop, Monday 28September 2015

5.1 SESSION 1 - Introduction

5.1.1 Opening

The 5th ODIP Workshop (and 1st of ODIP II project) was opened by Helen Glaves (ODIP coordinator) at 09.00 hours, on Monday 28 September 2015, at the IBIS Alesia Montparnasse hotel, Paris, France.

Helen Glaves (BGS) welcomed the participants, thanked the organizers and explained the logistics and the arrangements for the social event (group dinner) of the meeting. She introduced the agenda and the format of the discussions which followed the framework of the previous Workshops e.g. the plenary sessions and the working break-out sessions.

Then, Helen Glaves invited people to introduce themselves by telling names, Institutes and their role and expected contribution to the project.

5.1.2 ODIP II: Overview including aims and objective

Helen Glaves gave a short introduction of ODIP II focusing on the contact obligations of the EU partners. ODIP is in its 2ndphase which was recently funded by the EU. The proposal was submitted September of 2014 and officially started on 1 April 2015. Its duration is 36 months. The basic concept is to support multilateral cooperation on research infrastructures in marine science. It is a collaborative project between Europe, USA, Australia and related international initiatives such as IODE, GEOSS and POGO. The key objectives are to: continue and extend the activities of the existing 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; develop common approaches for specific aspects of marine data management e.g. vocabularies, formats, sensor web enablement etc; development of joint prototype activities including the further development of the existing prototypes to fully operational systems to demonstrate this coordinated approach; extend the scope of the project to include other domains e.g. marine biology.

ODIP will facilitate organized dialogue between key organizations in Europe, USA and Australia involved with the management of marine data through a series of workshops. ODIP II will seek to engage organizations and data infrastructures dealing with marine data in other regions e.g. Canada, Asia. The ODIP work plan includes 4 work packages: WP2 ODIP workshops to bring together international experts; WP3 ODIP prototypes to be developed jointly by the European, US and Australian partners for the purposes of demonstrating; WP4 to assess the impact of the implementation of the prototypes for existing infrastructures and to attempt to develop potential solutions. The co-ordination between the regional initiatives will be demonstrated through the development of several joint EU-USA-Australia prototypes that ensure persistent availability and effective sharing of data across scientific domains, organizations and national boundaries.

The project management structure has not been changed as it has proven to be successful in the first phase. The management is done by the Project (NERC/BGS) and the Technical Coordinators (MARIS), the Project Office that is located at NERC. The new EU Project Officer is Agnès Robin.

The EU consortium has been expanded from 10 partners to 19 from 9 countries. From the USA side, the R2R partnership was not specifically mentioned because unfortunately the NSF will not support the fund of supplement for R2R. Four USA partners (SIO, WHOI, LDEO, FSU) are actually the representatives of R2R Project who will continue the interaction with R2R and continue the efforts to solve it out and find other sources to fund their participation to the ODIP Workshops. There is additional interest from Canada (Ocean Network) to be part of ODIP. The Australian contributions and membership as well as the international ones were then outlined. Funding is still a challenge for the Australian partners.

The group then discussed other possible contributions such as funded EU programmes such ENVRIplus, Organizations such as JCOMMOPS, ERIC or other associations and initiatives such as Marine Data Harmonization Interest Group of RDA.

The project effort is distributed across the WPs with a significant amount of efforts focused on the prototypes developments.

The membership of the Steering Committee was discussed. The Steering Committee is a strategic management board for the project. Membership is not static and may be modified as the project evolves to involve other interests. It constitutes by the project coordinators and the WP leaders The NOAA representative is not yet defined (it will depend on the NOAA delegates participating in each Workshop). Current membership does not adequately reflect the project consortium as a whole – biology not currently represented. The partners committee is represented by all partners and the relevant contributors. They will meet at regular intervals and it was agreed with the EU to be twice a year during the Workshops because of limited funds for additional separate meetings (in parallel ODIP people will try to make benefit of virtual meetings outside the ODIP Workshops, in other meetings).

The resources are mainly allocated to the effort. Other indirect costs are relatively high on this project due to the need for a significant amount of travel for those involved. Some of which is outside Europe to attend the workshops. Acceptance of ODIP funds to attend other meetings will be assessed by the project Coordinators. Participation at the ODIP Workshops is a priority. The 2nd ODIP II Workshop is scheduled for May-June 2016 while the 3rd early on 2017. Finally, Helen Glaves gave some final financial details concerning the project initiation such as the pre-financing payment mechanisms which are managed not by BGS but by third party.

5.1.3 ODIP II: development of potential activities

Dick Schaap (MARIS), ODIP II Technical Coordinator welcomed everybody to the second phase of the project. He was very pleased that the group is together again to continue the successful ODIP approach and its synergies that came out of this. There are many projects now either new or future ones that are related to ODIP and make use of its work. ODIP is a platform that brings together ideas, progresses and developments, trying to tune and implement these in other projects using the linking-pinch principle as ODIP does not implement by itself. More analytically the approach is to develop interoperability between existing regional marine e-infrastructures in which ODIP partners are involved in order to create a global framework for marine and ocean data management. Through the Workshops the related topics of interest will be presented and discussed in order to identify potential topics for prototype projects. ODIP II will bring together expert developers and managers of leading regional and global infrastructures. In addition to the content wise approach there are many IT challenges as new standards are coming up continually such as OGC, ISO. Also the internet of things will bring large changes and opportunities in using the network but now the challenge is how to deal with the flow and access of the plugged data and metadata for example from the observations sensors using the SWE. ODIP II will also try to connect the existing data systems which used so far a bottom-up approach, will try to combine the data from different sources and turn them into information and knowledge by using the technology. Dick Schaap encouraged partners to bring new ideas that they may have to be accommodated either into the Workshops topics or into the cross cutting activities. A number of prototype projects will be formulated and taken into development, largely by leveraging on the activities of current regional projects and initiatives such as SeaDataNet, EMODnet (EU), IMOS and AODN (Australia), R2R, US NODC, UNIDATA and US IOOS (USA) and in dialogue

and direct cooperation with global initiatives such as IODE-ODP, GEOSS and POGO. This process requires strong interaction with the development activities taking place in the regional and global infrastructures. ODIP will function partly as a “think-tank” with agreed solutions carried forward by the related infrastructures for further development, testing and, if successful, wider implementation and operation.

Dick Schaap then overviewed the long list of possible ODIP II topics from which additional ideas may come forward through the Workshop discussions within the week. During the timing of the project, these new ideas will be distilled into the prototypes and cross-cutting activities. The first phase three prototypes (SWE, collaboration with the global data systems and the cruise reporting system), will be continued and extended within ODIP II. The NetCDF standardization (there are many flavors of it) and especially the cooperation with UNIDATA, the originator of NetCDF, will be included. The extension of use of the data statistical analysis UIG/DIVA software in applications other than the SeaDataNet will be explored as users more and more are interested about data analysis products instead of data themselves. The usage of controlled vocabularies is today a necessity in all platforms and domains and the ODIP activities on this cross-cutting activity will be continued. Along with the vocabularies the standardization of geographic marine names, including ocean basins, seas, seamounts, sandbanks and other sea features will be brought forward. Another item of interest is the harvesting from several data sources, automated aggregation with duplicates elimination, gridding with on-line visualization tools, quality control issues and prototyping aggregations of marine resources. The work on data citation & publishing as a mean to encourage researchers to open and publish their data will be continued. The WPS for processing of near real-time data streams, the clouds systems 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. Provenance of data from different sources for versions control or environmental/management/policy use becomes more and more important and common standards for capture of provenance information will be explored. The interoperability between operational marine observations systems, the multidisciplinary interoperability/System of Systems approaches such as GEOSS and GCI, Earth Cube BCube will be also checked because performance will be needed.

Dick Schaap then invited partners to bring forward new ideas and propose additional activities either by presentations or orally in order to wider the basket of the potential future topics.

5.1.4 Discussion

The group agreed a list to be maintained throughout the meeting where partners can add new items, comments and ideas. Communication with RDA on relevant topics has to be ensured so as the information to be replicated between the two groups, ODIP and RDA. The group then discussed for several key people from RDA interest groups (such as Adam Brown for instrument data) that could be involved. Also there are a number of EU projects dealing with sensors, SWE, etc and Jay Pearlman proposed ODIP to make known to the manufactures of instrument and platforms what it is needed. Helen Glaves noted that the representation of manufactures is missing in ODIP as well as in RDA and this linkage should be done. Thomas Loubrieu noted that the people from the “Ocean of Tomorrow” projects should also be invited in RDA. The group also discussed the engagement of the manufactures and how to make them to be interested as they want an acceptable business model and a certain guaranteed amount of purchases in order to implement specific standards and their customers (governments, met offices) cannot always offer such commitments. A common interest is needed by both sides. Another issue is that the ocean community is not yet in position to propose to them a unified approach, as different groups and the private sector use different standards and this will be one of the challenges for ODIP. Dick Schaap informed the group about a Workshop on SWE for Oceanography, at the Oceanology International 2016, organized by Eurofleets project, middle of March 2016. Partners from several projects and initiatives will discuss how to develop common marine profiles of OGC SWE standards. Manufactures are invited to share their views on adoption of these standards.

5.2 SESSION 2 - ODIP Prototype Development Task 1: plenary

5.2.1 ODIP 1: aims, activities and progress

Dick Schaap, ODIP Technical Coordinator overviewed the progress of ODIP Prototype 1 especially for the new partners. The aim is to establish interoperability between the Data Discovery and Access services of the three leading regional systems of Europe, US and Australia making use of the brokerage service and interact with the leading global ocean portals. The three systems are SeaDataNet in Europe, the US NODC in USA and for Australia the AODN. The targets to supply the data/metadata are the GEOSS and the IODE/ODP portals. It is used the GEO-DAB Brokerage System that is being used largely by GEOSS that aims to realize a community by community system approach. How it works: it takes the XML metadata output of the local system and converts it to the generic model that is populated then into GEOSS. The Brokerage has been developed in several projects in Europe and USA. CNR is involved in running the Brokerage System. The agreed approach was the three systems use the Brokerage and share with GEOSS and ODP at collection level and not at the individual granule level (of millions of records). The access to the granules is done by the local systems. The GEO-DAB Brokerage makes use of a detailed generic brokerage scheme. The initial work plan (Deliverable 3.2) was to start the connection with SeaDataNet and then apply the same steps to the other two systems. SeaDataNet is a distributed network with more than 100 data centres connected using OGC-ISO and INSPIRE compliant standards to the discovery and access service. Currently there are 1.8 million entries at granule level of data sets. The whole collection of granules is turned into 400 collections by aggregating by data type, data providers, discipline and geometric type (point, track, area). REST web service has been set up (IP – IP protected) to the brokerage which harvests in a dynamic way the XMLs and converts these to the generic xml which populate the GEOSS and ODP portals. In the current OAI-PMH interface due to a misunderstanding with CNR, SeaDataNet has been put as a domain in front of the other systems and looks like SeaDataNet is harvesting from them while it should be individual regional systems. This will be corrected. There is also a CSW service up, differentiation is needed on how to get only SeaDataNet. Dick Schaap then presented how the process model was implemented between SeaDataNet, GEOSS and ODP. The GEOSS portal harvests dynamically from the CS-W service and imports the SeaDataNet collections into the GEOSS portal, while the ODP portal harvests from the OAI-PMH service using jOAI. The SeaDataNet collections are now included and maintained in both the GEOSS and ODP portals. Triggers exist so as if something change at the source, then the system automatically is updated through the chain to make sure that the population of the content is up to date. He then showed how users can use the global portals to discover these collections and through dedicated URLs can drill down to the SeaDataNet portal for further detailing at granules level for formulation and submission of requests for access the data.

The same approached was used and implemented for the US-NODC. US-NODC provides services both at granules and at collections level (about 28 000 collections entries). The collection definition although different from SeaDataNet, fits for the purposes of ODIP. For US-NODC a collection can be data from an individual scientific project while in SeaDataNet, collection is an aggregation/envelop of data sets from one data center collected from many projects e.g. thousands of geological samples of one Institute at a whole marine region. There is a plethora of ways to support users to access the data: OPeNDAP Hyrax, THREDDS, Live Access Server, ftp and http links, which are provided as links to the metadata collections. Comparable links exist at granules level; the only difference is instead of www links (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. The Brokerage at CNR harvests the collections from the US-NODC (28 000 entries), they convert them to the generic XML and populate those forward to GEOSS, ODP portals. He showed the interface of the Brokerage, where on top there is the OAI-PMH interface where one can find the CSW interface with the specific links to the US-NODC catalogue entries.

The same was done for AODN. They use the GeoNetwork system. They also have collections, less and of different definition than the other two cases and provide these collections as CSW, OAI-PMH and OpenSearch endpoints for discovery. At present there are about 110 collections which change over time as more data and collectors come in. Dick Schaap then showed the web page of the IMOS Ocean portal interface with the 123 steps to get the data, the catalogues collections and the

individuals with links to the data. The principle is the same as in all three systems: harvesting of metadata collection, putting forward to the global portals, finding there the collections and then getting back to the regional portals where more details and data access is provided. AODN uses the Dublin core metadata profile and not ISO because the ISO 19139 does not give links to the data. He showed the Brokerage page of AODN which propagates to the global portals and how the AODN as well as the US-NODC and SeaDataNet data can be found from GEOSS and ODP. In case of ODP, the ODIP block of data should be renamed to the regional data systems names.

In conclusion the ODIP Prototype 1 has been successfully finalized. What's left is to a) check with CNR to specify the CSW URLs and not the SeaDataNet domain instead of US-NODC and AODN regions and data, b) to check if the numbers that the regional portals send are the same with the ones that being harvested and reproduced at GEOSS and ODP, also to check that the maps on global portals are representing correctly the areas of the collections, and c) report the Prototype 1 achievements in Deliverable D3.4.

Then Dick Schaap invited partners for questions or further suggestions. Jay Pearlman asked if it is harvesting metadata only or metadata seek and data access? Dick Schaap replied that the Brokerage is used only for metadata discovery but from the GEOSS portal and the discovery metadata you can go down to the granule metadata and then the data itself and the access is done not by the Brokerage but by the individual regional systems. Jonathan Hodge asked if there are any plans to in cooperate some data services endpoints such as OpenSearch or WMS and not only harvesting data. Dick Schaap replied that currently these services are provided for the individual entries but have not lifted yet at data interoperability and global level. It could be an OpenSearch or WMS portal building on the regions. Currently the work was only at discovery level using geo-brokerage in between, to unify and do the conversion. But it could be the future of this Prototype. The group then discussed several issues related to the data interoperability such as semantic interoperability or user registration. Thomas Loubrieu noted that the user registration issue was identified for the impact analysis at the previous Workshop e.g. to go from metadata to data interoperability, some management for the user identification regarding the data access restrictions is needed. Dick Schaap noted we started from metadata interoperability because it was easier. Data are not always directly usable because there are different formats, different perception, there is not semantic interoperability, and sometimes users are not satisfied when they have direct access of data from different data centers because they simply cannot use them. Brokerage helps towards this aspect as it customize data for users.

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

5.2.2 ODIP 1: report on impact assessment

Thomas Loubrieu (IFREMER), WP4 leader, introduced the outputs of the impact analysis of the first phase of the ODIP project. An inventory of cost/benefit of possible implementation the impact for each prototype was compiled. The document was sent to all partners prior to the final fourth Workshop of the first phase (April 2015) for updating and reviewing. It included all the valuable information of the brainstorming discussions during the prototypes sessions concerning the positive impact, the implication costs and changes that need to be implemented at regional level. Since the 3rd Workshop in Australia (August 2014) demonstration use cases and performance indicators were identified. The possible enhancement for the second phase of the project could include: to identify demonstration use cases to eventually collect success stories; to evaluate impact with indicators; to define simple targets for each of the prototypes so as to be easy to measure their efficiency; to have roadmap as a reminder of the activities throughout the project.

One feedback from ODIP 1 was that depending on the prototype the readiness of the activities was quite heterogeneous. Prototype 1 is almost operational, we are not far at all from having an operational portal merging all the metadata descriptions. But this is not the case for the 3rd prototype which is in a more research/innovation phase. To address the prototypes readiness level even if they are not yet operational and to measure the impact which was very valuable in the case of ODIP project as it brought people together to discuss, the schema of the concept of Readiness Levels of the Framework for Ocean Observing was used by Thomas Loubrieu. This schema fits well to the ODIP case. According to this, the first level is the concept of the Regional prototypes which is already

available, and its impact is that ODIP is valuable for research, innovation and technological expertise pooling during the discussions from experts. The second level, the pilot trans-regional demonstrations can be achieved by technology pooling (sharing software, standard profiles). At the end, the final target of this activity is to decide an implementation which would lead to set up an operational trans-regional infrastructure (not in ODIP).

For the impact assessment the following benefit/cost impact classification is useful. During the concept and regional pilot phase we do not expect much cost, instead the benefit is research and innovation pooling. During the trans-regional pilot phase there might be a cost as the regional systems may have to maintain and operate new infrastructures, software interfaces etc to manage transition but the benefit would be sharing the software (e.g. for prototype 2 is the GeoNetwork). Reaching the operational implementation phase, an end-user assessment is needed of what we expect from an end-user point of view, and is expected higher operation costs at data center level. But the expected benefit is to enhance end-users services and make operational costs at regional and national data centers levels lower. Two examples to illustrate this: in Europe there is the EMODnet check point EU project to assess the fitness of purpose of data services. One expected benefit of ODIP prototype 1 project would be to define metadata that contain the quality information required to assess fitness for purpose of the data sets. Another example is: in order to enhance the provenance information in observation metadata, prototype 3 could propose common implementation of SWE so that provenance information is homogeneously encoded at trans-regional level in a pilot.

As a summary, demonstration use cases should carefully be defined so as to have success stories at the end and be able to report these success stories. Also, simple/accurate targets should be identified to show user's or operators benefit.

Regarding prototype 1, the targets are to populate GEOSS and ODP with EU, US and AUS observation datasets description records. The work is concrete and the status is "toward operational implementation". The CSW services are working well and a demonstration use case was drafted to support CCAMLR to establish a MPA in the high seas of the Southern Ocean. Thomas Loubrieu invited partners for additional ideas for other use cases. Some performance indicators were also drafted to quantify the number of datasets added to ODP and GEOSS thanks to ODIP. The numbers will include contributions from all three regional data systems, SeaDataNet, US-NODC and AODN.

Finally Thomas Loubrieu concluded the results of impact analysis of prototype 1 by presenting the implications for ODIP II. Some of them are cross-cutting such as references services (vocabularies, further population of EDMO). The conclusion was that maintenance and upgrade of NVS and EDMO was needed at European level, mapping with NVS and EDMO was needed at USA and AUS level. Another conclusion of the impact assessment was that the federation of identity would be of interest to work on for interoperability on data access. At EU side, there is collaboration with EduGain (used in GEO) and Marine-ID (developed for SeaDataNet). The datasets description standards profiles were also identified as impact, with ongoing work within RDA working groups, IODE/ODS process there is room for collaboration, especially the issue of obsolescence management (deprecation and supersede) and the granularity of the datasets. Finally, the last impact identified was the operationality of the brokerage service which is operated now by CNR. At the previous Workshop it was discussed to extend the metadata interoperability of prototype 1 to the semantic interoperability with proper connections of metadata formats with vocabularies. Also to extend metadata interoperability to data access.

Thomas Loubrieu invited partners for comments and asked for other use cases to be used as success stories. Justin Buck commented that there are scientists who do not know that they can find combined datasets to ODP neither that ODIP is contributing to that. Dick Schaap replied that this is GEOSS and ODP mission not ODIP itself. These data portals propagate the populations of the datasets and create one shop stop. ODIP aim is to make that work. The GEOSS, ODP should do more to inform the community that the data are there but still users cannot easily find what they are looking for. Thomas Loubrieu commented that the users can benefit from ODIP work if they are trans-disciplinary users otherwise they go to the thematic portals and the Kim Finney's MPA case is such an example of trans-disciplinary usage. Dick Schaap replied that this is what the prototype does, helps users to find potential relevant data, and brings potential customers to potential data and not to deal with the data individually. Youdjou Nabil commented that data workflows linked with data services would help, not

only human discovery of data resources. Dick Schaap agreed that this was his initial point, make resources discoverable and bringing them together is happening now but it does not always help many users, further services are needed so as users to find data of common formats, processed data, and more generic products along with the data. Youdjou Nabil asked if the minimum metadata are provided by the data providers, and Roy Lowry noted that we use the ISO 19139 standard which has more high mandatory information content than the Dublin Core profile. Thomas Loubrieu stressed that for all the above reasons we need to find such use cases in order to demonstrate all these. Dick concluded that more steps are needed to make data more usable because usually users cannot find what they want at the big portals and machines are needed on the top of these portals to digest, to process the millions of data and make them more easily available.

5.2.3 Discussion

Dick Schaap introduced the discussion about where prototype1 should go in the second phase of ODIP. So far, the prototype was focused on metadata only, but there is need to go deeper to the data brokerage and services so as to make data more easily available and later, on top of them to have product services. The first step is to make it fully operational and become fully dynamic so as new entries to propagate through the system. The next step is start exploring the data brokerage and the services on how to help users and take part of the work load for them by using machines/robots to harvest and aggregate data. The group then discussed the issue of metadata. Jonathan Hodge commented that it would be of interest to check the question of bringing closer the document focused metadata approach with the linked data approach, to check the linkage of provenance management systems with metadata somehow and if the provenance metadata could replace the lineage elements in metadata records. Dick noted that from the data managers perspective we need more metadata for users while the data providers cannot always provide these metadata because they do not have these especially the historical ones. In ISO 19139 the mandatory fields is such a compromise while in SWE we can overcome this compromise as a lot of metadata are being set since the beginning. It is not only a technical issue to join the metadata that are automatically produced by systems like Argo with the data but putting together metadata from many different sources is related mainly with governance issues Dick commented. Thomas Loubrieu noted that this is similar with the data ingestion systems e.g. how to streamline inputs from providers. Rob Van Ede noted that prototype 3 and 1 could come together and evolve to something bigger. It could be services of data generation and connection to metadata and not only static documents production, Jonathan added. Sergey Belov commented that the creation of concrete federate search facilities needs further discussion because harvesting of metadata is a nightmare! Dick agreed that still there are many problems to the automated metadata harvesting. About the semantic interoperability in the ingestion systems and other cross-cutting activities, Roy Lowry mentioned that at the moment AODN uses the DCMG science keywords, SeaDataNet uses the discovery vocabulary, there is a partial mapping between them, but the GCMD is not mentioned yet and although it is too coarse it is not sure that it is drifted away.

Dick Schaap added that a follow up activity of one of the elements of prototype1 could be the harmonization of vocabularies as so far in prototype 1 we deal only with metadata and use the vocabularies as they come. There is yet no harmonization of the vocabularies at GEOSS or ODP level. This could be a small but important step to make although still at metadata level.

Rob Van Ede asked if we know how many of the discussed metadata are covered by INSPIRE. Dick replied that ISPIRE is an EU directive for harmonization of geospatial data, making them discoverable and accessible, and from the discussions with the research team working on INSIRE it came out that it is not only SeaDataNet but also the nations should report at and populate INSPIRE.

There is a Marine Pilot which has adopted the SeaDataNet vocabularies (P01 and P02 for the marine domain). Work is done with the INSPIRE team for a win-win situation so as EMODnet that is a policy approach to become fully INSPIRE compliant and get the stamp of fully quality controlled accepted by INSPIRE. This would be a success for EMODnet/SeaDataNet and at the same time a success for INSPIRE team since the marine community in Europe follow the same INSPIRE rules and used as an example for other communities. The only remaining issue is not vocabularies but the data models, since INSPIRE is defining some data models and further work is needed to become compliant with

that models. But INSPIRE catalogue does not have all the metadata. They are also set-up a data portal for Europe (a big discovery service). Thomas Loubrieu added that in RDA there are some groups for some advanced data registries (description of feature types) able to describe data types in registries so as everyone can access the data, visualize it, or having the metadata information.

Dick Schaap concluded that already new things came forward and during the Workshop sessions more ideas will come for the continuation and extension of prototype 1.

5.3 SESSION 3 - ODIP Prototype Development Task 2: plenary

5.3.1 ODIP 2: aims, activities and progress

Friedrich Nast (BSH), ODIP II prototype 2 leader, welcomed the participants in session 3 and introduced prototype 2 which was led by Bob Arko during the first phase of ODIP. The advantage of this prototype is that it is very much focused in one theme, the CSRs. The partnership, discovery and access of the CSRs are possible in all the three continents: in Europe through SeaDataNet, in USA through R2R and in Australia through the MNF system. The goal is to bridge the three regional systems and it seems to be trivial. Nevertheless there are a lot of things to overcome. Before giving the floor to the presenters of the session he recalled the evolution of the CSRs. Initially it was the ROSCOP forms on paper which were sent by polls to ICES. The next version was a “word” version but still it was free text to fill in. During the EU/SeaSearch project an on-line system was built using web technology. The next one was in XML and GML versions. Today we are on the way to get automatic generated CSRs (it is still a dream!). Nationally in Germany there is the project MANIDA (BSH) to help chief scientists to get automatic CSRs. At European level there is the Eurofleet project to generate automatically the CSRs from the ship system. Still there are some problems with the event logging that have to be solved but much has been published already. The next step is to harvest CSRs from different originators and then go from the three systems to a global portal (POGO) to discover the CSRs and access information. The best next step would be to link the CSRs with the data through the metadata. There will be further discussions on the possible next steps during the impact assessment and the dedicated discussion session.

ODIP 2 development task: progress and results

Bob Arko (LDEO), gave an overview of the results of ODIP I from the U.S. side. He reported that there is progress since the Liverpool Workshop. 130 new CSRs from U.S. vessels have been published so far to POGO, this was the end goal from the U.S. side because there was no much activity on that. Two vessels were selected for that and BSH was very helpful. One good outcome of this effort was the advantage that came out from the mapping of R2R terms with the EDMO codes and the SDN vocabularies of ports (C38) and devices (L05). He showed an example of a recent cruise of Kilo Moana, with all the included information such as environmental sensors used. The next steps include: 1) publish remaining older cruises (~4600) and the new cruise routinely going forward quarterly or annually (~400 to 500 every year, part of the R2R mandate and depending on the funding), 2) improve cruise records by populating Sea Areas using the C16 vocabulary, the P02/P03 Discovery Parameters, at least the P08 Disciplines, also detailed cruise abstracts and all scientists (not just Chiefs and Co-Chiefs) plus ORCIDs attached to every NSF funded investigator so as hopefully to be linked with the data publications records, and 3) upgrade the GeoNetwork portal using the last release of May 2015.

Bob Arko concluded that good progress has been made from the US side and workflows now are working to publish in POGO.

Lesley Wyborn asked how R2R gets people to fill the ORCID information if it is poor. Bob Arko replied that the only thing they do is that given a R2R cruise id there is an ORCID attached to it. Scientists have to fill in the appropriate information. There is a lot of interest (and ODIP is contributing to that) and R2R will focus on this the next 3 years so as the thousand of scientists in the R2R catalogue to see their cruises attached to an ORCID and then attached to DOIs. Ana Macario asked if R2R would

be a trusted entity for issuing ORCID for those scientists who do not yet have ORCIDs. Bob replied that R2R does not mint ids. Only a link to ORCID is needed and some guidance about the minimum information required by scientists, not just first and last name. Helen Glaves commented that BGS is now recommending the staff who is registering to ORCID what minimum information to put in their profiles such as scientist name and Institution name.

The discussion will be continued later on the session for the identifiers.

Roy commented that C16 is poor compared to C19 and that R2R should use C19 vocabulary for the Sea Areas. SeaDataNet CSR uses C19, C19 includes C16. Any kind of spatial serving is automatically connected with WMS, WFS and is included in the C19

Concerning the Australian progress, Dick Schaap commented that there is a lot of mapping between the Australian Organizations and EDMO. Sebastien Mancini noted that there is a lot of work on vocabularies but not much work by MNF to map the cruise reports. Bob Arko added that MNF uses the MCP profile to publish ISO records and probably MCP can be used for publishing and not CSRs.

Friedrich Nast asked Jonathan Hodge to be the link with the AUS side.

CSR harvesting: update on progress

Anne Che-Bohnenstengel (BSH) reported on the status of the CSR harvesting. She first explained that until the end of 2014 the CSR submission is possible using either an online Content Management System (CMS) or sending XML records via Email or FTP. Since the beginning of 2015, there is in addition a weekly CS-W harvesting of the CSRs from the connected centres (every Tuesday). The requirements for harvesting are: a) creation of CSRs in ISO 19139 format using MIKADO or other house software, and b) implementation of OGC CS-W service at the data centre. The workflow is as follows: after the XML records have been put on the CS-W server, the data centre contacts BSH and then the harvesting will be tested. If it works, the records are entered into the entry database. The next step is the quality control. There are automatic checks on mandatory fields and vocabularies and if these basic requirements are fulfilled the records will undergo further manual/visual checks on the contents. In case of inconsistencies the data centre will be informed to make the necessary corrections. When the CSR is valid, the second step is the insertion into the master database, the central CSR inventory for publishing. Each record has an unique BSH identifier and a local identifier which is defined by the collating centre and thus is also unique in combination with the EDMO code of the collating centre. Comparison is done between existing and newly harvested records. If the record already exists it will be updated, otherwise it will be inserted as new entry in the central database. The records of the central CSR inventory are automatically published at the SeaDataNet and POGO websites. Anne then showed an example of the POGO website with the recently published USA and one AUS records. She also presented the harvesting statistics (new and updates) since the beginning of 2015 for several data centres (IEO, OGS, HCMR, IFREMER). There are 3419 updates from IFREMER since the beginning of 2015. Anne explains that the number of update refers to every update of existing records. The present work nconcentrates on the monitoring of the harvested records from the connected centres. Data centres can check their own records after harvesting to find out if the content is OK (track charts, etc). All the harvested records are also available at the BSH GeoNetwork website (http://seadata.bsh.de/geonetwork-sdn/srv/ger/find) and can be downloaded from there in ISO19139 format.

Next steps include: a) connection with more partners, the next candidates are the Belgian Marine Data Centre (BMDC), BODC(UK), Marine Institute (Ireland), and hopefully R2R (USA) and Australia, and b) improve QC procedures for the harvested records for plausibility check with status display on monitoring portal.

Ana Macario asked if all the new CSRs have track line geometry. Anne replied that all the CSRs submitted in the new format can provide the cruise track in GML. Roy Lowry commented that BODC is planning to improve its system and put geometry to CDIs but not into the CSRs yet.

Partnership for Observation of the Global Oceans (POGO)

Lesley Rickards (BODC) presented POGO, it is a consortium of major oceanographic institutes worldwide, represented by their Directors. It is a forum in which Directors meet together annually to discuss issues of mutual concern, interest or matters of influence. It is supported by member's fee and grants by charitable foundations. The goal is to promote the completion of a sustained, integrated, global system of ocean observations for the benefit of society, without duplicating efforts just facilitating people being together, so POGO is a high level activity. The membership is 33 members in 19 countries (there are some notable exceptions like Canada, New Zealand, or most of the African countries have not joined).

Promoting observations is one of the main activities to improve scientific knowledge and interpret scientific results to policy makers, to enhance public awareness of oceanic issues and then provide capacity building in training and technology transfer. There is a structure that optimizes flexibility and provides links to the research community. The 3 pillars of POGO are (it reminds the EMODnet): Promoting ocean observations, capacity building and influencing policy. Promoting ocean observations needs to promote the availability of the data collected. This has to be done in a global way and as fast as possible and this is the linking with ODIP. She then presented another view of how POGO works by supporting various observing systems (not financial but facilitate advocacy, etc), by linking with other partners (GOOS, SCOR, GODAE, etc) and working partnership with them and then providing a leadership role up to global systems like GEO. POGO also formed the Oceans United so as Oceanographic Organizations could speak with one voice where it is needed.

Lesley Rickards then introduced the Blue Planet activity of POGO. POGO is a Participating Organisation in GEO and led the creation of the Task "Oceans and Society: Blue Planet" in 2011, for inclusion in the 2012-2015 GEO Work Plan. POGO continues to be the lead organisation and point of contact for Blue Planet, and submitted a proposal for inclusion of Blue Planet in the next (2016) GEO Work Programme. New Vision and Mission adopted following 2015 Blue Planet Symposium in Cairns, to make sure that society recognizes how important the ocean is and is committed to the stewardship of the oceans for a healthy, safe and prosperous future for all. The Blue Planet activity is the possible area of collaboration with ODIP. Blue Planet brings together many ocean, coastal and inland water observation organisations and programmes, seeking to add value to existing work rather than duplicating it. Lesley explains the diagram that shows the flow from sustained ocean observations through data collection, data&info management and models towards products and services, societal applications and ultimately societal benefits, as well as the importance of capacity building (as the basis for everything else). There are different components of Blue Planet activity like "developing capacity and societal awareness" component (C1), "sustained ocean observations" component (C2), "Data access & visualisation" component (C3), etc. ODIP could play a lead role in the C3 component. Jonathan Hodge, the representative of C3, noted that "Data access & visualisation" is new for ODIP and that Blue Planet does not want to recreate or duplicates efforts but built on things and bring the groups together for the benefit of the society. Roger Proctor presented ODIP to the Blue Planet Symposium.

Part of POGO is the Cruise Programme Database that is the link with the CSRs and thus ODIP. It runs since 2007. Initially it was funded from the Census of Marine Life and NOAA but not anymore. Now there are funds from the EU/Eurofleets project to do the Cruise Programme for Europe but with efficient effort this could work for the rest of the world. The idea behind Cruise Programme Database is to have cost saving, efficiency and have people working more closely together. For example, if you need more deployments somewhere and you know that someone is going there, you do not need to take your ship, you can use the other's ships. Or built a database of where people have been and this is directly related with the CSRs. The main problem was to get the information from other and populate the database. In EU it works because of the project funding but not for the rest of world, even if it appears to be very simple information. The most difficult is to get into the system the geographical information. Non-public information (e.g. from USA ships two years ago) is an additional difficulty. Currently there are 2965 programmes from 2007 onwards for 20 countries and 60 ocean going vessels (more the 60 meters long). Apart from the link with the CSR database maintained by BSH there is also a Cruise Vessels Database run by EurOCEAN. Lesley stressed that the earlier the information is inserted into the system and more countries join, the more useful it would be.

Lesley Rickards concluded her presentation by presenting another area of interest of POGO that could be the possible third area of collaboration with ODIP: 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 by e.g. geography, parameters measured, length of the time series, etc.

Dick Schaap asked what the message to the ODIP group is. Lesley replied keep existing collaboration on cruise Programmes and for the future work more closely with the data themselves like model data workflows, big data, the Blue Planet.

Jonathan Hodge confirmed that and Jay Pearlman informed the group about the Oceanic Engineering Society Symposium, September 2016, in Monterey. Dick Schaap noted the Blue Planet is GEOSS initiative and there should be an EU Call on Blue Planet.

Friedrich Nast commented that Blue Planet looks like a blue print for IODE and its future. He also added that in Germany they use the Cruise Planning as a data tracking system as chief scientists report what they are planning or achieved using the CSRs but it is not the same with other parts of the world where scientists are not so keen to reveal where they go because of the pirates.

Thomas Loubrieu asked the relationship of POGO with GOSHIP. Lesley replied that GOSHIP is for particular ocean sections and GOSHIP is a subset of POGO.

Sebastien Mancini asked what the tool to populate the Cruise Information database is. Lesley and Dick replied that it is very simple to do so, using the MIKADO tool.

5.3.2 ODIP 2 report on impacts assessment

Following the same methodology as for the ODIP Prototype 1 during the morning session, Thomas Loubrieu reviewed the impact analysis results for Prototype 2. The target is to populate POGO with CSRs from the regional systems USA, AUS and EU. The EU system is already connected to POGO through the BSH, work has to be done for USA and AUS. Kim Finney identified and drafted demonstration use case for the Southern Ocean Observing System (SOOS) for improving the information on what cruises have been done in this area. Sebastien will present later what is available in the area under Session 11 on SOOS Field Project Portals (paragraph 5.11).For the performance indicators, the available number of cruises in each regional system have been checked. At the end of ODIP first phase there are 7250 EU cruises, 1229 USA and 10 AUS cruises. A decrease to the EU entries compared to the October 2014 is because some ICES and BODC records in the old format had not been checked then for duplicates. Today these duplicates have not been inserted into the database. The implications for prototype 2 identified during the brainstorm sessions and reported in the Deliverable 4.2 are: a) on the standards and profiles used to transmit the CSRs and specifically for the ISO19139 format it was identified that the vocabularies references should become more matured using anchors-linked data and use GML versions. Also ISO19155-1 should become compliant with ISO19115-2. SeaDataNet3 project or POGO can deal with this format change and b) at regional level it could nice to have POGO face-lift by BSH within ODIP II project. At USA two interfaces have to be maintained, one for the national services and one for POGO. For AUS to federate additional cruise summary reports institutions, other than CSIRO. As a general implication, POGO should manage vessels with length less than 60m. Lesley Rickards said that POGO is doing that for EU and could extend to smaller ships. Roy Lowry commented that it is easy to do for ships that are politically agreed to POGO, just add these ship names lists to POGO lists and then BSH can easily make the CSRs for these ships part of POGO. A simple email with the ships list to BODC is enough. Dick noted that the Eurofleets R/V database can be included. This would enlarge the scope of POGO as this database includes additional information such as ship capabilities, mass, etc.

Bob Arko noted that there will be potential implications for prototype 2 by the change of the ISO19155-1. The groups asked if it 19155-1 or 19155-2. Roy noted that ISO 19115 is a content standard not a schema, 19115-1 is a list of fields and elements names. ISO 19139 implemented 19115 as XML schema based on GML. It is actually what happened with SeaDataNet. The group then discussed about the ISO changes, extensions and the next generations of standards. A critical question is whether it will end. Up to now we handle only with discovery metadata and not usage metadata to describe usage and provenance. CSR does not carry this information. But O&M and

SensorML do. The new versions of 19155 also does. ISO is being upgraded and this will have implications to the whole chain. CSR has to be upgraded as it includes ISO standards. The point is not to create different standards but common pathways should be followed in the future to avoid interoperability problems.

Thomas Loubrieu concluded and invited partners to identify other useful demonstration use cases.

5.3.3 Discussion

Friedrich Nast summarized discussions so far about the future of CSRs that includes more partners, federated systems, more inputs from USA and NOAA and invited partners for more ideas about the future of CSR using ODIP II project to formalize advances commenting that the future includes mixing of cruise records with descriptions of how measurements and chemical analysis were done.

Dick Schaap replied that Bob's idea on data linking to go from the cruise to the data at granule level (using CDI or other granule). If the data are included in the discovery system, the CSR could be the link to the data. IFREMER has already started with the work, and other SeaDataNet partners will follow. R2R is doing the same, cruise is an intermediate way to help people to go to data and the scientists are involved in this through the publications. It is feasible, it is a short term activity. In CDI, this work has been done, but from the other side, e.g. in CDI .you can find the cruise and later on you can go from the cruise back to the data.

Friedrich asked how far the ODIP partner's systems are to automatically generate and add new information into the CSRs. The goal is that the chief scientists do not have to fill-in any forms in the future. As it was mentioned in the last SeaDataNet Plenary meeting, collaboration with manufacturers of ship systems is needed in order to generate CSR by pressing only a button. Bob Arko replied that this is done in R2R for the environmental sensors. Spain, Italy (OGS) are doing this through Eurofleets project. Friedrich added that CSRs can be used not only as data tracking systems but the underway CSRs are also a tool to find out the research vessel activity in near-real time.

Dick Schaap noted from the discussion so far it seems that there are in place all those ingredients to make a nice project e.g. how to make CSRs more efficient, how to get more data in, more ships and operators, getting access to the data, there is also POGO and the Blue Planet initiative for transforming data into knowledge and information. All the above are components of a good project. From his point of view what we could try to do is not the same prototype but an ambition for a prototype use case for a specific area in the ocean, at an area to demonstrate how all these tools and information that we have can be used together as a good case for information and knowledge.

Lesley Wyborn noted that there are many coastal activities without CSRs which are not captured by this ambition. How do we structure the scientific observations from platforms other than cruises, in coastal measurements? How do you define what you are observing and measuring? Jay Pearlman asked if the question is how we address citizen science. Lesley replied that citizen science is for shallow waters. Thomas said that small experiments at sea with no CSRs do not reach the data centres and work is needed for that. Roy said that standards have been developed for CSRs but are being extended to other platforms. ICES is an example which started from ships and it now covers a wide range of platforms. He suggested a metadata standard to describe the deployment of a data collection platform. Friedrich mentioned that the Indian Ocean Experiment 2015-2016 could be such a use case that incorporates many platforms.

Friedrich closed the discussion session by informing the group that the Canadians will adopt the SeaDataNet infrastructure to their system so a lot of new cruises will be inserted into the CSR system. Helen Glaves thanked the new leader of ODIP prototype 2 for his excellent chairing.

5.4 SESSION 4 - ODIP Prototype Development Task 3: plenary

5.4.1 ODIP 3: aims, activities and progress

Jonathan Hodge (CSIRO), ODIP II prototype 3 leader, introduced the prototype and the latest inputs. These are actually a number of experiments from different Organizations on SOS and OGC, examples on OGC services for performance time series services, etc. A key part of this session will be the discussions for the future.

Sense OCEAN Developments

Alexandra Kokkinaki (BODC) presented the developments of BODC in the SenseOCEAN for retrieving biogeochemical data in a standard format from sensors. Autonomous ocean observation is massively increasing the number of sensors in the ocean. Best practices for data management need to evolve to ensure that key metadata and technical data from novel sensors are never lost, data are efficiently processed, archived and delivered in a seamless way. In order achieve this we need interoperability and a pre-requisite for interoperability is to apply standards from the sensor through to delivery. A problem is that sensors are attached to legacy platforms that cannot transmit OGC SWE formats such as SensorML. A solution could be a transmission of a unique id of the sensor that would be referenced to a NERC linked data server and provides RDF, SensorML or JSONLD and apply standards to the file format, on how we will query and show the data and apply standardized ontologies and languages. To do that, BODC asked manufacturers to provide data. Alexandra then presented an example of the received data and their technical characteristics. Then they modelled the data. In linked data and semantic web community it is beneficial to re-use what other expertise already developed and not re-invent the wheel. The SSN (Semantic Sensor Network) ontology 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 extended the SSN ontology to serve its sensor descriptions. An example of a wind sensor was given. It was used the Library for Quantity Kinds and Units, the Dublin Core ontology, and the Good Relations ontology to describe serial numbers, manufacturers, make and models. They also used the P01 BODC Parameter Usage Vocabulary, P06 BODC Data Storage units and C75 vocabulary for Organizations. The ontology design was based on the received 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 received. Alexandra then explained the model, the URI design and the RDF content. The final step is to publish to SPARQL endpoint, RESTful interface, ELDA, and Mash up application. The future tasks are: RESTFul Publication; Metadata publication in VoID, PROVO (Adam gave some hints); Effective discovery (CKAN); Align with PROVO ontology, Link with O&M; Produce sensor descriptions in SensorML/Json LD; Create persistent Identifiers (pURL).

Sensor Web Enablement integration

Thomas Loubrieu (IFREMER) presented the SWE realization within the SeaDataNet project. The tasks were to provide a graphical editor for observations systems and to demonstrate with a 52°North application. The editor - a web application, is a flexible system and includes a drawing tool so as the data providers/scientists who make field observations can describe their observation system. The tool includes preloaded sensor models. The Sensor model descriptions are extracted in SensorML format from the EMSO sensor model directory (aka FixO^3 yellow pages, http://www.the .com/). The list of preloaded models is extensible. Descriptions of the sensor models are provided as “sml:typeOf” information within the sensor instance description. The user can drag and drop icons of sensor or platform model to create instances and can link them together to describe complex systems. Links are oriented and means input/output relationship, type of connection 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. It is free text information. Vocabulary references from URI (linked data) can be used. Auto completion is also provided. The export is offered in SensorML and it

would be interesting to explore with BODC the export in JSONLD. The tool is available online as demo: http://snanny.ifremer.fr//webgraphiceditorDemo/.

Thomas Loubrieu then overviewed the demonstration that was presented during the SeaDataNet plenary meeting by connecting two 52°North servers at OGS and IFREMER. These implement the SOS GetObservation with O&M as response format using the 52°North SOS server. A test with SensorML output was not performed. In the CDI records URIs were added that point to the 52°North Sensor Web REST-API. The result was a good looking 52°North client but for time series display only. At the CDI portal the REST-API URLs are transformed to human readable so that and the user can access the time series visualisations. Thomas concluded that it was very efficient to work with 52°North and this API. The idea of REST was in conjunction with the three standards was explored. For the future it is needed to access data from vertical profiles and trajectories, extend the demo for sensor and system descriptions, and re-merge with core SWE standards. Work goes on in: Eurofleet2, ODIP2, JERICO-NEXT, NeXOS, FixO^3, ATLANTOS, EMSO-DEV; Ingestion system, SDN3, ifremer/csic/ogs internal Business; and on the clouds.

Simon Jirka mentioned current activities of 52°North.

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 too much software available, most software is experimental, and most software requires extensive configuration. The FME is a data loading and manipulation tool (ETL tool) for translating data between several formats and do geographical manipulation between them. It supports more than 300 formats which is a good if data transformations are needed. It manipulates contents quality issues and no coding is required. But SOS is not supported out-of-the-box. Perhaps future versions can support this like it is possible for WMS and WFS. Rob then showed the visual tool with the input/output and translate tools, an easy and flexible tool to use and explained the method he developed. The first step was to map the data to the NVS vocabularies. Then JSON requests generated from the Oracle dataset and posted to SOS. For every request there was a response. He showed an example of a visual response and the overview of the dataflow. Current work includes the insertion of observation requests. The next future work includes cleaning up the workbench and publishing in FME store as (free) custom transformer (supporting the InsertSensor and InsertObservation operations). The far future work includes implementation of more requests and data retrieval as now much work is needed to visualize the SOS requests with GIS visualization packages.

Adam Leadbetter noted that there are R packages and the visualization of SOS requests is very straightforward.

The tool use the 52°North SOS implementation.

IOOS SOS Activities

Shane St Clair (Axiom Data Science), representing IOOS presented the work during the last three years in SOS activities. IOOS is a US federal/regional partnership for ocean data monitoring in US to enhance, organize, analyze, and provide access to ocean data and tools. It is the federal parent organization (IOOS) and 11 regional associations (RAs) for specialized issues in coastal areas in US. Prior to 2012, the OOSTethys project was the initial attempt to adopt OGC SOS standards. There was progress but scattered adoption, various implementations with differences in behaviour/responses. In 2012 a meeting took place between all regional associations to develop formal IOOS SOS application profile and software implementations. The decisions were 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. From 2012 to 2014 a SOS application profile was developed. But there was slow going after initial meeting, there were complex requirements (nested assets, feature types, etc.) and certain problems discovered only after implementations were

developed. Version v1.0 of the application profile was finalized in 2014. This included response templates for GetCapabilities, DescribeSensor for network and station asset types, GetObservation for time series and time series profile. It used the standardized vocabularies (CF via MMI, IOOS, etc). It was developed OGC CITE style test descriptions, and WSDD (web service description document of about 200 pages). All the documentation is available at the GitHub page: http://ioos.github.io/sos-guidelines/.

The effort developed by Axiom Data Science is the i52n-sos. It is a software implementation of IOOS SOS application profile v1.0. It is Java web application using a database at the backend. It uses 52°North SOS 4.x as core and includes support for multiple database management systems (PostgreSQL/PostGIS, SQL Server, Oracle, HSQL), the transactional SOS operations, SOS 1.0 and 2.0 compliance, an installer and administrator web interfaces, 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 that 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 implementation developed by RPS ASA (Applied Science Associates) is the 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). Due to THREDDS internals, 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 tools developed: 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 to SOS servers using sos-injector (many US sources including NDBC, CO-OPS, USGS, etc.), c) the sos-injector-db, to inject data to SOS from an existing database, d) the ioos-sos-compliance-tests, for OGC CITE (teamengine) tests for IOOS SOS v1.0 implementations, and e) compliance-checker, a Python tool to check dataset (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 IOOS catalog (http://catalog.ioos.us/). Beyond SOS there is also the Sensor Scalability Experiment (http://axiomdatascience.com/maps/ioos) for harvesting "all" available sensor data (mostly U.S., adding more) handle large data volumes. Hex binning is used to show trends at high zoom levels. There are about 90 million observations over 14 days in memory. It provides high performance statistical binning and analysis. The AOOS demo is available at: http://goo.gl/pFpRAR.

Shane St Clair finalized his presentation with the lessons learned: prioritise client library development - lack inhibits adoption, provide a simple data format option (many people want CSV, which can still co-exist with SOS), use a small group to propose drafts of standards (quick initial proposal, larger community can adjust proposal (or reject)), develop pragmatically (release software with basic functionality early, add new features prioritized by user demand, do not wait for complete implementation, software benefits most from being used!)

SensorCloud

Jonathan Hodge (CSIRO) gave a quick update on SensorCloud system. It is a times series data aggregation, ingestion and serving system, built of Java messaging getting system in the middle, and the configured data sources and signals presented as API on top of it. It is not an SOS implementation, but it is guided by some OGC standards such as O&M and a simplistic way for describing sensors (StarFL). It gets entries from a lot of different data types and moving platforms (e.g. sensors on animals, ships, etc). It allows the data collection from sensors but it does not require completing a full SensorML document. It holds information on deployments and can stuck tracks

through specific ids and serial numbers of instruments which is useful for QA/QC. There are some developments for the locations, there is the concept of RelativeLocation: a platform can have location and individual sensors can have relative locations (e.g. the anemometer is mounted 10m above the platform). For the calibration information, there is conformance test or field sensor calibration and the calibration event recorded. There are some changes in the structure of stream of 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.

MongoDB is used for data storage. It provides: document storage, flexibility in data types, GridFS, distributed storage for files (images), an aggregation framework (distributed computation of aggregations), geospatial indexes and queries (for mobile sensors), it is easy to horizontally scale, and there is experience within development team. Using V1 for 1.1. Billion observations it came out that scalar data is small in volume (index size large vs data) and nodes use a lot of RAM, performance degrades as data/ram ratio increases. V2 offers: Aggregated storage, multiple observations in a single 'document', smaller indexes for high frequency (faster than hourly) data streams, MongoDB aggregation pipeline unwinds results, and am improved sharing to balance data streams across nodes in cluster. Some technical details of the ingestion system and the Sensor Messaging Gateway (SMG) were then presented. The data sources can be a configurable generic polled file import (CSV, TSV, fixed width, FTP, HTTP). The existing library of data sources is: Campbell Scientific; Libelium; PACP (DPF/CSIRO); AgIsp (DPF/CSIRO); ROS (Robotic platforms); BoM/SILO. There can be custom data sources (Java, Phyton) and others.

He then gave an interesting example from the data point perspective: a Web JavaScript Streaming application using a STOMP interface in a RabbitMQ system which serves oyster heart bits in real time (20 points/sec (Hertz)).

The authentication and authorization includes: HTTP 'Basic' authentication; All SensorCloud metadata and streams belong to one or more groups; User 'roles' defined permissions for a group, e.g. John works on a project called Aquaculture. That project has its own private sensors, and all needs access to some private sensor data. The role 'aqua_researcher' allows read access to all data in the 'public' group and read access to data in the 'private_aqua' group.

The system now can handle model workflows and provides raw sensor data by reference. The model services provides a framework to: 'Wrap' existing models as web services, model inputs and outputs described in REST API, and currently support for Keplar, R, Python, Java. The gridded data services are: THREDDS for data management for netCDF data sets and provision of catalogue and services (WMS, WCS, OpenDAP, NetCDFSubsetService, and HTTP)

AWI – ODIP II Prototype 3

Ana Macario (AWI) presented the AWI activities relevant to prototype 3 and explained that the Computing and Data Centre group has a tradition on developing different information systems for data acquisition and especially on board of the Polarstern. Only the last couple years a systematic effort started to adopt OGC standards relevant to prototype 3 as means for supporting the automated data flow from sensors to PANGAEA. Mostly they are focused on devices and sensors on board of the two main German vessels Polarstern and Heincke as well as the land-station Neumayer, trying also 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. The current statistics are: about 100 ship-mounted sensors, about 500 sensors from other platforms, which give about 10 M measurements per year in distinct data format. Every information related to the sensor is being stored in a near real time database (in PostgreSQL) and can produce SensorML 2.0 and there is also an interest from scientists for versioning. Also a series of and REST-based access interfaces have been developed to put the information of users interest at a dissemination level. In terms of monitoring environment, it is important to keep the range of each sensor stored in the database thus the AWI SensorML profile includes range values for each

sensor. These are used in automated QA/QC procedures (e.g. measurements out of expected range are flagged) and in monitoring dashboards. Ana then showed an example of a monitoring dashboard to illustrate the pragmatic view of SensorML, where given certain range values they can get an alert when temperature reaches a certain value. For accessing near real time data they offer at the moment web services (encoding JSON, tab-delimited) with track line geometry and atomic dissemination for selected sensors. Support of the O&M standard is planned.

The Planned ODIP II activities include: a) adoption of SeaVox vocabularies for parameters (not a trivial task for PANGAEA) and EDMO manufacture vocabularies (which is not ready yet), b) share the AWI SensorML 2.0 profile in a GitHub repository/ODIP, c) O&M encoding for selected sensors. These will be engaged in the activities of Eurofleets II, FixO^3, AtlantOS.

There will be a tight cooperation with 52°North which will include the installation of SOS 4.x core server including the Sensor Web REST-API extension. In particular AWI is interested in the integration of the legacy data infrastructure and information systems. For that purpose “connectors” have to be developed as the data are very complex and in many formats. There is interest to use the O&M encoding and the SOS not only for near real time data, but also go back in time, tap the information and disseminate it into PANGAEA. This will be a challenge and will not be trivial. It is expected about 10 million measurements per year from near time data and about 10 billion measurements from PANGAEA.

Jonathan Hodge commented that there is a lot of activity within several projects that could contribute in prototype 3 and we have to discuss how we will proceed in ODIP II.

5.4.2 ODIP 3 report on impacts assessment

Thomas Loubrieu (IFREMER) reminded that the target of prototype 3 target is to merge together resources from the three regions and this prototype is more a regional and regional pilots status category oriented. There are many results based on different implementations. Three standards were identified as good for the conceptual approach of what should be described in terms of observations systems and observations. Also the implementation of a RESTful JSON approach might be more efficient and would facilitate the development of Web clients that will be used on top of these services and the drafted demonstration use case that was identified did not achieve yet that. The decided performance indicators allowed to consider that depending on the type of SOS service, there are two different approaches, one is to have one ncSOS service per file (or per single granule) and the other which is being developed now, is to have a SOS services on the collection level of data sets (52°North, oceanotron, sensorCloud are doing this). The identified implications for prototype 2 are the standardization of the profiles of the observation systems and the observations themselves and specifically the restful API profiles. At regional level the impact is that work needs to be done on the implementation of the standards, but it seems that big steps have done since last Workshop at Liverpool.

5.4.3 Discussion

The group discussed possible ways forward.

Jonathan Hodge commented that the question is what the implications for prototype 3 are and how ODIP could proceed to bring the several efforts together. Either we continue an informal approach where the groups meet and present their activities and try to learn from each other or try to construct something more formal with targeted efforts. From the AUS side the challenge is the funding for an ODIP flavored activity. An option would be to spin up an Asia-Pacific-USA project or another ODIP-tagged activity and try to find some funds.

Dick Schaap noted that a report from the first phase of ODIP project is needed and agreed that during the last years big steps have been achieved from local systems which started from scratch. The report should describe all these developments. Although the ambition was higher at the beginning, now we can describe how ODIP will proceed as these systems are evolving and become more

mature and new groups are entering into the project such as 52°North. From now on the ambition could be a nice demonstration case to show how all these SOS services cooperate together and their overlapping as it was done in prototype 1 with the regional systems and the global portals. A more formal approach like Blue Planet initiative or other would be better of course as it would bring some funding.

Helen agreed that it would be a good outcome of the project to build this use case. The group then discussed several cases that could be good examples to be used as use cases, e.g. different SOS services, sensors descriptions, or working with manufactures. Another challenge is the leadership of prototype 3 by Australia in ODIP II. Helen Glaves noted that balance is needed between the three prototypes as Europe now is leading already the two of the three prototypes.

The discussions continued during the second day of the Workshop.

Day 2 of the Workshop, Tuesday 29 September 2015

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

5.5.1 ODIP prototype development projects

Each group gave a feedback on final outcomes and potential further developments in ODIP II.

ODIP 1

Dick Schaap (MARIS) noted that there is some remaining work for prototype 1 such as: a) separate name spaces for the three regional data systems of SeaDataNet, US-NODC and AODN, and b) check the numbers that are exchanged and are being harvested. He then summarized the issues that came forward from the yesterday discussions as possible next steps for the continuation of prototype 1:

- Make it more operational and fully dynamic because now it is more as demonstrator, so as any change at the source to be sure that is propagating to the system.
- The three systems are using their own vocabularies which are harvested and so far they are pushed forward to GEOSS and ODP as they are. So, a task is to check the harmonization of the vocabularies especially in the brokerage, semantic brokerage could be added by having ontologies between the three systems. This needs cooperation between vocabularies and brokerage people.
- Explore data brokerage as now we have only metadata brokerage, check what the existing projects are, what plans can be done with them and how far it can go since ODIP is leaning on leverage with existing activities. ODIP can formulate actions but needs other project to do part of the work.
- Learn on user's requirements. Check uses case to see what is the impact by the current federation, how far it can help users, and more is needed because users are not interested in discovery and access but more and more are interested in aggregated data services and added-value services.
- Check provenance and data profiles as the three discovery services give access to data but are still autonomous systems and we have to provide more information to users to know what they are looking at and what potential services are available.
- We are now harvesting metadata automatically but manual control is needed afterwards to check if the harvesting is OK and in many cases the results are not satisfactory. We may need to check the federated search which is another approach e.g. take the 3 bases and we do not bring them in one discovery service but an engineer looks the three and gives back the results. But this is completely different that GEOSS and brokerage is doing now, but we will check it, but we will look at.

Dick Schaap then invited partners for further ideas.

Data quality is an issue when combining data from different systems. Atlantos could be used as a use case where its data will be used both for scientific purposes (requirements for higher quality) and

operational purposes (feeding into models, Copernicus, US and Canadian model) and look what is the added- value when connect everything and what users want.

Semantic interoperability and how it could be approached (for example in case of parameters): each region uses different lists, SDN uses P02, AODN uses a mix of BODC and AUS vocabs, and US-NODC other (to be found). For semantic interoperability what is needed is the regional systems metadata to carry URLs for the parameters they are using, then we have the basis for semantic interoperability. These URLs resolve to a RDF document and then through translation from one list to another, cross-search could be done. A common vocabulary is not needed. The client chooses what list to use. Semantics will lead us to the data as well. Additional vocabularies (for biology: GBIF, WoRMS, ...) can be included in the model by mapping with P01 and by extended the data model. When the vocabularies are found and agreed, the next fundamental step is to markup them to the fundamental metadata documents.

ODP is moving to the same approach, it uses URLs but for the content mapping the technical group of ODP will need assistance. ODIP can assist it. GEOSS is in the process of handling semantics and can focus on marine coastal applications. GEOSS does not do any content harmonization to the brokerage. In coastal areas the biology community is different than the water community and GEOSS could focus there. This cross-domain homogenization would be a big step for ODIP.

Biogeochemistry from estuaries could be a use case, but yet it is not clear if ODIP should extend outside the marine domain or if it is feasible.

The group further discussed the users need (such as the MPA demo case) for geo-search at the granule-level and not only at collection level as global portals currently are doing. This is still difficult to realize when working with federated systems and metadata collection (and not point) search.

A link between EDMED and CDI references could make EDMED a tool for discovery at collection level. The user define the polygon, gets the list of EDMED, EDMED lists resolves down to the CDIs, each CDI is carrying a geometry and you can filter that geometry to the targeted interest.

Dick Schaap wrapped up the discussion and suggested not to create too high expectations that ODIP could not succeed. Semantic is a step that ODIP can take.

ODIP 2

Friedrich Nast & Anne Che-Bohnenstengel (BSH) summarized the first day discussions on prototype 2 work. The general comment is that the Cruise Summary Report is an integral part of the POGO infrastructure and that there is a strong role in multidisciplinary observations on board if the cruise has a summary or a survey of what has been achieved during the cruise. Also, from the data management view, the CSR can be used as a strong tool for data tracking.

The CSR cookbook and documentation for CSW-harvesting can be found at:

- CSR schema plus minimal requirements: http://www.seadatanet.org/Standards-Software/Metadata-formats/CSR
- Manual for CS-W: Deliverable SDN2_D92_WP9_CSW_harvesting.pdf
- GeoNetwork: ftp.ifremer.fr/ifremer/sismer/donnees/SeaDataNet_Software/CSW/geonetwork-sdn.war

In detail, the feedback received was:

- Connection to data: by connecting CSR with the CDIs e.g. using CSR as a discovery tool for accessing data
- Link to underway sensor data
- Extension of CSR schema/standards to include O&M or introduce next generation of standards: it is a big job to do that extension because the big buckets for parameters that CSR now have may be too rough to go to data (*and don't want to compare apples and peers !*)

- Extending CSRs to other platforms, e.g. small barks, gliders, mammals, or put CTD in turtles (*it is interesting to see ship codes for turtles!*) ….
- Include more ships less than 60m to POGO: it is an outstanding issue
- Blue Planet Mission
    - transferring data into knowledge: use CSR in some products (e.g. Caribbean) to show what is the potential of CSR
    - Find use case to see how these tools work together

From the EU side the next steps include:

- POGO face-lift
- Link CSR to CDI
- Spain and Italy continue with the automatic CSR generation – in general these tools to be applicable for other ships? Could it be documented and be a standard for next generation of research vessels! (helpful in case of ship system manufacturers)
- Introduce ORCID for scientists
- Introduce DOI for CSRs

In the USA the next steps include:

- Maintain CSR interface and continue populating. This will bring thousands of new CSRs.
- Include Ocean Area (C19!), parameters and more detailed abstracts
- CSW-Harvesting
- Submit CSRs for NOAA ships. It is an outstanding issue and contact with NOAA people is needed.

For Australia:

- Federate additional CSRs from other Institutes
- CSW-Harvesting

Friedrich Nast asked for any missing points.

The standard format of Argo floats includes at the heading a lot of the platform metadata (codes) needed for the CSRs. With the proper cooperation (Argo, JCOMMOPS, Copernicus) the CSRs then could be automatically created in real time mode. All the observing programmes use flavours of CF netCDF. But is there any global agreement to extend CSRs beyond research vessels? (It is a big can of worms). Gliders people ask BSH/CSR group for platform codes and they are open to do that. The group discussed if it is better to keep CSRs for R/V and not mix with other platforms or if there is really an interest by the community to use CSRs to find data (if so, will the cruise databases be fluid by floats and gliders?). The French NODC for example uses EDIOS and not CSRs to describe Argo floats. On the other hand, a cruise can include many platforms and by using the appropriate filter you can find the data of your interest. BODC case: they have a metadata system no matter what the platform is. They also maintain platform deployment metadata that are sent to BSH and BSH filter them out for inclusion or not into the CSR.

Identified Actions:

- A general use case to ask scientists what they want (**Action**)
- A specific use case to go deeper and ask in more details what is expected from ODIP, how they want to find what they need (N. Ocean, with fine grain search) (**Action**)
- Put WMS-WFS on top of CSR to make visible to other systems what CSR includes and for linked data purposes (**Action**).

The most common request in BSH/CSR is: give me the data from a specific cruise.

ODIP 3

Jonathan Hodge (CSIRO) noted that a decision has to be made on how prototype 3 will progress in ODIP II.

- Option 1: continue operating what we have and treat it like a RDA IG by reporting activities of partners and exchange ideas trying to align activities.
- Option 2: look for a topic that could be turned into an actual prototype activity and more structured.

There are challenges of coordinating different contributors of ODIP and challenges for creating some opportunities for funding. Dick suggested to keep the programme as is and to try to explore with partners (GEOSS, RDA) some funding opportunities for a joint approach.

At the moment in Australia there is nothing towards SOS or projects of common interest with ODIP domain that could push towards it. Australia could contribute to find some commonalities at the output such as defining some OGS, O&M specs able to be applied to different systems.

A question is, how this shall be structured and to investigate if there are issues with ODIP to have something less formal. Helen replied that the format of the activities is not an issue. Dick suggested that we continue what we are doing keeping each other informed but also look for the low-hanging fruits, look at commonalities, differences and if something is worthwhile try to do something together.

But there is currently no active SOS work in Australia. Roger Proctor suggested that another region should take over the coordination of this activity. But regional balance is needed. The group discussed re-scoping prototype 3 or creating new tasks where AUS could contribute such as sensor data, model workflows and big data. On the other side there is interest for SOS and SensorML in the other regions. Shane St Clair suggested to change the output to SensorML and explore interoperability formats.

Decision: prototype 3 will be re-formulated and instead of focusing solely on SOS, it will have a broader focus on interoperability, led by 52°North (Simon Jirka).

Alexandra Kokkinaki proposed to use ontologies to map to SensorML and to expose the sensor description in RDF. The group then discussed Alexandra's' proposal on how to get what we want from sensors.

Alexandra Kokkinaki will write a summary of what exists and a proposal of what could be done next. It will be then turned into a **use case** (**Action by A. Kokkinaki**).

Helen summarized and Dick added that is important partners to indicate other projects related to this work so as ODIP to make the links.

5.5.2 Discussion

Included in paragraph 5.5.1 above.

5.6 SESSION 6 – Vocabularies: plenary

5.6.1 NVS Developments

'One-armed bandit semantic model'

Roy Lowry (BODC) presented the NVS developments. The L22 instrument vocabulary extended for harmonization with IMOS and R2R. 89 additional concepts added to L22 since the last ODIP workshop. The work with IMOS is completed. There were some handful of 'difficulties' with the R2R mapping was one problem but is almost finished. For example it wouldn't map combination of instruments labelled as instrument. In P01 757 concepts added since April. Now 13350 P01 concepts are marked up with CAS. CAS is a Chemical Abstract Service registry number, an identifier given to a

chemical substance by chemical abstracts and it gives the chemical manufacture sector. When a company starts manufacturing a new compound they register it with a CAS.

The semantic model for P01 became a 9-wheel of one arm 'Bandit' and so far it covers chemical substances in biota. Roy Lowry then 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. A huge amount of combinations can come out of the combination of these 9 elements and green dogs can come out (the green dogs is the new cartoon that scientists have adapted!). Developing the Chemical Substance wheel was the most work. After checking of P01 concepts, a good definition of a chemical substance is that it is an element, an isotope, a compound or a mixture. For example Chl-c is a mixture of c1, 2 and c3, Chl-a is a pure compound. An ideal scenario for defining a substance wheel (Simon Cox proposal) was to find a substance wheel that was a resource outside NVS, and the resource needed to fulfil certain criteria: it has be comprehensive, be an authoritative collection of substances; Each substance has a URI for reference to it at the RDF triples; URIs resolve to RDF documents. But this is a Cloud Cuckoo Land (unrealistically idealistic state where everything is perfect) and Roy Lowry could not find it. Two possible candidates were the Chemical Abstracts Service (CAS) Registry Number and the Chemical Entities of Biological Interest (ChEBI). Both had issues that prevented Roy to adopt them universally. With CAS there is a small number of duplicates and ambiguities (a Chinese chemical manufacturing company had few registrations on radar that already had been registered); Poor coverage of mixtures because the chemical manufacturing industry do not deal with mixing catalogues; Poor coverage of compounds of research interest that are not manufactured commercially, for example sterols from mushed scallops. The issues with ChEBI are: there is patchy coverage of some substance types such as large organic molecules, isotopes, mixtures. ChEBI wisely tends to avoid large organic molecules, and some isotopes and mixtures (Chl c) are absent from ChEBI. There is a confusing range of entities, for example Cadmium atom, Elemental cadmium, Cadmium molecular entity, have different identifiers. There is a huge number of replicate IDs, Roy shoed and example of three different URIs which resolve to the same thing. There are also multiple URIs not helpful for semantic interoperability.

So, the decision taken (using also Simon's Cox advice) was to create a Chemical Substance Wheel in NVS (S27). The objective is to guarantee coverage for every chemical substance (about 1200) in P01 and maintain operational (trigger-driven) mapping to external resources of ChEBI, CAS and eReefs so as when a new P01 parameter is created and the substance is in ChEBI, then the trigger sticks automatically the URLs into the RDF document. Alexandra will show a demo on this later on. The population work is in progress, currently 191 out of 400 concepts are covered and the target is about 1000-1200 concepts. The downside is that every substance has yet another URI. This could not be avoided because P01 could not covered otherwise.

The Maris vocabulary client has done sterling service for SeaDataNet and many other projects. However, long usage has shown it to have some limitations: Limited search behaviour control causes hit floods; No management of deprecated concept; No ability to locate vocabularies by searching concepts; Dependent on BODC Oracle back office for cache refresh; Doesn't cover all (190+) vocabularies in NVS; Display not optimised for mobile devices. A new Search Client has been developed by BODC to address these limitations (next presentation)

Thomas Loubrieu asked that there is also another identifier for chemicals entities, the INCHI and how it is related. Roy replied that ChEBI carries the INCHI identifiers as well and that there are several identifiers, each has its own coverage and is feasible to extend the mapping to these if people find it useful, technically it can be done. The difficulty is to mint all this information.

Dick Schaap asked if there is contact with Simon Cox for this different movement. Roy replied that Simon is not involved directly. In Liverpool Simon was hoping that CheBI would do the integration in the substance wheel, but as this did not happened, progress should be made. But anytime ChEBI or else can provide full coverage, it can be plugged in the RDF replacing S27.

Cyndy Chandler asked if Simon is aware of the duplicates entries. Roy replied that they've been in contact not specifically for the duplicates but inconsistencies and broken URLs in eReefs have been reported and Simon has fixed them.

NVS search client

Alexandra Kokkinaki (BODC) presented the NVS2 Search Service that developed for searching inside the NVS2 Vocabulary Server, especially into the collections URLs (http://vocab.nerc.ac.uk/collection/) and inside each one collection (http://vocab.nerc.ac.uk/collection/XX/current). The tool developed to help users locate easily the codes and the related vocabularies, of the terms they are interested in. The technical architecture was presented: it talks to a sparql endpoint, so there is a sparql triple store with all the collections and the concepts inside, the sparql queries that ask the sparql endpoint through Javascripts/JQuery and the landing pages. There are two types of search: to locate controlled vocabularies that contain a search term and to locate a search term inside a certain vocabulary. Two types of users, the simple ones and the Roy type apply the two types of searches. Alexandra then gave a live demonstration of the Vocabulary Search Client (https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search), with search simple and advanced examples, explained how to sort, download results, search by catalogue and case sensitive, how to narrow down the searches to single collection (where deprecates terms are showed), exclude terms, etc.

The service is now fully operational and can be put on to the SeaDataNet site.

NVS Linked Data demonstration

Roy Lowry (BODC) demonstrated how he thinks that linked data is: a series of links diving into different resources providing information, all driven by RDF, the fundamental standard of linked data.

The (http://vocab.nerc.ac.uk/collection/P01/current/VLZJ0092/) is a URL to a single concept within P01. It looks like html but it is a RDF. The 5 wheels exposed (S06/observed phenomenon, S27/substance, S02/relationship between the substance and the matrix, S26/matrix, S25/biota) are URIs. Clicking on the S27 link gives: links to other P01 'total iron' concepts, and links to ChEBI, CAS and eReefs (Simons' world). Clicking on the S25 link gives: links to other P01 'Asteroidea' concepts, and links to WoRMS and LSID RDF.

Currently the demonstrator is containing 'total iron in biota' P01 concepts. Work is in progress to bring to operational status. It involves: Population of 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.

Finally, Roy Lowry presented the changes to the BODC vocabulary team as well as personal changes due to his retirement on November 1, 2015, after 35 years of work.

The group applauded Roy, the wizard of vocabularies!

Use of Controlled Vocabularies, ODIP II USA partners

Cydy Chandler (WHOI) reminded the USA activities. It is work by 4 Institutes collaborating with R2R (LDEO, FSU, SIO, WHOI). Besides the controlled vocabularies listed below, R2R, and BCO-DMO encouraging scientists to use also ORCIDs. The vocabularies are exposed to RDF to support link data as well. The R2R has a different scope than BCO-DMO and uses different vocabularies but both use 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 (R2R, BCO-DMO):

- 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)(US NSF now suggests PIs obtain an ORCiD)

R2R: http://www.rvdata.us/voc, (US academic fleet data)
BCO-DMO: http://bco-dmo.org, (NSF funded marine ecosystem research data)

NVS vocabularies used by R2R
- SeaDataNet measurand qualifier flags (L20)
- SeaDataNet Ports Gazetteer (C38)

Matching and mapping in progress:
- SeaVoX Sea Areas (C19)
- BODC Parameter Usage Vocabulary (P01)
- SeaDataNet Parameter Discovery Vocab. (P02)
- SeaDataNet Disciplines (P08)

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)

5.6.2 Report on AODN and ANDS vocabulary developments

Sebastien Mancini (GA) gave an update on the work on vocabularies of the VOCRAM project (started Sept. 2014), and explained how eMII has been using the VOCRAM tools to build vocabularies and implications for ODIP. The work started 3 years ago to improve the 1-2-3 Portal functionality meant harnessing some controlled vocabularies (and mandating their use). eMII and IMOS adopted MCP 2.0 metadata schema and built a small number of AODN vocabs to support content population (where possible by re-using existing terms managed by others, e.g. UK NERC). Till recently creation and administration of these vocabularies was restricted to internal management by eMII (not a useful model for encouraging community participation). eMII approached ANDS for a national-in-scope project VOCRAM. He showed a schematic representation of how they use the ISO19115 metadata at the 1-2-3 portal.

ANDS already had some infrastructure that could be used to support vocabulary services but they weren't properly integrated and the creation/editing functionality was missing. VOCRAM used to source existing editing software (Pool Party) and to bind all components together to form an integrated, user-friendly vocabulary services tool suite. Sebastien then presented the Pool Party tool, a commercial tool but with academic license for reduced cost for all research Institutes in Australia (5000 dollars/per year), its team is based in Austria, and managed by ANDS. It is a web based application, project based (one parameter is a project, a scheme is a project), and explained its functionalities: how to manage the vocabularies, to create new concepts, the wiki viewer, its associated sparql query endpoint, how to publish SISSVoc, etc. Few issued found during data integration. ANDS built additional tool on top to do some cleaning and then data were exported Organizations can be populated by EDMO. Dick noted that there about 50 AUS EDMO entries in USA cruises that need to be confirmed that are not duplications.

Sebastien concluded: AODN has now a tool to publish vocabularies covering: Parameter (167 terms), Instrument (236 terms), UoM (62 terms), Platform (324 terms), Organisation (366 terms) with links to EDMO entries. Most of these also have at least one published classification scheme. They have just met with AODN community representatives and will work with them to increase content in existing

vocabs and add new vocabularies. Need to pilot an appropriate community moderation and governance model.

AODN Community can now (as of 22nd September 2015): discover available vocabularies through ANDS Portal, download (versioned) vocabularies through ANDS Portal, link directly to a published concept definition through URL reference, use the ANDS Tool suite to create and publish their own organisation-level vocabularies. eMII will be seeking collaboration to make sure that organisation-level vocabs created in the community, that potentially overlap with AODN vocabs, have appropriate mappings. Ideally adopting AODN basic vocabs is the goal. They are in better position now to look at (automated) metadata mapping between MCP 2.0 and SeaDataNet CDI.

The group then discussed how AODN could apply a similar to the Roys' spinning wheels approach to allow scientists to do their mapping outside the PoolParty tool.

5.6.3 Report on RDA VSIG activities

Rob Thomas (BODC), as present at the 6th RDA Plenary meeting last week, reported for the Vocabulary Services Interest Group. The Group is chaired by Adam Shepherd (BCO-DMO), Simon Cox (CSIRO), and Stephan Zednik (Rensselear Polytechnic Institute). Adam Leadbetter (MI) is co-chair. Fifty people attended the meeting, 3 presentations to set the scene. Examples presented were earth science/marine centric. ODIP represented by BCO-DMO, CSIRO & BODC. (A second BOF meeting with 28 attendees was the same time with the harmonization group). People from diverse domains interested on vocabularies and best practices. Potential action items for the group to take are Wiki to list showing:

- known vocabularies
- known services and tooling
- on vocabularies services:
    - list 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 aways:

- Quite a few people interested in meeting up by telecon (outside of RDA plenary).
    - Mention of twice a month as a possibility with consideration of global coverage.
- People interested in identifying best practices for vocab services
    - Driving towards this should identify problems
    - Problems areas can become focus of the VSIG for solving

Additional material:

- 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 (signup for mailing list)

5.6.4 Discussion

Outcomes of discussions combined with the output of the break out working group (see paragraph 5.10.2)

Day 3 of the Workshop, Wednesday 30 September 2015

5.7 SESSION 7 – Model workflows and big data: plenary

5.7.1 Model workflows and big data

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

Adam Leadbetter (MI) outlined the presentations plan and introduced the session.

Big Data is a lot of different things associated with it, it is complex data, social data, storage, technology associated with it. A lot of the common definitions are focusing on the five Vs of big data. We attend to focus on volume but there is a lot of technology around, velocity if you move data around, speed if you move data through streams on different systems, there is variability (data of different types), variety and veracity of data (what is their quality and how fit for purpose in terms on analytics). There is an overlapping in the middle and there is where the Big Data is. This session is dedicated to these things, how they overlap and to find some applications for them in the marine domain (punch cards nowadays are not enough for Big Data). Lesley's Wyborn favourite definition 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."

Australian perspective – what has already been achieved and more (811-0107, 03.17 – 36.22)

Lesley Wyborn (NCI) presented the Australian perspective on Big Data control and the volume aspects related with these. She explained that Big Data is a relative term where the volume, velocity and variety of data exceed an organisations storage or compute capacity for accurate and timely decision making. The problem is the scale of the storage and moving data around (and not the data). Combined and integrated, the NCI collections are too large to move: bandwidth limits the capacity to move them easily; the data transfers are too slow and too expensive; even if our data can be moved to public domain, few can afford to store 10 PB on spinning disk. So, it was needed to change the focus to: moving users to the data; moving processing to data; having online applications to process the data in-situ. The call was for a new form of system design where: storage and computation are co-located; 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 to new, more complex Data Mining. We are moving from: "Give me the file, the whole file, and nothing but the file and let me process it locally on MY KIT", to: "Please enable me in real time to discover, access and process only those parts of multiple files and/or databases that I need and let me do it online using YOUR KIT and let me drive it from my iPAD or my SmartPhone". For the marine and oceanographic community, the impact of moving to this new environment is that will loose degrees of freedom since common storage formats should be used for the large collections.

The Australian Government invested 375 million dollars to build a Big Data research infrastructure to make available for their publicly funded national data through the Research Data Services initiative (RDS), which supports over 40 PBytes of multidisciplinary data at nine nodes around Australia. The marine/oceans data are 1319 TB (without the marine satellite that used for SST). One of these nodes is at the National Computational Infrastructure (NCI). NCI has established a powerful and comprehensive in-situ petascale computational environment to enable both high performance computing and Data-intensive Science across a wide spectrum of national environmental and earth science data collections. The platform is called the National Environmental Research Data Interoperability Platform (NERDIP), and includes the data, data management, data catalogues and data services for a comprehensive platform to enable access by a variety of communities for multiple use cases.

Over 10+ PB of data have been co-located at NCI and comprise major national and international data collections from social to space data. The collections are called the National Environmental Research Data Collection (NERDC) and they span from the core up to astronomy and comprises 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).

Combined, they offer unparalleled opportunities for geosciences researchers to undertake innovative Data-intensive Science at scales and resolutions never before attempted, as well as enabling participation in new collaborations in interdisciplinary science. Lesley presented some examples (Larger Scale Inversions, higher resolution drive new science). A unified data platform, the NCI National Environmental Research Data Interoperability Platform (NERDIP), 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, VL's and Data Portals. Two examples were presented: 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 Data (HPD) sets that are structured to facilitate uptake in HPC environments. Once incorporated into NERDIP, the next challenge is to ensure that researchers are ready to both use modern tools, and to 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).

Issues for ODIP:

- 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 if you are lucky 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

EU perspective – Streaming data processing

Adam Leadbetter (MI) presented the velocity aspects of Big Data for getting data back in real-time using some of the technology of some big companies that pay for it. We know how to do batch processing but we do not yet how to do it real time or near-time more in proper scale. How you scale the real time streaming data? He explained the Unix philosophy (McIlroy, Pinson&Tague, 1978) and how the Marine Institute applied the Unix way of programming and tools to the real time data feeds. They explored the stream composition through the context of the Galway Bay cable observatory project. He gave some background of 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. Docker container is on a shore station server with serial2kafka app running. Shore station Kafka holds on to the data for a fortnight. Kafka queue replicated across the network to HQ. Raw data stored in Cassandra are available through ERDDAP. Or, some augmentation through stream processing in Storm - back on to a Kafka queue, exposed through WebSockets.

Another issue that MI is looking is how to provide engines to build the above flows through messages queues or processing tools without writing code. NiFi is a web flow based programming tool. It is based on drag and drop plus configuration in a workbench, i.e. as little code as possible. It includes ability to fire off, say, the individual R processes from the previous slide. However, may be mainly of use for ingesting data as far as a message queue, but not make any composition of data.

How streams of data are related to the Internet of Things (IoT). IoT is the network of physical objects or "things" embedded with electronics, software, sensors, and connectivity to enable 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.

JSON messaging becomes very interesting, presented in OGC SWE Domain Working Group – 16th September 2015 – Simon Cox (& Peter Taylor). Now there is a proposal for an O&M encoding in JSON. The encoding includes URIs, and SOS can be easily built on top of it.

Further work includes: Look at the architectures for reprocessing historic data; Incorporate Sensor Web Enablement, OM-JSON; Deploy on vessels/mobile, remote platforms; Investigation of Apache NiFi.

The group discussed to set up a structures space on the ODIP website to share tool, libraries and put the relevant resources and links to other external material. Dick Schaap will create the space.

Jonathan Hodge then gave some online examples of different types of model workflows environments and how they have been implemented in Australia such as: the CoESRA virtual experiment environment (http://www.tern.org.au/CoESRA-pg29647.html) for data complex 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, fisheries models with other local ones along the coast for better understanding of the ecosystem. The effort is similar to crowd sourcing (or science sourcing). The used framework is html plus java for the website, postGIS database, Threads server, NetCDF CF compliant files and all data services use OGS WMS/WFS standards.

Addressing Variety and Veracity with GeoLink: a US perspective

Cyndy Chandler (WHOI) addressed the Variety and Veracity challenges of Big Data for the work done within R2R and started from the work done in BCO-DMO Project for marine ecosystem data. By variety it is meant the need to integrate vast array of data types. The working framework has changed from working with data collected by scientists themselves and with different data types from a distributed environment. There is broad temporal and geographic ranges and scales, there are in situ observations and measurements from hypothesis-driven research (new instruments), model results, laboratory experiments, integration of social science data in addition to the full range of natural geo science data for ecosystem analysis with relevance to the society, new data types (e.g. metabolomics). Veracity is quality and mainly two aspects of it. First, integration requires high quality metadata, being able to trust the metadata, describe the data resource of interest and present this information to the scientists (fit for purpose). Provenance information is a huge challenge especially for data coming from many different sources. Efficient federation requires resource accurate matching between repositories with complementary content (vocabularies aspect). Secondly, we want to make information available in a machine-interpretable way. The client is no more human but a machine, the scientific work is driven by machines now.

Cyndy Chandler gave an example of the NSF EarthCube GeoLink Project (Semantics and Linked Data for the Geosciences) where she is involved with Bob Arko on how Semantic Web Technologies offer some solutions for meeting Big Data “Variety and Veracity” challenges. The Project aim is to bring together experts from the geo sciences, computer science, and library science; to develop Semantic Web components for geoscience research data; and support discovery and reuse of data and knowledge. Semantic web technologies usage and stacking is used instead of re-eventing them. The partnership is a range of data environmental data providers representing a broad range of geoscience research community. The domain focus so far is marine ecosystems, starting from cruises, both sensor and sample data from observing networks to the “long tail”, and informed by activities in many other projects and communities including ODIP. Two use cases on ocean ecosystems and seafloor morphology were used to drive the developments in the Project. Ontology design patterns (ODPs) were used to harmonize content of different repositories, trying to identify the essential attributes and properties that describe the main concepts of the information. For example, a field expedition is called cruise but it could be another organization or unit collecting data for a specific purpose. There were different kind of information not only data but journal publications from peer reviewed journals, conference presentations, abstracts, PhD thesis, funding awards. The ODPs used as a filter to publish subsets of their content as Linked Data (according to W3C) and developed an

integrated knowledgebase and exercise it against science use cases. The “Linked Data” principles were followed and the bench mark was “The Web is the API.” (Bob Arko, March 2015): use URIs (network names) to identify things (using controlled vocabularies definitions); Use HTTP URIs, so these things can be referenced by both humans and machines; Describe these things using standard languages such as RDF and SPARQL; and include URIs (ie. links) to other related things. In particular, R2R and BCO-DMO both define what cruise, platform (e.g. vessel) and instrument are; match and link them to NVS terms and use ICES to define platform codes. This is a simple idea that it works when you put a lot of information together [Terms from the NERC Vocabulary Server (NVS) are important for federating content from distributed systems (Leadbetter et al. 2013a)]. The vocabularies terms were mapped and linked to R2R, BCO-DMO content by using the URIs from NERC, for example R/V OCEANUS is defines as: http://vocab.nerc.ac.uk/collection/C17/current/32OC/ [Leadbetter, A., R. Arko, C. Chandler, A. Shepherd, and R. Lowry (2013b), Linked Data: An Oceanographic Perspective, The Journal of Ocean Technology Vol. 8 (3). pp. 8-12]. Not only R2R and BCO-DMO, but MBLWHOI library is now using NERC vocabs and publish URIs in their RDF. It would be interested to connect in such as way data and publications (open data access is required for that).

Cyndy Chandler gave an example from VERTIGO Project where a Cruise Description and Trackline at R2R is connected with a Multiband Sonar Dataset at R2R/NOAA and a Sediment Trap Flux Dataset at BCO-DMO/WHOAS through the cruise id. The dataset is linked with a Journal Article on VERTIGO using DOI and a NSF Funding Award. The chief scientist encouraged to get an ORCID, he hadn’t one before. So, if all of the different repositories have at least one global PI common, the web could make the connections. Another example using the physical data is a cruise (with R2R cruise_id, and ICES platform code) which is connected with a publication available at USGS NGDB, which uses IGSN to identify physical samples. The conclusion is put PIs to as many things as you can, at instances level.

Cyndy Chandler concluded that Semantic Web Technologies offer some solutions for meeting Big Data “Variety and Veracity” challenges. The particular group using controlled vocabularies and Linked Open Data (RDF/SPARQL) are important parts. Standards are important, BCO-DMO supports ISO19139, FGDC, GCMD DIF, schema.org Dataset extension, formal data publication with a DOI, and RDF with semantic markup including PROV, FOAF and more (such as SKOS, OWL). The challenges (which are related to ODIP) are: lack of key vocabularies published online using OWL with URIs; lack of gazetteer data (eg. physiographic features) published online with URIs and proper geometries; lack of universal Person and Organization identifiers (with sufficient metadata); and need to map/match instances manually, at least in the beginning.

Helen Glaves noted that ORCID does not require other information than the name. Cyndy said that we need identifiers that unambiguously connect the PI with the right person. Alexandra Kokkinaki asked if they publish their own ontologies if they do not find the originals for Phds for example. Cyndy said that they use URIs internally to connect with their knowledge base but they do not make them publicly available. Alexandra said that using others’ ontologies, they can extend and publish. Cyndy replied that she will discuss it with Bob Arko. Lesley Wyborn commented how they maintain and publish vocabularies in Australia.

OUTILS HYDRORUN: MarsWeb

Thomas Loubrieu (IFREMER) presented the MarsWeb, a service that give access on the local HPC in Brest to people who are doing monitoring of the coastal environment and are spread along the coast and in other overseas territories. MarsWeb purpose is for coastal monitoring and is a web interface which enable scientists to run models with configuration of inputs on atmospheric conditions, hydrology, bio-geo-chemistry, and tides inputs. There is the possibility to monitor the run of the model at HPC at Brest, to visualize the results on line and have advanced products as model outputs. Some screenshots of the configuration of the model, the nesting of models on the map, the monitoring of the runs, the visualization results as well as the architecture were given. Thomas Loubrieu then presented the demo on: snanny.ifremer.fr/dashboard.html. More than 2 million points are indexed. Sub setting of the data sets is possible, the example includes merging of Argo profiles and navigation of research vessels from IFREMER. Switch from the density map to the actual measurements and visualization of

the individual measurements is possible. Currently there is a cluster of seven servers and the perspective is to expand it next year in terms of CPU as FREMER plans to build a Big Data infrastructure.

5.7.2 Discussion

Adam summarized the discussion points:

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

The group discussed first the CF standardization.The NetCDF is already an OGC standard written by CNR (Stefano) and UNIDATA (Ben) group. Any proposal for CF should come from this group.

There are several CF attributes lists which cover different topics. There is no association between the the groups for example between bathymetry and geophysics, also there is an issue with satellite. All these groups should coordinate to move forward together. Lesley Wyborn proposed to find which topics and data sets the ODIP covers.

Some of these groups extend the CF terms but not in a compliant manner. In CF there are a huge number of degrees of freedom. The guidelines are loose, the set of conversions are branched, there are two different versions dealing with gridded data, 1.5 and 1.7. For point data is 1.6 version. There are other issues also, some people label data as CF compliant but it is not, even if they pass the CF checker. The official checkers are for gridded data only and not for point data.

AUS has 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 needs.

It cannot be one global implementation of CF conventions in great details because different communities have different needs. For example the SeaDataNet CF profiles cannot carry the extra bits specific for bathymetric data. Roy had proposed a layer structure: the CF conventions at the bottom, then a layer with the specific community conventions for all data types. This was done in SeaDataNet where a part of the profile included an attribute with the SDN parameter codes in addition to the standard name. Other layers on top can be added for specific types of data, e.g. for a like bathymetry profile. Roy proposed that ODIP could help to have a grid CF profiles for specific data.

New prototype identified:

The group agreed to make a prototype with an inventory of the CF profiles that are being used within the ODIP partners, an inventory of checkers, who are using them (IMOS, IOOS, ...), what can be plugged in, and make it an OGC standard. The prototype will be led by Australia (Sebastien Mancini) (**Action: new prototype)**.

Some developers do not like NetCDF and prefer to store data and metadata in a database. But NetCDF is meant here as exchange and not as an archive format.

Other issues that can be checked by the prototype are the feature types, like the number of instruments, platforms per file andwhat the best practice is. Usually people put on instrument in one NetCDF files, but in other cases like OceanSITES they use multiple instruments and it is difficult to manage (OceanSITES is the official profile in the in situ TACS and lot of improvements are needed).

The extension of the ODIP web site by setting up of a wiki (or any other tool) to host and manage all the ODIP standards (not only CF) and available resources with appropriate ownership and a responsible for the content management (clear governance structure), will be explored (**Action by Dick, Adam, Jonathan)**.

The next discussion topic was the merging key data sets (LIDAR, bathymetry and coast line). Lesley Wyborn asked if there are countries that have problems with acknowledged coastline and cannot

merge LIDAR with bathymetry. Helen noted that in UK they use the “white line” around the island which is where the land stops and the sea starts and integration with bathymetry requires huge efforts. Dick noted that in Netherlands there are many different coastlines of different definitions (high/low water mark) depending on the use. The issue is not the definition but the methodology used during merging that affects the results. EMODnet bathymetry uses the OpenStreet water map as coastline as it not a definition. A new EMODnet project called Coastal Mapping which will deal not only with the methodology of merging but also with the definition of a common coastal mapping in Europe (UK is not part of it).

The next discussion point was the high performance data sets. The group agreed that this issue with the first one (data and workflows at scale) and all the issues that are related such as how to build workflows, the standards in between, the handling of big data packages, performance of formats and visualization tools could be an interest topic and become an ODIP best practice that could be used by other projects (**Action: tentatively new prototype led by Australia** (to be discussed and confirmed by the AUS partners).

Dick Schaap gave an example of interest. The previous EMODnet phase had the right methodology but the performance of the tools was not right. EMODnet could be lifted a lot as a product machine by using this ODIP best practice. Then the powerful new tools (and the new data) could convince the MFSD and regional conventions to use these for impact assessment calculations. In this way, the authorities, decisions makers (as end users) and not only the scientists become part of the process and become owners of the outcomes (returning line).

5.8 SESSION 8 - Data publication and persistent identifiers

5.8.1 Plenary

Introduction

Justin Buck introduced the presentations of this session.

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

Helen Glaves (BGS) presented the COPDESS initiative that intends to bring together the Earth Science publisher and repositories to help translate the aspirations of open, available, and useful data from policy into practice. The drivers behind the creation of COPDESS are: Open access and open data mandates widely acknowledged and being addressed; Data are increasingly recognized as part of the scholarly record: data citation is coming of age; Cyberinfrastructure & eScience developments require access to data in standardized formats; Growing number of repositories - need to adhere to best practices. She explained then the data publishers’ perspective: Many have had supplements for some time (Difficult to deal with/costly); PDF’s mostly (not searchable, poorly indexed, variable quality); Require authors to comply with data availability policy; Little guidance on community standards; Want to use and promote repositories, but not well integrated except for a few exceptions; Worried about repository funding and stability. The data repository perspective is different: Provide important quality control and discovery; Valuable for discovery and integration; Poor connection to publications, often ad hoc or case-by-case; Want better integration with publishers; Much data not being collected; Worried about funding.

COPDESS was founded in October 2014. It is a permanent international coordinating conference of publishers and data facilities & consortia on Earth and space science data publication. The structure of this coalition will be more clearly defined over a series of upcoming meetings. A draft Statement of Commitment has been drafted and released on 15 January 2015 saying that repositories and to a less extend the journals adhere to a best practice regarding data sharing and archiving and how these two bodies of expertise will interact.

There are a number of ongoing actions trying to develop COPDESS and how to move forward: Build an online directory of Earth and space science data repositories that can be used by journals and authors; Promulgate metadata information and data standards; Develop common workflows within the repositories that support the peer review ..and within the editorial management systems that will ease transfer of data to repositories. Other actions are: Promote referencing of data sets using the Force-11 data citation principles; Promote and implement 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; Promote use of other relevant community permanent identifiers for samples (IGSN), researchers (ORCID), and funders and grants (FundRef). Next activities include: COPDESS Directory of Repositories to be released in August 2015; Workshop in Europe (funded by Sloan & NSF), October 20/21, Oxford, UK, with main focus to organize training on the use of the directory, alignment of journals and integration with editorial managers.

The group raised several questions and issues such as: how a repository can get accreditation (=through WDS criteria), or if the information on reviewers is included (=yes, along with other metadata to enable reproducibility). Other points were that the biggest challenge for COPDESS will be to structure the directory of repositories, and if a repository should keep the streams to the raw sensor data or the processed, although is not feasible always to keep raw data due to the cost.

Depending on the data type (satellite, water samples, etc) there is no point always to keep raw data if you trust the instrument (the question raised what is raw data). Today there are intelligent sensors that process the data and there is no point to keep the signal, perhaps is meaningful to keep the provenance metadata on the standards used for the processing.

RDA Marine Data Harmonisation IG/Data Citation WG

Helen Glaves (BGS) presented the outcomes of the joint session of the RDA Marine Data Harmonisation IG/Data Citation Working Group. The Marine Data Harmonisation IG was set very much in parallel with ODIP to bring together all those who had interest in research on the marine environment, from biology to the social scientists. Its objectives are: to promote a common global framework for marine data management through the use of common standards and best practices; inform the activities of other RDA IG/WGs with relevant input and feedback from the marine domain including providing suitable use cases; disseminate the outcomes of relevant RDA WGs/IGs to the marine data management community. The group is working closely with other related initiatives like ODIP, Belmont Forum, IOC-IODE.

Within RDA there are several Working and Interests Groups with relevance to the marine community. Currently the work within the metadata IG/WG is to: develop 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; identify a small number of suitable use cases; if successful develop a strategy for wider adoption of these solutions; and develop a proposal for and RDA Collaboration project.

Citing data may seem easy: from providing a URL in a footnote via providing a reference in the bibliography section to assigning a PID (DOI, ARK, …) to dataset in a repository. The reality is that citing data is more complex. There are issues about the granularity of data to be identified/cited such as: databases collect enormous amounts of data over time, researchers use specific subsets of data, and need to identify precisely the subset used. The current approaches seem to be: Storing a copy of subset as used in study (scalability); Citing entire dataset, providing textual description of subset (imprecise/ambiguity); Storing list of record identifiers in subset (scalability), not for arbitrary subsets (e.g. when not entire record selected). What increasingly people want is to be able to indentify & cite precisely the subset of (dynamic) data used in a study.

Citable datasets have to be static, fixed set of data, no changes (no corrections to errors, no new data being added). But research data is dynamic. We add new data, correct errors, enhance data quality and changes sometimes are highly dynamic, at irregular intervals. The current approaches are: Identifying entire data stream, without any versioning; Using "accessed at" date; "Artificial" versioning by identifying batches of data (e.g. annual), aggregating changes into releases (time-delayed!). The solution proposed by the RDA WG on Data Citation was: to cite precisely the data as it existed at certain point in time, without delaying release of new data. This solution requires the data to be time

stamped and versioned and be able to recover these snapshots from the servers, dated and stamped. To do so you need to store the queries and not the data subsets. But this solution cannot work because the databases are dynamic, and the data change, the queries should be dynamic and this creates performance issues with the databases (processing intensive). The purpose of this session is to provide feedback to this group why this solution does not work. From the Argo experience it came out that what is needed to freeze versions of data and not keep track of the changes (you cannot re-create history but frozen it). Argo now has found another solution (Plan B). In AUS neither solution cannot work for legacy reasons, they cannot recreate any governmental database archive more than 7 years old.

Perhaps frozen local copies is a solution, and one query would be enough (as we cannot restore history but we can freeze it and manage it). Storage is not a problem now neither at the future (with cloud computing).

As a conclusion, the RDA WG Data Citation recommendations cannot be followed because our systems and infrastructure have been built in such way that cannot meet these criteria. Dialogue with this group should be initiated as well as a more representative participation with members outside Europe.

A 2-page flyer with the 14 recommendations on data citation is available on the web. There is a RDA wiki also for input and comments by the community. BCO-DMO will upload some use cases for data sets that meet these criteria, for time series it will be more difficult. Justin will also contribute. Shane commented that there is a difference between pointing to a specific time point and static monthly snapshots and proposed to define a use case with the difficulties on getting to an arbitrary point in time rather than restoring versioned static copies from the database or monthly snapshots of data. Lesley noted that as these extractions have legal use they should be labeled. Justin recalled the example with the scientists who were led to the court because of the misprediction of the earthquake in Italy.

Argo DOI progress

Justin Buck (BODC) introduced partners what the Argo global array is. It is a global array of more than 3,000 free-drifting profiling floats. Each measures the temperature and salinity of the upper 2000 m of the ocean and every 10 days it send back this profile. This means that the data set is continually growing. At the same time the scientists and operators quality control the data and this change the time series that already exist. So, effectively there is a time series going back to 1998 which change and grows, and effectively the whole ARGO data sets changes and grows. Currently there are more than 2000 publications in 15 years citing Argo data and none of these citations can unambiguously get the data as they were at the point of time of their analysis and this is a significant problem.

ARGO GDACs were created 15 years ago to have the current versions of data. But US NODC outset decided that they need snapshots of the entire data set every week. IFREMER, who has done a lot to assign initial DOIs to the data, had to go back at monthly granularity. So, we cannot reproduce the data at any point in time at roughly weekly granularity. So, rather Institutes versioning the GDAC data (resources are needed for that) it was decided to exploit the snapshots that are held.

At the moment IFREMER assigns a separate DOI to every single snapshot. The goal of the ARGO steering team and the scientific community is to have a single DOI, tracking and cite it is easier and more transparent. The proposal (with the help Simon Cox) 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:
- e.g. http://dx.doi.org/10.7289/[Argo_accession_DOI]?[time_slice _information]
    - # Client/browser side snapshot resolving service via a specific javascript for the accession
    - ? Server side snapshot resolving service, preferred but not currently supported by DataCite.

Where 7289 is the NOAA or IFREMER DOI prefix code.

Finally, the identifier with the date at the end, is:

- http://dx.doi.org/10.7289/argo_doi_identifier?result_time=2005-01-11T16:22:25.00

Justin Buck then showed an example of an Argo snapshot in the Sextant metadata catalogue and how to cite these data. Also the catalogue provides information to data managers what metadata are available and how to mint a DOI. There is a debate if the citation should appear as well. Peer reviewers say that the citation should go to the references of the journal so in many cases the citation appears to the acknowledgments which make it harder to trace especially the acknowledgments in Elsevier. If you search Elsevier by google, you do not find the specific DOI.

The recommendation from RDA for the single DOI with a time id (with a # or ? between), was not to put date at the end but an off scatted identifier. Maybe it is not such as a bad idea as we do not resolve the data every second in time.

Because snapshots are held at two locations, another issue is that we need to resolve a location based on the nearest time.

Justin Buck strongly encouraged to apply for some RDA funding to investigate this work (will check with IFREMER and US NODC.

Thomas proposed the data management systems to keep track of the changes in data sets within the snapshot strategy. This requires homogenized databases, one version copy with additional text. The group then discussed the sustainable snapshot management (volume issue), should use a version control repository system (e.g. GIT).

AUS report on Dynamic Data Citation & IGSN

Lesley Wyborn (NCI) gave the AUS report on DOIs, She presented the Dynamic Data Citation from the dynamic data perspective. First, she explained what is Dynamic-Dynamic Data and how to preserve dynamic queries and assigning persistent identifiers. Dynamic-Dynamic Data are Data that is dynamically changing and is being accessed by queries that are dynamic (different and often unique). The ability for dynamic queries on data has been explored with web services. Issues raised with RDA Dynamic Citation Group as follows: There are at least two use cases where new data are dynamically added to an existing data set:

- Use case 1: new data are regularly and systematically appended to an existing data set over time, e.g., with outputs from a satellite or sensor, and 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 a select number of attributes/components of the existing data set.

In the case 1 (appending): it was felt that RDA approach was rather data base centric and did not apply to large volume raster arrays. It is easily resolved: time stamp the source data and save the query and the time of the query.

In case 2 (subtle changes made to an existing large volume data set): With large volume raster arrays that can be over a Petabyte in volume and in multi-petabyte climate models, storing multiple time stamped snap shots of these is not feasible, fundamentally due to cost of the infrastructure. The working solution is:

- The data sets need to go through a controlled release process, similar to software and the exact changes to the data set are documented, so that if required (e.g., legal case), a data set can be recreated.
- It is recommended that provenance workflow engines are used, 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 would have a persistent identifier, as would all components of the workflow.

An example was given of how to preserve dynamic queries in a dataset using a provenance workflow engine, the Virtual Hazard Impact and Risk Laboratory (VHIRL) service. VHIRL aim is 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 selected data. It provides natural hazard researchers with access to an integrated environment that exploits eResearch tools and Cloud computing technology. The service captures data service information (hosted on RDS), subset details of data selected, captures code utilised along with "how" it is used (template/input files) and location (PID) of where input files/scripts are persisted. The finalised outputs are persisted with PIDs on RDS and captured in provenance information. After job is completed, finalised provenance record is published to provenance store. PROV record endpoints could be registered in ANDS RDA alongside output data!!!

Lesley Wyborn gave an updated on ANDS minting:

1) 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) although the total number of ANDS DOIs is still low
2) Manual minting was introduced in December 2014, to augment machine-to-machine minting with 9 institutions taking up this option (only 1 is doing both machine-to-machine and manual)

The update for NCI (Who actually mints the DOI ?)

- they have M2M minting ready to go
- no production DOIs minted yet (for social reasons not technical: who does 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 (ie 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: It's probably fair to say that 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, not the Node.

Finally, Lesley Wyborn outlined the IGSN Project. It is funded by Research Data Australia (RDS) for Petascale data challenges, seeking to bring in data on physical samples that can calibrate the petascale, proxy data sets. There are 3 IGSN allocation 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. These local agency 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
- formulate agreed governance models for registering Australian samples.

Justin asked what a sample in the OGC content is. Lesley replied: it is totally based on the O&M model, it is designed for sampling features and samples taken from this feature (e.g a borehole and all the samples from it and the derivative samples taken from the sampling feature). Bob Arko noted that it is allowed to assign an IGSN for both a physical specimen and the feature of interest (e.g. a drill in the ocean, if you revisit it 10 years later is and deepen it, you can make reference of the original identifier from 10 years earlier). IGSN can be used to search repositories and find what work has been done with this sample. Cyndy mentioned in WHOI it is used to identify water samples from NISKIN bottles (not only rocks samples)

IGSN: International Geo Sample Number

Bob Arko (LDEO) gave a quick overview on the IGSN for citation of physical samples. The rationale behind the persistent identifiers for physical specimens is to have discoverable, accessible and citable physical specimens as digital datasets:

- Discovery and Access for Re-use and Reproducibility
  - Samples need virtual representations
  - PIDs need to resolve to these virtual representations
- Sample Citation
  - Sample collectors need to get credit for the intellectual effort and resources they put into their collection (especially in the ocean), preparation, and curation
- 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 utility of these data is substantially higher when combined
- Sample Management
  - Tracking of samples & sub-samples

IGSN was introduced 10 years ago (http://www.igsn.org/). The idea is to guarantee a unique identifier in every specimen and can resolve to a landing page, as a DOI does. These numbers have started to appear to the literature just now.

A working meeting two weeks ago reviewed the metadata that are needed for a sample and recognized that there is an essential set of metadata that represent the sample (like a birth certificate), and needed to describe it and make it discoverable at the search engines. Curatorial information about all the things that happen to the sample through time distinguished from the core metadata. The essential metadata will be turn into an updated ISO&OGC O&M schema (some extra information like sampling type, material type, sampling method, etc that are used in DOI) and will be published following the DataCite metadata principles.

An example was presented on how to use IGSNs to track the provenance: first the hole gets an IGSN, then the neighboring holes get IGSN, in each piece of material. Years later you can put more content in the same hole using the same IGSN (allowing sometimes to track the relationship between parents and children). Initially the effort was focused on rocks and sediments but now extends to fluids and gases, and now biology. Funding agencies advice that to attract publishers IGSN has to accomplish both bio and geo samples.

IGSN now appears in peer publications and when published online you can get the samples birth certificates through the IGSNs links.

For ODIP, one potential use case is to link IGSN with sampling data from Research Cruises:

- Geological samples (core, dredges, grabs, etc) catalogued in Index to marine and Lacustrine Geological Samples, with IGSNs,
- Linked to R2R Cruise IDs for U.S. vessels; sample inventory included in Cruise Sumamry Reports published to POGO.

5.8.2 Discussion

The group discussed the need for well-organized local management systems with local samples identifiers which are linked with the global IGSNs. It was also discussed the link of IGSNs with sensor data.

Justin Buck summarized the next session discussion points and ODIP II possibilities:

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

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

The group split in three parallel break-out session working groups on: vocabularies (rapporteur Alexandra Kokkinaki), data publication/citation (rapporteur Justin Buck), and model workflows/big data (rapporteur Adam Leadbetter). Each partner should attend in two groups.

Day 4 of the Workshop, Thursday 01 October 2015

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

5.10.1 Model workflows and big data

Adam Leadbetter (MI) summarized that the discussions focused on Tools and Platforms that are used in Big Data and model workflows. The group also identified some use cases that could be developed forward in ODIP II.

Tools and Platforms:

- 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

Use Case 1:

- T/S Climatology
    - NCI platform
    - SeaDataNet harvested T&S
    - US, Australian data
    - Choose a focus area
    - Build workflow for creating the climatology
    - Visualisation, performance, scalability are all issues to be addressed

Concerning the Australian data, Sebastien Mancini noted that recently they have put together all IMOS T/S data along the shelf as point data. Dick Schaap noted that for the use cases the most interest is not the product but the methodology (workflows, tools from the 3 regions and how to put them together, it is the interoperability things).

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 ***MAY*** occur

5.10.2 Vocabularies

Alexandra Kokkinaki (BODC) presented the outcomes of the vocabularies working group. A wish list compiled focused on 3 areas to investigate:

- Further development of mappings
- Further development of content
- Further development of tooling
- Best Practices (including best practices for including vocabs in NetCDF files)

For the mappings:

- Implement unit conversions through rich predicates like
  - 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 works currently with version 6
  - mapping to GCMD version 8 URIs
  - Investigate how GCMD is working right now URL's

For the content:

- ODIP to expand C19
  - SeaVoX salt and fresh water body gazetteer
  - Add more content to the geometry server
  - ODIP members send emails to BODC Enquiries to submit content originating from the relevant authority
    - Preferred Label for the sea (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
  - Anyone uses NVS1? (
- Create a self-service governance to help users create their own P01 one arm bandit vocabularies (Dick Schaap commented that this is in the to do list for long time)
- Add richer predicates to P01 to map with P07?
  - One example?

For the tooling:

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

For the Best Practices:

- Develop best practices for embedding vocabs (parameters, units, instruments) in netCDF files

Adam Leadbetter commented to put the mappings back into MMI. Alexandra agreed to do that.

5.10.3 Data citation/Persistent identifiers

Justin Buck (BODC) reported that the break out group focused on aspects on dynamic, trying to find some use cases for ODIP and what is needed to do to progress things.

Use cases that create dynamic data citation are:

- IMOS – multiple dataset versioning
- Marine Institute – NRT platform chlorophyll
- IFREMER/BODC - Argo
- BCO-DMO – file versioning
- IOOS – workbench data
- NCEI – Argo

Between these there are some more 4-5 good use cases to check.

The group also discussed some of the details of the dynamic data citation. The model is: http://dx.doi.org/10.7289/[Argo_accession_DOI]?[time_slice _information]. It was discussed whether to use "#" or "?":

- IFREMER has worked with CNRS to make the use of # possible
- Also needs to go to DataCite and CrossRef for wide adoption in issuing authorities

Another aspect discussed was the Opaque or transparent time information:

- Advantages from user perspective with transparent but potential citation of ambiguous data state
- Further clarification to be sought from DataCite and CrossRef

There is a funding available from RDA to develop prototypes and test the method proposed by Rauber *et al.*

The bid would need to address two themes:

- ODIP to liaise with DataCite and CrossRef to address implementation issues
- Develop and show the viability of prototypes

Decision was taken to attempt to bid as an ODIP consortium rather than individual data/observing entities. Possible partners in the proposal:

- IMOS – database type implementation
- NCI, raster array bid?
- Can this be linked to either EU or US funding
- + ANDS as minting authority

- Marine Institute – NRT platform chlorophyll
- IFREMER/BODC – Argo
- Using RDA Europe funding
- + plus Andi and Ari?

- BCO-DMO – file versioning
- IOOS – workbench data
- NCEI – Argo
- Attempt to obtain matched RDA USA funding

5.11 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 dynamics and change of Southern Ocean systems through the design, advocacy and

implementation of cost-effective observing and data delivery systems. Two key tasks are: building tools to share data once it's collected (the GCMD metadata platform will be used), and building a field project planning portal to spark conversations before the data are collected. Most of the SOOS people are involved in ODIP also. The data office is in UTAS. The Field Project Planning is a tool with a publicly-editable spatial database to show who's doing what and where in coming years. Many oceanic and Antarctic research communities want to develop similar things (e.g. ICED, COMNAP, AFIN, ASPeCt, SCAR, SOCCOM, BCO-DMO, Argo) 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. It is based on a modular approach: different users entry in an easy form, what they are going to do, where, what instruments will be used, Principle Investigators, etc so as to combine efforts resources (like 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

How ODIP can help on: Design the infrastructure; Design a data input form; Build a robust back-end; Build an intuitive user-interface; Code testing; Web hosting; Code maintenance; ?

POGO, Eurofleets have addressed the above issues and will give feedback to Sebastien and Pip Bricher on data@soos.aq (**Action for Lesley Rickards, Dick Schaap**).

5.11.1 Discussion

Helen Glaves (BGS) invited the group to discuss if there are other cross-cutting issues to be included in the agenda for the next Workshop and what additional issues came out from the discussions this week that are not covered by the current ODIP II list of possible topics.

Data ingestion will not be broadening as a separate topic in ODIP II. It is integrated in other topics such as like SensorML, or at CF standardization needed during the ingestion into systems like NCI, unless other needs arise from other activities such as citizen science.

Biology postponed to the next meeting because key people could not attend this time.

It is a matter of time other communities like the climate community to be interested for the ODIP work but ODIP DoW is not designed to extend in other domains. Its mission is interoperability and common standards in the marine and ocean domain. ODIP is not an end2end project, does not communicate with the end users and cannot cover all end users perspectives. ODIP is feeding other projects and by these projects is reaching end users. Also ODIP is a technical project rather than oriented to dissemination and outreach activities. This is the RDA role. ODIP is leveraging with existing projects and through these projects ODIP can have access to end users and find success stories for its impact assessment.

But is important ODIP to understand the users need and do not work on isolation. ODIP and RDA have similar roles, but RDA scope is much wider. Cooperation with RDA, EGU, OGC and other groups helps ODIP not to become isolated. It is useful to invite representatives of relevant communities in ODIP Workshops. Already, many of these groups are already have been invited in (UNIDATA, 52°North, OGC members like CNR).

The management of polar data is a possible next topic and this community could be invited in the next Workshop. This will expand the ODIP community in other regions also like Canada.

A key point discussed was how to share the knowledge that is produced in this group. This should be managed carefully without adding extra technical work load. Tools like wiki could help.

The group also discussed how to address the topics of the proposal as well as those identified during this meeting, as new prototype(s) or as cross cutting activities. The lack of funding for partners outside Europe is an issue. The difference between the two is the prototype has a clearer target, less

freedom and more commitments while a cross-cutting activity is a presentation of what partners are doing between the Workshops but it is not work together as a group.

The three ongoing prototypes will continue in a certain direction. The group agreed to call the new identified topics as cross-cutting for the moment and see later how to proceed. USA, AUS partners are willing to continue work together but cannot commit now, and will explore if they can fund these activities. For the fact that the contract requires new prototypes, Helen Glaves explained that EU acknowledging the funding complications with NSF is flexible in the way that the Project will be delivered are as long as they are well informed about any modifications in the deliverables.

There is interest to make prototype 2 fully operational. If ODIP freeze it as it is now, one could argue that prototype 2 is in production phase, so existing R2R funding could be used to keep R2R into ODIP with routine production and enhancement of cruise reports.

The same approach (success story) could be used for NERC vocabularies in Australia as now these are used in IMOS.

An additional way could be that beyond the prototypes and the cross cutting activities, the separate use cases and their components could be used together as one strong use case with more impact and communication outside. It can be called as a new prototype (and not use case anymore).

5.12 SESSION 12 - Workshop wrap-up

5.12.1 Plans for next 8 months

Helen Glaves (BGS) presented the commitments for the finalization the first phase of the Project. Phase II started in April 2015 and there are some obligations for it also. There are seven deliverables, the key one is D3.4 (with input from partners in the coming one week).

For ODIP first phase:

- 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)

Work package reports should be sent until 12 October 2015, final cost statements from partners until 30 October 2015. All deliverables have to be completed and submitted to EU until 30 October 2015.

The ODIP final review will take place on 13 November 2015, Brussels, with the WP leaders only.

For the second phase the deliverables are (some of the dates have been shifted such as the first workshop and the related deliverables):

- D1.1, 6 month progress report (M7: October 2015)
- D1.2, Minutes of ODIP II steering committee (M6: September 2015)
- D1.5, Operational extranet (M3: June 2015)
- D2.1, 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 been identified:

- AGU Fall Meeting 2014: 15 – 19 December 2015, San Francisco, USARDA plenary: 22 – 24 September 2015, Paris
- AGU Ocean Sciences: 21 – 26 February 2016, New Orleans, USAIODE
- EGU General Assembly: 17 – 22 April 2016, Vienna, Austria
- 7th RDA plenary: 29 February - 3 March 2016, Tokyo, Japan
- Others ??

The 2nd ODIP II workshop will be held at USA, Boulder, Colorado, hosted by UNIDATA/SIO. The provisional dates are for May or June, and will organized back to back with the R2R annual meeting.

5.12.2 Closing remarks

Helen Glaves thanked Sissy Iona and IFREMER people Thomas Loubrieu and Béatrice Milosavljevic for organizing the meeting as well as all partners for their participation. Dick Schaap thanked participants for the exciting Workshop and the mutual exchange of information and ideas. He expressed his hope that ODIP group will continue to meet and work together.

Terminology

Term	Definition
CCAMLR	Committee for Conservation of the Antarctic Marine Living resources
CDI	Common Data Index metadata schema and catalogue developed by the SeaDataNet project
COOPEUS	EU-NSF funded project promoting open access and sharing of data and information produced by environmental research infrastructures
CSR	Cruise Summary Reports is a directory of research cruises.
iCORDI	Now renamed RDA-Europe is an international forum driving convergence between emerging global data infrastructures with a particular focus on Europe and the US
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
GitHub	A distributed revision control and source code management (SCM) system (GIT) repository web-based hosting service which offers all of the distributed revision control and source code management (SCM) functionality of Git as well as adding its own features
IMOS	Integrated Marine Observing System: Australian monitoring system; providing open access to marine research data
INSPIRE	Infrastructure for Spatial Information in the European Community
MNF	Marine National Facility: The Australian cruise reporting system
MPA	Marine Protected Area
ODP	Ocean Data Portal: data discovery and access service, part of the IODE network
ICES	International Council for the Exploration of the Sea
IOC	Intergovernmental Oceanographic Commission of UNESCO (IOC/UNESCO).
IODE	International Oceanographic Data and Information Exchange (part of IOC)
JSON	JavaScript Object Notation: an open standard format that uses human-readable text to transmit

	data objects consisting of attribute–value pairs
ODV	Ocean Data View (ODV) data-analysis and visualisation software tool.
O&M	Observations and Measurements: OGC standard defining XML schemas for observations, and for features involved in sampling when making observations
OGC	Open Geospatial Consortium: an international industry consortium to develop community adopted standards to “geo-enable” the Web
SensorML	OGC standard providing models and an XML encoding for describing sensors and process lineage
SDN	SeaDataNet: EU-funded pan-European e-infrastructure for the management and delivery of marine and oceanographic data
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
SOS	Sensor Observation Service: a web service to query real-time sensor data and sensor data time series. Part of the Sensor Web
SPARQL	a query language for databases, able to retrieve and manipulate data stored in a Resource Description Framework (RDF) format
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
R2R	Rolling Deck to Repository: a US project responsible for the cataloguing and delivery of data acquired by the US research fleet.
RDA	The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data.
WebEx	On-line web conferencing and collaboration tool