Ocean Data Interoperability Platform

Deliverable D3.3: Cross-cutting Themes

Work package	WP3	ODIP prototypes
Author (s)	Dick Schaap	MARIS
Author (s)		
Author (s)		
Author (s)		
Authorized by	Helen Glaves	NERC - BGS
Reviewer	Roy Lowry	NERC - BODC
Doc Id	ODIP_D3.3_Cross-Cutting_Topics	
Dissemination Level	PUBLIC	
Issue	1.0	
Date	25 October 2015	

Document History				
Version	**Author(s)**	**Status**	**Date**	**Comments**
0.1	Dick Schaap	DRAFT	18 October 2015	First draft
0.2	Roy Lowry	DRAFT	23 October 2015	Edits
0.3	Dick Schaap	FINAL	25 October 2015	Finalised
1.0	Helen Glaves	FINAL	26 October 2015	Sign-off

Contents

Executive Summary....4

1 Introduction....6

2 Controlled Vocabularies....7

3 Data publishing and citation....19

4 Unique persistent identifiers for people....29

Annex A Terminology....31

Executive Summary

The ODIP description of work (DoW) includes the formulation of further prototype projects that should form the basis of deliverable *D3.3 Definition of Prototypes 2.* However, due to the complexity and large scope of the first three prototype development tasks, it was decided during the second year of the project not to formulate further prototypes in the present ODIP project but instead to focus on a number of cross-cutting topics that were of direct relevance for the on-going prototype development tasks. These cross-cutting topics were included as recurring agenda items for the ODIP workshops starting with the second workshop. The topics that were included were identified by the ODIP partners as being of importance for the marine data management community. These topics were:

- Vocabularies
- Data Publishing and Citation
- Unique persistent identifiers for researchers

These topics are considered to be cross-cutting because these are relevant for data management in general and in particular for harmonising metadata and data descriptions and achieving semantic interoperability between different regional systems through mappings and ontologies. In addition the topics of data publishing/citation and the assigning of persistent identifiers to data sets are very relevant for encouraging researchers to release and publish their data sets in scientific literature and obtaining academic credit by means of citations.

Vocabularies: all three of the ODIP prototype developments are concerned with vocabularies and have been supported during the project by developments on vocabularies which have been coordinated by NERC-BODC (Europe), CSIRO (Australia), MMI (USA) and R2R (USA). Considerable progress has been made on exposing the model behind the SeaDataNet P01 vocabulary (usage parameters) to facilitate easier mapping of terms, and also with the further development of the SISSVoc search service (Australia) including an underpinning RDF demonstrator and a SPARQL endpoint. SISSVoc now provides access to around 60 vocabularies from the NVS 2.0 as well as from other authoritative vocabularies. Recently NERC-BODC has also released a new search facility for its NERC Vocabularies Service (NVS 2.0) which is also fully based on SKOS and RDF. This new facility makes it easier to retrieve terms in mappings and also to follow deprecations. In addition to these developments, the vocabularies team has also given extensive support to the ODIP prototype teams for adding new terms to the existing vocabularies, building mappings between regional vocabularies e.g. SeaDataNet and GCMD, SeaDataNet and CF standards, They have also created new vocabularies where needed e.g. for instruments to support SWE.

Data Publishing and Citation: this has also been identified as a topic of general interest for the ODIP community. There has been synergy between the relevant activities in ODIP and those being undertaken by both the Research Data Alliance (RDA) and the Belmont Forum. Focus has been on best practices for using Digital Object Identifiers (DOIs) as persistent identifiers for data publishing, citation of dynamic data sets such as ARGO floats, and mechanisms at regional and global scales for minting and managing DOIs. The discussions addressing this topic have been coordinated by WHOI (USA) with support from NERC-BODC (Europe), ANDS (Australia) and NCI (Australia).

Unique identifiers for researchers: discussions during the 2nd ODIP workshop confirmed that having persistent identifiers for people that are available in some form of catalogue is important for the marine data management community. There are currently only a few systems available for assigning unique persistent identifiers to researchers and these have examined in more detail by the ODIP community. Persistent identifiers for researchers will continue to become increasingly relevant in the marine data management community for use in metadata for data publications and in cruise summary reports, which at present make use of a free text field for this information. As the ODIP project has progressed it has become evident that the ORCID system for assigning unique persistent identifiers for researchers is the preferred option for many applications.

The cross-cutting themes that have been addressed as part of the ODIP project will also be addressed in the ODIP II project because the developments in these areas will continue to be very relevant for the marine data management community and the development activities in the new project.

As highlighted above, the activities to address these topics have been carried out to support the three ODIP prototype development tasks. As a result this deliverable has been refocused to document these activities and has been re-named *D3.3 Cross-cutting themes*.

1 Introduction

The Ocean Data Interoperability Platform (ODIP) project is managing an Europe / USA / Australia/ IOC-IODE coordination platform, the objective of which is to establish interoperability of ocean and marine data management infrastructures, and to demonstrate this coordination through several joint Europe-USA-Australia-IOC/IODE prototypes that ensure persistent availability and effective sharing of data across scientific domains, organisations and national boundaries.

The ODIP workshops, which are organized as part of Work Package 2, are instrumental in bringing together the representatives of the regional data infrastructures and other relevant experts for the purposes of reviewing and comparing the existing regional data management systems and the associated standards in order to identify the commonalities and major differences between them, and propose how to overcome these inconsistencies through the development of interoperability solutions and/or the use of common standards.

The ODIP description of work (DoW) includes two phases of prototype definition and, as a result of the first phase of this activity, three prototype development tasks were initiated that are described in deliverable *D3.1 Definition of prototypes 1*. However, due to the complexity and large scope of the first three prototype development tasks, it was decided during the second year of the project not to formulate further prototypes in the present ODIP project but instead to focus on a number of cross-cutting topics that were of direct relevance for the on-going prototype development tasks. These cross-cutting topics were included as recurring agenda items for the ODIP workshops starting with the second workshop. The selected topics that were identified by the ODIP partners as being of importance for the marine data management community were:

- Vocabularies
- Data Publishing and Citation
- Unique persistent identifiers for researchers

These topics are considered to be cross-cutting because these are relevant for data management in general and in particular for harmonising metadata and data descriptions and achieving semantic interoperability between different regional systems through mappings and ontologies. In addition the topics of data publishing/citation and the assigning of persistent identifiers to data sets are very relevant for encouraging researchers to release and publish their data sets in scientific literature and obtaining academic credit by means of citations.

As highlighted above, the activities to address these topics have been carried out to support the three ODIP prototype development tasks. As a result this deliverable has been refocused to document these activities and has been re-named *D3.3 Cross-cutting themes*. Details of the presentations and the outcomes of the discussions of these topics during the ODIP workshops are also reported fully in deliverables *D2.4, D2.6* and *D2.8* which are the minutes and actions of the ODIP workshops.

.

2 Controlled Vocabularies

2.1 Introduction

Use of common vocabularies in all metadatabases and data formats is an important prerequisite towards consistency and semantic interoperability. Common vocabularies consist of lists of standardised terms that cover a broad spectrum of disciplines of relevance to the oceanographic and wider community. Using standardised sets of terms solves the problem of ambiguities associated with data mark-up and also enables records to be interpreted by computers. This opens up data sets to a whole world of possibilities for computer aided manipulation, distribution and long term re-use. Common vocabularies have to be controlled for consistency, which includes both content governance (concept population and semantic descriptions) and technical governance (content storage and distribution) services. Developments on controlled vocabularies for the marine and ocean domain are undertaken in all three participating regions (Europe, USA and Australia).

2.1.1 Potential actions as identified in the 1st ODIP Workshop (February 2013)

Vocabularies are very important resources for the ODIP prototype developments and the wider marine and ocean communities. In the brainstorming at the 1st ODIP Workshop a list of potential actions with respect to vocabularies was formulated:

Action1-1: Implementation of SPARQL technology and mappings between vocabularies (SKOS):

- Establish further SPARQL end-points for the exposure of controlled vocabularies; allowing simultaneous submission of queries to different vocabularies
- Organise these into a federated network
- Build user-friendly tools to query this federated network
- Setting up a pilot portal with mappings to demonstrate SPARQL

Action1-2: Establishing thesaurus-based semantic aggregation of data marked-up using the NVS 2.0 /SeaDataNet parameter usage vocabulary (P01)

- Develop a well-governed controlled vocabulary of terms for aggregated data products, with particular reference to EMODnet mapped to P01.
- Look for other applications for this approach across the ODIP community

Action1-3: Formally document vocabulary governance within the NERC Vocabulary Server (NVS 2.0)

- NVS 2.0 is a very important resource for the wider ODIP community
- Prepare documentation for content management and governance, include tracking history

- Refer to the ISO19135 governance model
- Guarantee sustained service by NERC for at least 10 years

Action1-4: Harmonisation of the conceptual models and controlled vocabularies used for event logging on research vessels with particular reference to Eurofleets and R2R

- Compare and harmonise conceptual models
- Harmonise controlled vocabularies used for events
- Establish governance for these controlled vocabularies

Action1-5: Develop a unified approach to the utilisation of controlled vocabularies under NERC Vocabulary Server governance in GeoNetwork.

Action1-6: Develop a unified approach to the utilisation of controlled vocabularies under NERC Vocabulary Server governance in other metadata standards such as O&M and SensorML

Action1-7: Develop and expose a conceptual model for the SeaDataNet P01 Parameter vocabulary

- P01 consists of concatenated terms, following a conceptual model. The number of concepts can be increased considerably, e.g. for water quality and contaminants in biota.
- There is a need to increase the visibility of the underlying model and make it more accessible. This will make it easier for data centres to map to these vocabularies and also submit new entries, including the possibility of using external vocabularies in components (e.g. WORMS for marine taxonomy)
-

2.1.2 Progress on vocabulary developments to support ODIP

During the ODIP project a number of the actions identified during the 1st ODIP workshop were addressed by the project partners and subsequently reported at the ODIP Workshops. Progress on these activities as reported at the workshops is documented below.

- **Status of controlled vocabularies at the time of the 2nd ODIP Workshop (December 2013)**
  - **Status in Europe Roy Lowry (NERC-BODC)**

Common vocabularies were set-up and populated by SeaDataNet (see http://www.seadatanet.org/Standards-Software/Common-Vocabularies and http://vocab.nerc.ac.uk/). Vocabulary technical governance initially was based on the NERC (DataGrid) Vocabulary

Server (NVS), which was originally developed in 2006. The service includes multiple lists of standardised terms of relevance to the oceanographic and wider community. In order to support the requirements of the user community, several enhancements were required to the existing NVS, and therefore a version 2.0 (NVS 2.0) was developed by NERC-BODC. The major upgrades delivered by NVS 2.0 consist of:

- a move to the accepted version of the World Wide Web Consortium's (W3C) Simple Knowledge Organization System (SKOS) specification for encoding the data dictionaries and taxonomies served through the NVS
- the ability to serve multilingual titles and definitions for resources
- the provision for mappings to external resources enabling the results of ontology extension to be delivered.
- correction of flaws identified in NVS 1 that allowed multiple URIs to be assigned to a single concept.

NVS2.0 has been operational and stable since early 2013 and SeaDataNet has been migrating to using NVS 2.0 from October 2013 onwards. In its first year of operation NVS 2.0 has received 860,000 calls from 1268 IP addresses with a balance between the use of the SOAP and REST protocols. All SeaDataNet repositories (CDI, EDMERP, CSR, EDMED, and EDIOS), formats (ODV, NetCDF), tools (MIKADO, NEMO, ODV) and services now use NVS 2.0 which has better version management and truly unique URNs for concepts. The parameter description vocabulary used by SeaDataNet (P01) includes approximately 28,000 terms (December 2013) and more have also been added for handling biological terms in SeaDataNet. A new vocabulary for aggregated parameters defined as mappings to P01 terms (P35) was initiated and populated with one test term. It will be populated further with the priority being given for parameters in use in the EMODNet Chemistry project. The P35 vocabulary supports a fully automated aggregation and validation process for data sets marked up using P01 vocabulary gathered from multiple sources. This is a follow-up from Action 1-2. Progress has also been made on Action 1-7 concerning exposing the semantic model which underlies the P01 terms. The semantic model (O&M concept) was agreed and both JSON and RDF representations have been developed. Further work is planned on developing the user interfaces to make it easier for data centres to query and identify relevant P01 terms. Users should also be able to propose P01 extensions using a form which can be processed more easily as part of governance. The P01 vocabulary has already been extended by mapping 1779 ICES contaminants for the biota chemical/matrix combinations used in UK data. This has been done as part of the EMODNet Chemistry project which is aiming to populate the SeaDataNet portal with a complete set of marine chemistry data.

Development of the P01 semantic model was undertaken in consultation with CSIRO (Australia) who also provided feedback on the RDF representation. In addition CSIRO demonstrated the use of a SISSVoc facade over the NVS content which underpins the interoperability of the NVS 2.0. Interactions with partners from the USA provided information on the US NODC SKOS, and the mappings between the NVS 2.0 and MMI-ORR for SeaVox platforms and device categories have been created and loaded on both services.

For Action1-3 there is an on-going discussion in the NERC Information Strategy Group which is establishing a mission-critical register of services. NVS 2.0 is recognised as a candidate service and additional resources for development are also being put in place.

Further work is needed to address the NERC Vocabulary Server (NVS) governance with reference to the ISO 19135 standard.

- **Status in Australia by Simon Cox (CSIRO)**

CSIRO has set up the SISSVoc service (see http://www.sissvoc.info/). This is a follow-up of Action 1-1 and 1-2. SISSVoc is a Linked Data API for accessing published vocabularies. SISSVoc provides a RESTful interface via a set of URI patterns that are aligned with SKOS. These services provide a standard web interface for any vocabulary which uses SKOS classes and properties. SISSVoc provides web pages for human-readable views, and machine-readable resources for client applications (in RDF, JSON, and XML). SISSVoc is implemented using a Linked Data API façade over a SPARQL endpoint. This approach streamlines the configuration of content negotiation, styling, query construction and dispatching. There are many other vocabularies being served outside the ODIP community which can have a complimentary use. However there is a need to be cautious about different definitions and meanings. Examples of available vocabularies include OFKN which focuses on Linked Open Vocabularies, QUDT which offers a conceptual model for units, quantities, and dimensions via URIs and ChEBI which provides URIs for chemical substances. CSIRO has used these examples to show how a comparable semantic model to that previously presented by NERC-BODC for SeaDataNet P01 can be built. This has been converted to SKOS and is now included in the SISSVoc Search service (http://www.sissvoc.info/search.html) that CSIRO operates and has been developed within the framework of the eReefs project. It has URIs to authoritative vocabs such as ChEBI and DUTQ. Underlying SISSVoc are an RDF demonstrator and a SPARQL endpoint. CSIRO has also included around 60 of the NVS 2.0 vocabularies in SISSVoc which are accessed on-the-fly and not by local buffering. It had been concluded that vocabularies should be standardized, harmonized and published and that, where ever possible, the ODIP community should extend / re-use existing vocabularies. It is recommended that the P01 semantic model should be exposed by RDF via a SPARQL endpoint and NERC-BODC has agreed to work together with CSIRO on the relevant services required for the ODIP prototypes.

- **Status in USA John Graybeal (MMI)**

A leading project in this area is the Marine Metadata Interoperability (MMI) project (see http://marinemetadata.org). The MMI project was first funded by the National Science Foundation (NSF) in 2004. Today it continues to provide guidance, vocabularies and semantic services, with regular updates on events and news of interest to the community. There are also several other vocabulary development activities in the USA. The Rolling Deck 2 Repository (R2R) project is focusing on the vocabulary mappings required to publish R2R Cruise Summary Reports in the SeaDataNet 3.0 ISO schema. The BCO-DMO (WHOI) initiative has been looking at the concept of event logging for research vessels together with European colleagues (EuroFleets) as part of Action 1-4. A major challenge is that the existing conceptual models have significant differences and work is currently ongoing to overcome these issues. The two projects are also compiling a list of the terms used by the individual communities to create a single combined list. BCO-DMO is also undertaking several mapping exercises and developing a faceted search on top of SKOS. As part of the Shipboard Automated Meteorological and Oceanographic Systems (SAMOS) activities Florida State University is mapping the SAMOS QC flags on to those used by SeaDataNet. Scripps Institution of Oceanography (SIO) is making progress with mapping the R2R organisation and port vocabularies to those used in SeaDataNet (EDMO and C38

respectively). However, there are still a lot of missing mappings which should be populated as part of ODIP2 prorotype development task. US NODC is exploring using SKOS in their vocabularies and will also begin consolidating/ merging the controlled vocabularies used across the three NOAA National Data Centers (NODC, NGDC, and NCDC) as part of the merging of the three centers to NOAA's National Centers for Environmental Information (NCEI). MMI-ORR has successfully been migrated to TAMU-CC for operations and several new developments have taken place, such as VINE which is a vocabulary integration tool used for making mappings to external ontologies. There is also a shift from closed to open practices, such as using linked open data principles.

Note: Vocabularies are instrumental in all of the ODIP prototype development tasks. These vocabularies should be used and linked by means of concept URIs which return a document for each concept. The ODIP prototype development task should regularly check which vocabularies, concepts and mappings are already available and what additional ones might also be required.

- **Possible use of gmx:Anchor**

One issue is how to handle external references in metadata (e.g. ISO 19139). In the current SeaDataNet 3.0 ISO Schema for cruise summary reporting, use is made of different codings: some external references are stored with codeList while others are stored as text in a Character String tag, or directly as tags or attributes values. References to external directories or thesaurus could be done in a more homogeneous way by always using the **gmx:Anchor** tag which allows storage of a remote URL (as the external reference) and a label. Introducing this solution into cruise summary reporting would have significant implications for the service chain and tools and therefore it was agreed not to adopt it in the ODIP2 prototype development task at present. For the purposes of metadata validation (against vocabulary references) the reference can be checked by applying a schematron rule on a SKOS list or alternatively the URL in xlink:href status can be verified. If it resolves on the web this means that the reference exists, if not this means that the reference does not exist. This second option can also be easily incorporated in the schematron.

- **Status of controlled vocabularies at the time of the 3rd ODIP Workshop (August 2014)**

    - **Progress in Europe Roy Lowry (NERC-BODC)**

Further progress has been made with Action 1-7 concerning the exposure of the Semantic Model of the SeaDataNet P01 Parameter vocabulary. P01 is composed of multiple components. Exposure will make it easier for identifying mappings and missing entries. The primary objective is to build a set of 'one-armed bandit' reels. Then for mapping or new concept creation the P01 code becomes a 'spin of the wheels'. Each reel itself is a controlled vocabulary. These will populate the 'bandit' interface drop-down lists and form the basis of an RDF document describing the P01 code. So far the following wheels have been established:

- S02: parameter - matrix relationship (e.g. per unit wet weight of)
- S26: matrix (e.g. Water body [dissolved plus reactive particulate phase]) The building of this took nearly 6 months for various reasons
- S25: biological entity (e.g. Limanda limanda (ITIS: 172881: WoRMS 127139) [Sex: male Subcomponent: liver])

The next steps for the P01 exposure are:

- developing a manual 'one-armed bandit' mapping in cooperation with the EMODNet Chemistry project for describing contaminants in biota
- developing of the 'substance' wheel. Looking at integration of ChEBI into the semantic model
- developing of 'Parameter' wheel, which is fairly trivial.
- RDF encoding (needs a developer)
- 'Bandit' automated mapping tool (needs a developer)

More progress has been made with Action 1-2 on Semantic Aggregation. For EMODnet Chemistry two vocabularies have been developed: P35 – parameters and P36 – themes. Each P35 concept is mapped to the P01 concepts that may be aggregated to produce it. So far (August 2014) P35 is populated with 109 entries and growing. The ODV software is being enhanced by the Alfred Wegener Institute (AWI) in Germany to use the P35 mapping to automate parameter aggregation (which is currently a laborious manual process) when aggregating data sets from multiple sources. The P35 concepts may provide a common denominator for semantic interoperability.

As part of Action 1-6 efforts have been undertaken for Instrument Mapping to support sensor web enablement (SWE). The strategy is to extend the L22 vocabulary to cover all the devices in use by R2R and IMOS. This is a crude but effective semantic harmonisation. As of 01/01/2014 L22 had 700 entries. IMOS requested an additional 51 entries (plus a further 18 which were requested later). R2R requested an additional 120 entries. All entries also have to be mapped to L05 (device types). The L22 mapping is quite an extensive challenge and efforts are combined between BODC, IMOS and R2R. The initial work is substantial but it is expected that it will be possible to sustain L22 because the number of new instruments entering the market is not that large.

A new activity is Parameter Mapping between IMOS netCDF data variables and P01**.** This exposed a fundamental problem of OP (Observable Property) semantic labelling namely accurate identification of the OP! This has been an issue with BODC semantic mark-up for over 30 years. For instance there are different words for the same thing such as 'optical backscatter' and 'optical side-scatter', and loose labelling by scientists such as calling 'nitrate plus nitrite' 'nitrate'.

The last activity is SeaDataNet Format Linkages. The SeaDataNet ODV (ASCII) format has data columns mapped to vocabulary URNs in the semantic header, for P01, P06, L22 (optional), and L33 (optional). These linkages are solved by a specific expression, while the whole file may be linked to multiple external resources which are solved in the ODV format by using xlink. In the NetCDF format data channels are mapped to vocabulary URNs plus human-readable labels through SDN namespace parameter attributes, while whole file linkages might be solved by xlink expressions. This implies that there is an issue with linkages in SeaDataNet. It is proposed that linkage should include human readable labels (included in xlink anchor syntax), URL (e.g. xlink:href), and information telling the client what

to expect at the end of the URL (e.g. xlink:type). Improving needs to be evolution, not revolution with full backward compatibility because there are millions of files in the existing system. Possible next steps could include optional parameter attributes like sdn_parameter_url and optional human-readable labels into SDN_XLINK strings.

- **Progress in Australia with mapping AODN Parameter Names to Observable Properties Ontology Kim Finney (AODN)**

In AODN, Parameter Names generally encompass sensed or assigned properties of 'Objects of Interest'. Simplistic 'parameter' names just include a sensed property, like 'concentration' plus the Object Of Interest (e.g. Concentration of Carbon). But often 'what' the Object Of Interest is AND 'where' it is being measured is described (e.g. Concentration of Carbon in Seawater). Seawater being considered a Feature of Interest (FOI) in an O&M modelling sense. Mostly this naming convention accords with SeaDataNet's P01 parameter discovery vocabulary'. Other semantic entities that need to be closely coupled with a 'Parameter Name' are the methods used to determine the sensed or assigned property and units of measure, but in the AODN these semantic components are not generally aggregated to form a 'Parameter Name'. In IMOS Parameter Names (and other closely associated vocabulary terms) are being used to mark-up dataset metadata and are also used to map to locally named 'within dataset variables'. It is the aim of AODN / IMOS to ensure 'interoperability' between AODN data descriptions and that of other data publishers. This is done by trying to re-use vocabularies where suitable and requesting additional terms be added to these existing vocabs (e.g. BODC Instrument and Platform vocabs) when terms are missing. But AODN / IMOS also wants to be able to use the new Observable Properties ontology (and reuse some of its term instances) to act as a common bridge between their Parameter Names and those used by others. The analysis has led to a number of issues which might require a modified Observable Properties ontology. These issues have been raised by AODN/IMOS in order to get more guidance from CSIRO and NERC-BODC on how to apply the ontology because they want to use it. AODN applauds the development of OP and its simplicity of design in order to encourage easy uptake but feels that the simplicity/ flexibility in some cases is hampering its application. Therefore AODN wants to work with others to better understand how to apply it.

- **Progress in USA with MMI (update) John Graybeal (MMI)**

MMI has prototyped the CFSN - Climate and Forecast Standard Names Viewer. It gives fast and clean access to the CF standard names and presents a view across multiple semantic vocabularies. See http://mmisw.org/cfsn. It is a browse and a search tool, but also a concept exploration tool because the quick links to re-used CF concepts improves understanding. It is also a Term (URI) Interoperability tool linking standard names to US and EU resources. MMI maintains the Ontology Registry and Repository (MMI-ORR).

There is a CF standard names committee responsible for content (Roy Lowry and John Graybeal are among its members) and there is an officer in charge of making changes. There is a system to track proposed changes, and when the changes are accepted, the new versions of CF standards names are published at the CF web site. Then these are automatically rolled out at NERC NVS 2.0, while MMI manually harvests them and serves them via ORR.

There is the question where people should host their mappings: at MMI-ORR or at NERC-NVS 2.0. There is no direct answer because it also depends on the context. Both systems are well established and provide partly overlapping and partly complementary vocabularies and mappings. NERC-NVS's approach provides a regulated submission process

(vocabularies are validated), whereas MMI's ORR provides limited controls but more open submissions.

- **ODIP progress in USA for SAMOS Vocabulary Mapping Jocelyn Elya (FSU)**

SAMOS stands for Shipboard Automated Meteorological and Oceanographic System and it is supported by NOAA, U.S. National Science Foundation, and the Schmidt Ocean Institute. SAMOS records high-quality navigational, marine meteorological, and near-surface oceanographic observations from research vessels. The objectives as part of ODIP are to map their SAMOS controlled vocabulary terms to internationally served vocabulary terms (parameters and quality control flags) and to publish RDF resources for SAMOS controlled vocabularies. Additional work on Data File Access will publish RDF resources for the download locations of SAMOS data files and to make it searchable by controlled vocabulary terms, time, and location. Good progress has already been achieved: all 25 SAMOS quality control flags have been mapped to SeaDataNet measure and qualifier flags (L20), 27 out of 38 SAMOS parameters have been mapped to CF Standard Names (SeaDataNet P07), BODC Parameter Usage Vocabulary (P01) and SeaDataNet Parameter Discovery Vocabulary (P02). Also RDF resources have been created for these. For Data Access SPARQL endpoints have been set up that can be queried by time, location, and parameter. The RDF resources for data files contain URL for data download. Mapping 'Time' appears to be quite a challenge because of many components and options. Another challenges met was how SAMOS Quality control flags are mapped to the SeaDataNet Flag "bad value" flag.

- **ODIP progress in USA for R2R Vocabulary Mapping Renata Ferreira (UCSD)**

The University of California-San Diego has made progress for mapping R2R vocabularies to SeaDataNet for supporting the adoption of the Cruise Summary Report (CSR) system as part of ODIP2 prototype. Focus has been on four vocabularies in particular: Device model (SeaDataNet L22), Organisation (SeaDataNet EDMO), Person (person id systems) and Port (SeaDataNet C38). For Persons there are a number of professional systems that have been considered: ORCID, Researcher ID, Research Gate, Google Scholar, Scopus, Linkedin. Results so far (August 2014):

- out of 200 ports 193 were matched while 7 had to be added to C38
- out of 160 device models 24 were matched, 125 proposed for addition to L22 and 11 remain unmatched. The 125 submissions were fully prepared with description etc. The 11 remaining terms have no documentation (so far). There were a number of challenges such as devices with multiple components (i.e. MET stations) and historical devices with missing documentation.
- Out of 856 scientists there were 28 initial matches, 723 with no matches and 105 with insufficient metadata in ORCID. Due to the use of multiple leading person id systems a new direction was taken by R2R by asking every scientist to declare their registration identifiers in any of the leading systems (ORCID, Researcher ID, Research Gate, Google Scholar, Scopus, Linkedin). These are then registered in the R2R person database.
- Out of 405 organisations there were 70 initial matches to EDMO, 173 US organisations proposed as new so far while work is ongoing for the remaining organisations (also non-US). It was agreed that UCSD will prepare new EDMO

entries no matter what country and will submit these to MARIS as overall EDMO manager for further processing. US entries will be added to the R2R account in EDMO CMS for completion. Other non-US entries will be divided by MARIS over existing EDMO CMS accounts for uptake and completion. IODE OceanExperts might potentially also might provide a resource to identify organisations. MARIS has recently upgraded the EDMO User Interface to also better display the non-European organisations and extra display on Google maps and by matrix of associated services. See: http://seadatanet.maris2.nl/v_edmo/welcome.asp

The mappings will be used in the following R2R services:

- ISO Records - detailed description of each cruise, suitable for long term archives
- Web Feature Services – track line geometry for each cruise, suitable for GIS clients
- Linked Data - detailed description of each cruise, suitable for Semantic Web clients.

- **Status of controlled vocabularies at the time of the 4th ODIP Workshop (April 2015)**

  - **ODIP Progress in Europe by Roy Lowry (NERC-BODC)**

Semantic Model Exposure activity: the primary objective is to build a set of 'one-armed bandit' wheels. The wheels built to date are:

- S02: parameter - matrix relationship (e.g. per unit wet weight of)
- S26: matrix (e.g. Water body [dissolved plus reactive particulate phase])
- S25: biological entity (e.g. Limanda limanda (ITIS: 172881: WoRMS 127139) [Sex: male Subcomponent: liver])

The next steps for the P01 exposure are:

- Development of the 'substance' wheel. Looking at integration of ChEBI into the semantic model with proposed strategy
  - Validate/clean ChEBI/CAS mapping mined from eReefs
  - Expand P01 semantic model CAS coverage and include as an explicit field. CAS is becoming a very powerful identifier for chemicals.
  - Generate mappings between P01 URIs and ChEBI URIs based on CAS (will appear in P01 SKOS documents)
- Development of 'Parameter' wheel which is fairly trivial
- RDF encoding for the exposure of semantic model - draft based on Compound Observable Property (INSPIRE extension to O&M)
- 'Bandit' automated mapping tool
- Research reasoning mapping moderator

These activities will not be finished in the short term and therefore will continue in the ODIP II project.

Instrument Mapping: the mapping is quite simple because the instruments names are the same. The strategy is to extend the L22 vocabulary to cover all the devices in use by R2R and IMOS. Work is progressing and since August 2014 another 180 native concepts have been added to L22 and there are now 889 instruments described in the vocabulary, providing a relatively rich semantic resource. Work will continue between BODC, IMOS and R2R as resources become available.

Parameter Mapping, BODC engaged with ICES on a 'Bandit wheel' mapping exercise for the contaminant in biota database. There are over 800 combinations in ICES (150 for priority biota for EMODnet). As a result 683 new concepts were registered in P01 in less than 2 days. This is a very effective way of doing a mapping because rather than looking at separate long strings, the two semantic models can be abstracted and mapped. 8 ‘green A total of eight invalid model element combinations were identified in the process and eliminated from the ICES data base.

SeaDataNet Vocabulary Linkage Issues: a linkage to a vocabulary URI should include a human readable label (included in xlink anchor syntax), a URL (e.g. xlink:href), and information telling the client what to expect at the end of the URL (e.g. xlink:type). Xlink anchor does this job and is also becoming quite a popular solution. SeaDataNet ISO19139 XML documents use ISO codelist linkages. It is therefore recommended that an extra element is added e.g. Xlink anchor linkages (and keep the ISO codelist as it is) the next time the schema is revised . This will be a huge step forward for interoperability.

The NVS 2.0 RESTful interface provides URIs in the form of URLs for each controlled vocabulary (e.g. http://vocab.nerc.ac.uk/collection/P02/) and each controlled vocabulary concept (e.g. http://vocab.nerc.ac.uk/collection/P02/current/TEMP/). The URLs resolve to RDF documents that include content negotiation which makes them human-readable when accessed using a browser whilst remaining machine-readable by applications such as Protégé. The SOAP WSDL may be found at http://vocab.nerc.ac.uk/vocab2.wsdl. The SPARQL endpoint may be queried using the client at http://vocab.nerc.ac.uk/sparql/. Note that this link (and other links to vocab.nerc.ac.uk) may not work with Internet Explorer as the default browser.

The original SeaDataNet search client was built by MARIS and can be found at: http://seadatanet.maris2.nl/v_bodc_vocab_v2/welcome.asp. Recently NERC-BODC has developed a further NVS 2.0 search client

https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search/

This provides the following additional functionality:

Vocabulary search in addition to concept search

- Operation off a SPARQL endpoint
- Access to all vocabularies in NVS, not just the SeaDataNet subset
- Advanced user control over search behaviour to allow the hit count to be reduced (especially useful for EMODnet chemistry lot searches in the P01 vocabulary)
- Deprecated concept management
- Ontology browse functionality using all concept mappings in NVS

- **ODIP progress in Australia - Vocabulary Creation and Management (VOCRAM) Project Roger Proctor (UTAS)**

The VOCRAM Project aims to improve the end to end vocabulary management process. VOCRAM is an Australian project which started in September 2014. It was suggested by eMII, but is coordinated and led by the Australian National Data Service (ANDS). Other partners are: CSIRO, IMOS, TERN, ALA, BoM. ANDS coordinates all public accessible research data across Australia and provides tools and services to people for making use of these research data. IMOS does not have a mechanism for an and to end process for creating, managing, and publishing vocabularies. ANDS undertook the role to provide such services as they already had a tool (SISSVoc) for publishing vocabularies. The goal is to deliver 'software as a service' infrastructure that dovetails with other components ANDS is developing, or refining (e.g. vocabulary catalogue and publishing services). Once complete it will provide widely accessible vocabulary services that can be used to obtain further support for involvement in ODIP II.

Central in the process is the ANDS Vocabularies Services Cluster and part of the process is the functionality labelled 'Editor' for the management of the repository of vocabularies, before publishing. VOCRAM is mainly addressing the functionality 'Editor'. In addition to the Vocabularies Services there are also tools for creators and providers to put information into the Services and also tools for consumers to access these vocabularies. After evaluation of several software packages a choice was made to use Pool Party (https://www.poolparty.biz/), a commercial tool but with an academic license available at a reduced cost. There are constraints such as limitations to create URIs but in general it is a very flexible and useful tool. There is now an end to end process which starts with providing some vocabulary concepts, built in house, that are exported to skos files in Pool Party. These are then managed in Pool Party, imported to the ANDS repository and then published to SISSVoc. They are now building the interfaces between these components and the pilot service will be operational by the end of September 2015.

- **ODIP progress in Australia – CSIRO Vocabulary deployment patterns and governance challenges Simon Cox (CSIRO)**

A CSIRO project deals with deploying and publishing vocabularies. The methodology is not to create new vocabularies but to find what exists and, in the context of linked data, to create URIs for them. The sources can be published vocabularies as web pages, such as the International Units System (SI). In this case the URIs are actually addresses for web pages that describe the concepts and if these addresses change, then the link does not work. This was found for the SI definition of meter that changed in the last 6 months. Another case is the geologic time scale from International Chronostratigraphic Chart, that is published as PDF (coloured) with an html table behind it which is very rich in geological terms. GCMD is another possible source for getting URIs in RDF version through the csv published vocabularies. ChEBI (Chemical Entities of Biological interest) is also doing a good job of publishing terms. The database and ontology of ChEBI includes 30 000 unique chemicals and is a very good example of existing URIs. To list chemicals you only need to define subsets but not URIs because they already exist. The main challenge is how these can be formalized and encoded. Vocabularies are retrieved from various authorities, such as Bureau of Meteorology, National Archives, ABS etc. Focus is on vocabularies that are maintained in an Excel spreadsheet, and hosted online with no machine-readable information. The first step is to order the information contained in the spreadsheet in a manageable way e.g. break single columns that often contain multiple pieces of information

so that each column only contains a single property. The second step is to insert CSV to RDF 123. This program encodes the spreadsheet into an RDF format. This is done by importing the spreadsheet, creating relationships between the columns using the graph and providing it with the prefixes it will use. Finally some minor editing via a text editor and/or RDF editor by people using their judgement and the vocabulary is ready for publish online.

The related architecture starts with the source vocabulary (in csv, html, txt format) that is converted into a formalized vocabulary (skos/rdf). The formalization is done by people using the methodology described above. It uses a Linked Data Registry (LDR) tool to manage data and metadata in the database (triple-store). An API is used to load data into the database and keep tracks of the registries. The content is exposed through a SPARQL endpoint to a SISSVoc service. Two users interfaces are used, one for search and the other for the linked data.

The registration process is the management of definitions. Vocabularies are just lists of concepts and in the world of linked data what actually is needed is the sub-setting of these lists and their management in a transparent and organised way. This process is often referred to as 'registration' and the 'register' is a managed list (ISO terminology). The result of a registration process is as stable identifier. Identifier is issued when requirements for the register are satisfied. The content has to be valid, with no duplicates and adequate definitions. However, the adequacy of definitions cannot always be algorithmically tested.

- **ODIP progress in USA with R2R Karen Stocks (SIO)**

The R2R team is a collaboration of four Institutes (LDEO, FSU, WHOI, and SIO). Most of the USA ODIP funding has been to support five students undertaking work for ODIP. Jocelyn Elya has been mapping SAMOS controlled vocabulary terms (QC flags and parameters) of meteorological data from underway vessels to internationally served vocabulary terms. All quality control flags were mapped to SeaDataNet measure and qualifier flags (L20). A total of 27 out of 38 SAMOS parameters have been mapped to CF Standard Names (SeaDataNet P07) and BODC Parameter Usage Vocabulary (P01). For the unmapped parameters the challenges are that many "Time" terms are too broad or too specific and that attributes needed to be included in the parameters. The second student, Nkemdirim Dockery created SPARQL endpoints to allow the SAMOS data to be queried based on ship, time, geographic footprint, and parameter. This prototype has been completed and is to be published. Renata Ferreira mapped R2R terms to SeaDataNet vocabularies in support of the adoption of the Cruise Summary Report (CSR) system as part of ODIP2 prototype development task. The mapping of the R2R Port Vocabulary (UNOLS) to the SeaDataNet Ports Gazetteer (C38) has been completed, it was relatively easy and only seven new terms needed to be added to C38. The R2R Organization Vocabulary mapping to the European Directory of Marine Organisations (EDMO) has been completed with 242 new terms added to EDMO. About 125 new R2R Device Models were mapped and are in the process of being added with full documentation to the NVS SeaVoX Device Catalogue (L22). The fourth vocabulary was for people and an effort was made to map chief scientists to ORCID identifiers. Out of 854 scientists there were 28 initial matches, 826 with no matches and 108 with name match but the metadata were insufficient to confirm the identify of matches in ORCID. A new direction was taken by R2R by asking every scientist through an email and a web form to self-report their personal identifiers in any of the leading systems (ORCID, Research Gate, Researcher ID, Google Scholar, Scopus, and Linkedin). The initial results indicated that Research Gate and Scopus were the most widely used. ORCID, Researcher ID and the others followed with less use. Linyun Fu, the fourth student created a prototype with a query interface to Linked

Data using Elda (human and API). The Elda interface sits on the top of a SPARQL endpoint. This prototype has been finished.

- **ODIP progress in USA with GeoLink - Semantics and Linked Data for the Geosciences Bob Arko (LDEO**)

The GeoLink project is a current US activity funded by the EarthCube Programme. The project is related to the publishing of content as Linked Data in which the basic idea is that the Web is the API. The work plan includes: publishing a set of reusable Ontology Design Patterns (ODPs) to describe geoscience concepts; upgrading existing repositories to publish their content as Linked Data, using those ODPs; and populate an integrated knowledgebase and exercise it against science use cases. Some of the partners such as BCO-DMO and R2R are involved in ODIP so there is a direct relationship between the two projects.

The basic concept of the project is to model the content, import the ontologies in and use this model for discovery. Currently the project is focusing on ships and cruises but is expected to extend at broader cross-sections of platforms and expeditions types. The cause of doing this is to poll existing controlled vocabularies and classify contents. The success so far is that no new vocabularies have been created. Existing gazetteers have been imported such as GEBCO, the Global Volcanoes Programme, the SCAR Gazetteers for undersea features (south of 60S), existing NVS vocabularies for device types, platforms types, parameter types. In terms of people, there is currently no suitable vocabulary and therefore the US NSF Awards (from 1970 to today) catalogue is being used. Some of the challenges faced are: lack of key controlled vocabularies published online with URIs and useful definitions; lack of physiographic gazetteers published online with URIs and proper RDF geometries; lack of Person and Organization identifiers published online with URIs and adequate metadata.

2.1.3 Continuation in ODIP II project

In conclusion, vocabularies have been a fruitful cross-cutting topic for discussions and developments over the lifetime of the ODIP project and in support of the ODIP prototype development tasks. It is therefore strongly recommended and decided to continue these discussions and developments in the ODIP II project.

3 Data publishing and citation

3.1 Introduction

Data publishing and citation has also been identified as a general topic of interest for the ODIP community. There has been synergy between these activities in ODIP and those in the Research Data Alliance (RDA) and Belmont Forum.

Following the "**Ocean Data Publication Cookbook of UNESCO IOC, Manuals and Guides 64":** A formal publishing process adds value to the dataset for the data originators as well as for future users of the data. Value may be added by providing an indication of the scientific quality and importance of the dataset (as measured through a process of peer review), and by ensuring that the dataset is complete, frozen and has enough supporting metadata and

other information to allow it to be used by others. Publishing a dataset also implies a commitment to persistence of the data and allows data producers to obtain academic credit for their work in creating the datasets. One form of persistent identifier is the Digital Object Identifier (DOI). A DOI is a character string (a "digital identifier") used to provide a unique identity of an object such as an electronic document. Metadata about the object is stored in association with the DOI name and this metadata may include a location where the object can be found. The DOI for a document is permanent, whereas its location and other metadata may change. Referring to an online document by its DOI provides more stable linking than simply referring to it by its URL, because if its URL changes, the publisher need only update the metadata for the DOI to link to the new URL. A DOI may be obtained for a variety of objects, including documents, data files and images. The assignment of DOIs to peer-reviewed journal articles has become commonplace.

- **Status of data publishing and persistent identifiers at the time of the 2nd ODIP Workshop (December 2013)**
  - **Status in Australia Lesley Wyborn (NCI)**

Research data citation in Australia is led by the Australian National Data Service (ANDS). They started investing in the establishment of research data collections in cooperation with research institutes and universities during 2009. The aim is that Australian researchers can easily publish, discover, access and use/re-use research data. Data citation is important because it facilitates reuse and validation of data, it makes it possible to track the impact and reach of data, it recognizes and rewards data producers, it increases academic and institution profile, and it connects all research outputs. Data citations were previously quite short and imprecise, but now tend to be more structured and precise, and also often include a **DOI** (Digital Object Identifier) to facilitate retrieving the data set. For example CSIRO publishes such an attribution statement (Data Citation) as part of its data access portal. This approach is also being adopted in the USA where we see increasing access to the results of federally funded scientific research. The National Science Foundation (NSF) now allows for citable data (i.e. with a DOI) to be listed as an outcome of research in a similar way to journal articles. The data set itself is considered as a product which needs to be citable and accessible. There is a need to build awareness among researchers and create a culture of data citation. ANDS is running a community building campaign which includes videos and flyers. In addition, ANDS is providing a Data Citation Toolkit which includes general information and functions for minting DOIs for use in the ANDS data citing service. It can also be used for implementing data citation across a wide range of institutions and researchers. There are a range of DOIs for scientific articles and increasingly DOIs are also being adopted for data sets e.g. Datacite.org already has more than 2 million DOIs for data sets. DOIs can also be used for linking a researcher to individual datasets. By registering personal details in a catalogue such as ORCID and linking to DOIs for software, publications and data sets an individual can build their research profile. Publishers are also increasingly adopting this approach and encouraging authors to include DOIs for their data sets in papers submitted for publication. The Scientific Data Initiative which will raise awareness of data citation and urge researchers to publish data sets will be launched in the USA during spring 2014. This initiative will also stimulate the establishment of approved repositories. Work is now underway to develop reliable data citation trackers across the different media to count the number of individual data citations.

- **Status in Europe Justin Buck (NERC-BODC)**

There is a push from publishers and scientists for data citation: publishers want to link journal articles to the data, while scientists want credit for data set creation and usage. The adoption of DOIs is a good approach and in particular the use of DataCite DOIs. BODC can issue DOIs for datasets in collaboration with DataCite as part of a wider NERC and global approach to dataset publication. For the purposes of data citation, datasets MUST be static, fulfil strict (meta) data requirements and the data sets must become freely available when a DOI is issued. These DOIs can be found at the BODC Published Data Library (PDL) webpages and are also available in the SeaDataNet EDMED directory of data sets. However, the scope of EDMED is wider and also includes restricted data resources and those without DOIs. The Ocean Data Publication Cookbook has been produced jointly by UK, USA and IODE. It gives the criteria and best practice for publication and citation of data, and is freely available from the IODE portal:

http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=10574

There are a number of unresolved issues relating to data citation including how to attach DOIs to open time series and also whether persistent identifiers are the same regardless of the status of the data, versioning, and granularity. This can be solved by distinguishing between two separate timelines: event / measurement / OBSERVATION time; ingestion / update / STATE time. DataCite provides a dynamic data policy to deal with these kinds of data sets; however there are still some caveats. BODC and IFREMER are currently analysing data citation for the Argo programme which produces more than 200 publications annually. The problem is how to cite Argo data at a given point in time. To solve this issue for the real-time data stream, IFREMER has minted a Digital Object Identifier (DOI) for the Global Data Assembly Centre (Argo GDAC) as a whole. These are sufficient for Argo if long term reproducibility of the data is not required by the user. However, IFREMER is also minting individual DOIs for monthly granularity snapshots at the GDAC level to enable reproducibility.

- **Status in USA Cyndy Chandler (WHOI)**

Data publication involves domain scientists, data managers and library scientists. It provides the opportunity to strengthen the bonds between professionals working in those communities. The challenge is to develop a system that supports proper citation of intellectual work that also encourages increased sharing of research data. The SCOR/IODE/MBLWHOI Library Project (USA) is assigning persistent identifiers to data referred to in traditional journal articles which are stored in institutional libraries, and where the data held by data centers are packaged and served in formats that can be cited. The project has developed a number of use cases to identify best practices for tracking data provenance and clearly attributing credit to data creators/providers so that researchers will make their data accessible. This gives incentives to researchers to publish their data sets. Data citation metrics are also starting to be collected and will potentially be adopted by funding agencies as part of performance measurement. Libraries have used DOIs for a number of years and are now becoming the de facto standard for data sets. There are also a number of relevant related activities including: continued support and interest from IODE/SCOR; the Research Data Alliance (RDA) Marine Data Harmonization Interest Group (led by Helen Glaves), the RDA Data Citation Working Group and the CODATA Data Citation Standards and Practices Task Group. A next step is to address the issue of identifiers for people. The Open Researcher and Contributor ID (ORCID) initiative is building a registry of unique researcher identifiers which will provide persistent identifiers for named people and facilitate links to other resources/outputs created by the researcher.

- **Status of data publishing and persistent identifiers at the time of the 3rd ODIP Workshop (August 2014)**

    - **Update for Australia Andrew Treloar (ANDS)**

ANDS has been in existence since 2009 and currently has around 40 staff. Its focus is on research data (data that researchers produce and use). ANDS provides training, advocacy, services, and policy support. It aims to transform data resources into research data that is available for easy publication, discovery, access, and use/reuse. ANDS manages a number of online services such as:

- Research Data Australia,
- Cite my Data DOI Identifier service,
- Vocabulary creation/management service including an API,
- Research Activity identifier service including an API,
- Developer toolbox

ANDS uses various identifiers such as WoRMS, DataCite, DOI, etc. Persistence of these identifiers requires: Systems plus Processes plus People. Persistent Identifiers (PIDs) must be reliably available over time and can best be seen as an indirection layer that reduces brittleness in getting to digital objects. For example a DOI can point to an object in a data store via a DOI resolver system. A Domain Name Server (DNS) is also a good example of such a resolver system. ANDS initially provided the Handles service for object identifiers, which is a service to mint DOIs. It is also a foundation member of DataCite. It is the Australian DataCite registrar, minting over 2000 DOIs each month. This is only done by a Machine to Machine interface. The actual management responsibility lies with data holders. ANDS is building a culture of data citation in Australia. The space is complex for organisation and person identifiers. Market momentum seems to be moving towards ORCID. ANDS is member of the ORCID Datacite Interoperability Network (FP7 ODIN) project.

    - **Update for Europe Justin Buck (NERC-BODC)**

BODC makes use of DOIs for dataset citation and publication of datasets on top of its traditional serving of datasets. Current BODC enhancements relate to uniquely identifying people: initially by ORCID identifiers and readily extensible to other identifiers. There is a challenge with DOIs for dynamic data such as with the Argo floats. There are more than 250 scientific papers referencing Argo each year. The challenge is how to cite Argo data at a given point in time. DOIs are given to Argo documentation, to data snapshots for which a month of data is chosen for enabling reproducibility, and to the mutating/growing data stream. This Research Data Alliance (RDA) Data Citation working group has also made progress with citation of dynamic data. There is a position paper by Andreas Rauber, Ari Asme, and Stephan Pröll avaialble at: https://rd-alliance.org/group/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html. The RDA conceptual model uses a database data infrastructure to save data queries. This effectively enables the user to roll back the data state to the time specified in the saved query. The query or the reference to the saved query can from part of a citation. The conceptual model appears sound; it ensures data reproducibility and citations can be provided at the point of data delivery. However it is designed for database infrastructures and file repositories while legacy data

infrastructures were not addressed. Three use cases with different characteristics are to be prototyped by the UK: the UK National River flow archive, the UK Butterfly monitoring network, and the Argo data system (simplified). For the third prototype, relating to Argo data, the US-NODC approach for the long-term archive of Argo data was presented (as proposed by Ken Casey before the RDA summer workshop). US-NODC wants to mint a single DOI for the Argo data archive. The archive includes a weekly snapshot of the full Argo database for the last decade (the granularity of this snapshot at weekly intervals is more than sufficient for most research). To cite a particular snapshot the user can potentially cite a time slice of the NODC archive i.e. the snapshot at a given point in time has a single DOIs which allows reproducibility.

- **Update for USA Cyndy Chandler (WHOI)**

The Research Data Life Cycle: scientists must be involved early in the process as this will motivate them to provide metadata; a published data policy helps. In the USA there is a series of recent directives from US federal offices and agencies 'encouraging' data sharing and publication. A good reference is a presentation by Dr. Ross Wilkinson (ANDS) delivered during the 3rd RDA Plenary in March 2014 in which states that there is a need to build the research data infrastructure first, then the policies and then use the improved infrastructure to motivate the researchers to be compliant with the policies.

BCO-DMO (US NSF funded Biological and Chemical Oceanography Data Management Office) provides a recent case study. The goal of the original project funded by SCOR, IODE and the Jewett Foundation in the USA was to identify best practices for tracking data provenance and clearly attributing credit to the original data creators/providers. Support for proper data citation was expected to provide additional motivation for researchers to make their data accessible. The assignment of persistent identifiers, specifically Digital Object Identifiers (DOIs), enables accurate data citation. BCO-DMO automated the export of metadata from BCO-DMO for deposit, with a copy of each dataset submitted to the Institutional Repository WHOAS. BCO-DMO (data repository) requests a DOI from the research library. Partnership allows the Library to work with a trusted data repository to ensure high quality data while BCO-DMO utilizes library services and is assured a permanent copy of the data is associated with the DOI. In the case of BCO-DMO the following are used:

- Persistent Identifiers for Data: a DOI resolves to a dataset landing page that describes the data. Landing page includes a pointer to a static copy of the actual data
- Persistent Identifiers for People: ORCID (Open Researcher & contributor ID) which is a registry of unique researcher identifiers that provides unique persistent identifiers for people. It can also enable linking to other resources created by the researcher.
-

There are the following resources are useful for Data Publication and Citation:

- Force11: Joint Declaration of Data Citation Principles:https://www.force11.org/datacitation
- CODATA report on principles of data citation "Out of Cite, Out of Mind" https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article

- ESIP Guidelines http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines
- DataONE: http://www.dataone.org/citing-dataone
- IODE/SCOR Data Publication (MG 64)
- http://iode.org/datapublishing

- **Update for USA with DOIs in R2R Bob Arko (LDEO)**

R2R experience with IDs in the datasets and publishing DOIs: the position taken is that data preservation and making data citable is no less important upstream. To enable re-use the original data should be persistent and citable. It is also important to clarify that a DOI only implies identity, not quality. Quality should be mentioned in the metadata. R2R registers (not mints) DOIs, reusing their internal dataset IDs and never deletes an ID internally. They follow the ESIP guidelines for publishing data sets and include a checksum manifest embedded in the DOI metadata. DOI metadata embeds a license, because scientists want to see an attribution guarantee, no commerciality, e.g. license to propagate with the product. They engage downstream data systems (post-field products) to embed R2R's upstream DOIs (original field data). This is not done for including R2R DOIs in the publications but to capture the metadata behind DOIs and track back to the original source data. This provides a metric of success for R2R and it helps to build the provenance chain back to the original field expeditions. The provenance chain is important to enable reproducibility of research results.

- **Status of data publishing and persistent identifiers at the time of the 4th ODIP Workshop (April 2015)**

- **Update from Europe (NERC data centres) John Watkins (CEH)**

The activities and case studies on data publication and data citation for the NERC Data Centres is driven by the need to have open and reproducible science. There is increasing pressure for visibility not only for science findings but for the data behind those science findings and for access to them. There is also increasing pressure for published results to be reproduced by commercial organizations to explore science funding. There is pressure on funding agencies to make sure that research is backed up by the data that was used to obtain the results. In 2012, the report of the Royal Society on "Science as Open Enterprise" talked about science in the internet age and the expectation of visibility of the supporting material. The expectation is that all parts of the science should be open. In response to this requirement the number of data journals (e.g Nature Data) that enable not only access to data but access to peer-reviewed papers about the data themselves, is increasing. NERC data centre is publishing Data Sets and Data Papers. Data citation means that the data centre is enabling people to cite data. Data are curated by NERC using DataCite DOIs issued by the British Library that are 'minted' by EIDC data repository at CEH on request. Data publication is making datasets accessible. The EIDC do this via the CEH Information Gateway. This is done with or without a DOI but must have metadata, standard data formats and supplementary information. Finally, the data centre encourages the publication of

descriptions of datasets as peer reviewed papers about datasets with DOIs if possible. So the concept is data reuse through data papers: the data centre has datasets curated and deposited at the Centre which have DOIs which enable citation. Those data sets can be cited from individual science papers and associated with these datasets will be peer reviewed data papers that could also be cited from the science journal and the science papers. So, part of the science itself will be not only the science findings but also the peer review descriptions of data products that link back to data themselves. A data paper describes the dataset not the science. It gives details of its collection, processing, software, file formats etc. There is no requirement for novel analyses or ground breaking conclusions. It gives the when, how and why data were collected, what the data product is and its limitations. A number of data journals are appearing, such as Nature Data, Geoscience Data Journal, Earth System Science Data Journal (ESSD). They require data to be held in an approved repository preferably labelled using a DOI. EIDC and the other NERC data centres are approved repositories for these journals. A case study was presented where research scientists (Christel Prudhomme and colleagues) published an ensemble of hydrological model outputs (a large dataset) as a data paper in the ESSD journal. The dataset was ingested into the EIDC Hub repository and given a DOI that resolves to a landing page on EIDC Hub website (10.5285/1514f-119e-44a4-8e1e-442735bb9797). The dataset DOI is then referred to in the data paper. The data paper has its own DOI that resolves to the online abstract for the paper in the journal (10.5194/essd-4-143-2012). The scientist (Christel Prudhomme) who wrote this data paper publication received the same level of interest in it as the science papers, and this can be seen as a way to increase collaborations. NERC is being pushed to recognize data papers as having a similar standing to the science papers as an important part of research.

Dynamic data citation i.e. how to cite particular subsets or versions of evolving data, is amongst the topics of interest for the Research Data Alliance (RDA) and in particular the Data Citation working group. The DOI is a useful tool in the context of citing dynamic data. The basic principle is that DOIs for data sets should be from the results of queries and not static files. The DOIs should have the ability of time-stamping for re-execution against versioned database, the ability of re-writing for normalization, unique-sort, mapping to history, and the ability to hash the result set for verifying identity/correctness. DOI is developing as a useful mechanism for dynamic data citation. Dynamic Data Citation is needed to deal with big data and sensor networks and this is very much work in progress. The work being done by the RDA Data Citation working group is promising but the reference implementations need further development. The NERC Data Centres are currently adapting the RDA model to their requirements. DOI dereferencing and citation metric need to be negotiated to ensure these work with the agreed syntax.

- **Update from Europe for dynamic data citation - Argo DOIs & tracing DOI usage Justin Buck (NERC-BODC)**

Publishers want to link journal articles to the data and scientists want credit for data set creation and usage. Most of data DOIs are from DataCite and in particular as defined in: http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.0.pdf.
This enables reproducible research and ensures trust in scientific research (see chapter by Adam Leadbetter in Collaborative Knowledge in Scientific Research Networks. DOI: 10.4018/978-1-4666-6567-5). The related RDA workshop showed that there are several prototype implementations such as database or text file based implementations. These do not exactly fit the Argo model which is based on repository files that continually grow, but can be adopted for other case study implementations. Argo uses DataCite DOIs. For the Argo current model, Ifremer is managing the Argo DOIs. All Argo documentation (manuals, cookbooks, etc) are available from Ifremer hosted site (argodatamgt.org). Ifremer also has

minted a DOI for the GDAC that grows and mutates and evolves, it is not pure DOI but for the moment it covers the needs of the real time data stream. For reproducibility at the moment monthly snapshots are used. Every month an entire copy of Argo data is put in repository and given a DOI, different for each month. So actually DOI resolve the granularity of Argo data, e.g. one month. At the landing pages at Ifremer there is useful information on “How to cite” (references to data with DOIs), “Is cited by” (manual done for the moment), and “Short DOIs” (like time URL). At the moment there are 20 Argo DOIs, one for each snapshot going back in time but they want to move to a single DOI, as the US NODC is proposing, for the ARGO Accession. To cite a particular snapshot one can potentially cite a time slice of the NODC archive i.e. the snapshot at a given point in time. It is proposed that the timeslice information be appended to the DOI reference, see for example how the NODC Argo Accession (0042682) is cited:

http://dx.doi.org/10.[NODC_REF]/[Argo_accession_DOI]/[time_slice _information].

A similar example from an NODC archive – an SST data set (GLOB OSTIA) was presented where individual granules are cited within single DOI and the different versions links lead to different landing pages. The method of the single DOI approach was presented at the RDA workshop in San Diego, April 2015. The principle has been verbally agreed with publishers. Awaiting NODC to implement and mint DOI and expose snapshots (hopefully this US fiscal year), resolving a particular snapshot via citation method is additional work. Since 1998, 2000 papers have been published using Argo data which means that a citation mechanism is needed. BODC contacted publishing houses, Springer (currently assimilating NPG), Elsevier, Royal Society, Wiley, and all unanimous is saying we need to get Argo data into a data paper and all want it to be their data paper. Different publications were shown on how to use DOIs to go back to the data (tracking data usage with DOIs). The first example was a paper from the Royal Society with a reference on Argo data (cited DOI) in the references. By typing the DOI in the full text search, the DOI is being traced easily. A second example was a Nature paper on Argo, published in January 2015 in WebScience (Thompson Reuters) where the citation about the data is embedded in the body text and not in the references. Trying to get the DOI that used to cite the data to a cited reference search and return the paper, the result was “you cannot perform a 'Cited Reference Search' using a DOI reference”. This should be raised with the publishers, as it is a fundamental limitation. Another issue with Springer (merged with Nature) is that DataCite DOIs are unknown to CrossRef. Tracking dataset DOIs registered by one agency (DataCite) in STM publication DOIs registered at another agency (CrossRef), has as consequences that: the current automatic processes of doing lookups and cited-by linking via CrossRef will not "see" the DataCite DOIs. An other option is Research Gate. Its power is that scientists effectively building additional linkages and extra references. Research Gate did not find the paper DOI. Finally Google was tried. Typing the Argo DOIs, it brings the link of the Nature paper only. So, he contacted Google to discuss if the “Custom Search API” could be used to build a tool that will search across all Argo data DOIs. But there are big caveats to this: when one searches on a DOI, Google has access to the full text of all papers journal –even for the closed one, but they only return the hits where the references are for open access journals or closed access journals where the citation is in the references part because it is public. The citation in the body text of the paper of closed journal is a problem, even if scientist put a copy of paper in Google to see it, Google is not allowed to expose it due to copyright reasons. So, it is not allowed to get the data reference of a closed access journal. The indexing between DataCite and CrossRef DOIs works:

http://crosstech.crossref.org/2014/09/linking-data-and-publications.html.

Currently it is a nascent effort but is linked to RDA. A data paper for Argo is needed. Then users will cite data paper and snapshot DOIs, ahead of the single DOI and update if necessary. ESSD can be used as it is an indexed journal which will allow searching. Google

"Custom Search API" is an option but there are caveats. ReseachGate is subject to similar search result but may have constraints as Google.

- **Update from Australia by Lesley Wyborn (NCI)**

Persistent Identifiers (PIDs): assigning PIDs to data is quite complicated, the common practice is indexing and putting books on the shelf. The Cite My Data service of ANDS allows registered (trusted) clients to mint, update & retrieve ANDS Digital Object Identifiers (DOIs) that identify research data. ANDS does not manage Digital Object Identifiers; only provides the infrastructure that allows minting and updating of Digital Object Identifiers in the global DOI infrastructure. Updating is the responsibility of the client that minted the DOI. Processes and policies need to be put in place by those utilising the product to ensure that appropriate maintenance practices underpin persistence. Users of the service are expected to have automated methods to both mint and update identifiers. Australian environmental data originated from several sources are replicated from Governmental Agencies and other Research Institutes. There is a joint statement of principle from Universities of Australia, Australian Research Management Society, Council of University Librarians and ANDS: they are going to use ORCID, its advantage is that this is for working group also, not only for individuals. The ODIP group can have an ORCID Identifier. If someone collects samples, IGSN's are their birth certificate. They provide a persistent identifier throughout life and uniquely identify the sample in global space. They also enable parent child relationships to be preserved. They can be applied to drill cores. Links to publications provide metrics on value of sample/cruise. We have to start thinking that identifiers preserve the values of the samples we collect and on which the data rely on. Another example of why we need them is the EarthChem Portal includes 75 samples with the name M1 (or M-1) and a particular sample has different names used in the publications. The identifiers ensure unambiguous citation of physical samples. Facilitate interoperability and linking of data at the level of individual samples. We can cite the value of the sample and argue for its preservation in the repository. Another example is that IGSN can play the role of a DOI in a publication. It can be dynamically linked with pictures and other information of the sample. This is really an added-value to a data. IGSN nationally has a suite of allocating agencies; five of them are Marine Institutes. Another issue is that when a data centre starts to publish data it has to assign multiple roles to those who are associated with the data collections for example who is distributor, who is originator, user, etc. The following are use cases where identifiers can help with the duplicate copies, copy and change and dynamic data sets:

- Scenario 1 (duplicate copy): we get a data set in NCI, we do not change, and one can use existing owner-minted DOI and push owner catalogue entry to NCI. It is a simple case.
- Scenario 2 (Copy and change), when a new data set is created in the NCI data centre but cannot push it back to the originator who developed it, one can: Include <lineage> information showing relationship to original, NCI mint new DOI and push NCI entry to owner catalogue.
- Scenario 3 (Dynamic dataset), when a data set grows, one can timestamp DOIs so as to know when new data were added.
- Scenario 4 (Dynamic changing dataset), it is not so much issue with geophysical surveys or satellite data (in these you just add new data), but when we go back to time and upgrade the analysis or change it. This case also applies to a data platform. There is no answer to this yet.

Data citation and PIs can help solve issues of unintended data mutation. When we have a lot of versions of the same data set, Data citation and PIs can help to identify which is the real data set. DOIs can help with arguments of the type: it's my data set, I am assigning the DOI, this is the issue in AUS between organizations. Who assigns the DOI and when can be controversial when objects move between agencies. For these reasons a data management developed by the Organizations and submitted to NCI where they came to a federated agreement on the governance of the data collections e.g. who is minting the DOIs, when, etc.

- **Update for USA by Cyndy Chandler (WHOI)**

As Persistent Identifiers for Data, the Digital Object Identifiers (DOIs) were chosen for data, it should not matter who assigns and mints the DOIs (i.e. CrossRef or Datacite). There are other identifiers but DOIs were chosen because the publishers recognise them. A DOI resolves to a dataset landing page that describes the data. At the beginning this was not clear but is now globally recognized as a best practice. The landing page includes a pointer to a static copy of the actual data, or a logical chunk of time-series data. Although there is not yet any consensus how the data set pointer (link) could be labelled and that having a machine-interpetable link format is needed for automated tracking. ESSD Earth System Science Data (earth-system-science-data.net) is a journal for publishing data. It is now being indexed by Web of Science and this is a big step forward. Librarians are promoting ORCID Open Researcher & Contributor ID as Persistent Identifiers for People. It is a registry of unique researcher identifiers. It gives persistent identifiers for person names. It can enable linking to other resources created by the researcher (http://orcid.org/). The Software and Data Citation Workshop took place in January 2015, Arlington, VA (USA) (https://softwaredatacitation.org/). The Workshop is funded by US NSF and Alfred P. Sloan Foundation. Its aim is to support Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution. There were interdisciplinary discussion and exploration of new norms and practices for software and data citation and attribution. The key point is that software is now part of the discussion, not just data. The Workshop based on 22 use cases that participants submitted before the workshop, 4 for software, 13 for data, and 5 for both (software & data). Breakout groups addressed each use case and summarized the challenges, why it was important, why not solved, and identified 3-5 critical Actions that could be implemented or recommended by the community. From these use case, the most relevant for ODIP are the Interoperable Frameworks that defined the following critical actions: a) ask federal funding agencies to require every PI to have a permanent human identifier (e.g. ORCID, which resolves critical issues of identifying individuals), b) coordinate an agreed metadata model for both software and data; then each repository can define its profile of that model, c) at a global level, establish a "Scientific Solutions Center" (a system of systems) supported by a common (REST) API that brokers between trusted, distributed software and data repositories to better support "Scientific Discovery through agreed Norms and Practices for Software and Data Citation and Attribution", d) Focus resources on bringing together (coordinating and funding) experienced experts to enable greater interoperability and searchability across repositories of scientific data and software objects. The use cases were put on a critical action matrix of impact and likelihood. PersonID recommendation came out as high impact and high likelihood (tractable).

4 Unique persistent identifiers for people

Current person ID systems are:

- ORCID (provides unique ID with machine interpretable content)
- ResearchGate: URL with name string (not a unique ID)
- Researcher ID (Thomson Reuters)
- Scopus Author ID (Elsevier)
- LinkedIn (role: linking people's information)

It is the responsibility of researcher to have awareness of the fact that if he put info on the Facebook, Twitter, etc there is possibility that this info will be linked with ORCID. ORCID provides the unique person ID that can be used by other systems (ResearchGate, Scopus and Researcher ID all have easy ways to add your ORCID). ORCID provides unique IDs for groups as well. ORCID has been thought as a complementary tool to researchers and the INSPIRE also. Research Gate by including ORCID as identifier for the person says that their identifier to their entity is not a person but an activity profile and this is the person that did this activity.

By registering personal details in a catalogue such as ORCID and linking to DOIs for software, publications and data sets an individual can build their research profile. Publishers are also increasingly adopting this approach and encouraging authors to include DOIs for their data sets in papers submitted for publication.

ANDS has shared a paper on the reason they chose ORCIDs,

http://ands.org.au/discovery/orcid-joint-statement-of-principle.pdf.

Note the Landing page shows that a group can have a number as well ORCID http://ands.org.au/discovery/orcid-jsp.html.

From the discussions and practices it appears that ORCID is taking the lead: ORCID (http://www.orcid.org) is an open, non-profit, community-based effort to provide a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers. ORCID is unique in its ability to reach across disciplines, research sectors, and national boundaries and its cooperation with other identifier systems. It provides two core functions: (1) a registry to obtain a unique identifier and manage a record of activities, and (2) APIs that support system-to-system communication and authentication. ORCID makes its code available under an open source license, and will post an annual public data file under a CC0 waiver for free download. The ORCID Registry is available free of charge to individuals, who may obtain an ORCID identifier, manage their record of activities, and search for others in the Registry. Organizations may become members to link their records to ORCID identifiers, to update ORCID records, to receive updates from ORCID, and to register their employees and students for ORCID identifiers. ORCID records hold non-sensitive information such as name, email, organization and research activities. The ORCID community includes individual researchers, universities, national laboratories, commercial research organizations, research funders, publishers,

national science agencies, data repositories, and international professional societies, all of whom have been critically affected by the lack of a central registry for researchers. ORCID coordinates with the community through Working Groups and bi-annual Outreach meetings.

Annex A Terminology

Term	Definition
CDI	Common Data Index metadata schema and catalogue developed by the SeaDataNet project
CSR	Cruise Summary Reports is a directory of research cruises.
CSW	OGC standard – Catalogue Service for the Web
GeoNetwork	An open source catalogue application for managing spatially referenced resources. It provides a metadata editing tool and search functions as well as providing embedded interactive web map viewer
GEOSS	Global Earth Observation System of Systems
IOC	Intergovernmental Oceanographic Commission of UNESCO (IOC/UNESCO).
IODE	International Oceanographic Data and Information Exchange (part of IOC)
IMOS	Integrated Marine Observing System: Australian monitoring system; providing open access to marine research data
MCP	Marine Community Profile
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting
ODP	Ocean Data Portal: data discovery and access service, part of the IODE network
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

OWL	The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things.
RDF	Resource Description Framework: RDF is a standard model for data interchange on the Web. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. The RDF specification consists of W3C recommendations and working notes.
R2R	Rolling Deck to Repository: a US project responsible for the cataloguing and delivery of data acquired by the US research fleet.
SeaDataNet	SeaDataNet: EU-funded pan-European e-infrastructure for the management and delivery of marine and oceanographic data
SensorML	OGC standard providing models and an XML encoding for describing sensors and process lineage
SKOS	Simple Knowledge Organization System: an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web
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
Triplestore	A triplestore or RDF store is a purpose-built database for the storage and retrieval of triples through semantic queries. Triples are usually imported or /exported using RDF.
US-IOOS	US Integrated Ocean Observing System
WebEx	On-line web conferencing and collaboration tool
WCS	OGC standard – Web Coverage Service

WFS	OGC standard – Web Feature Service
WMS	OGC standard – Web Mapping Service