Oceanographic and Marine Cross-Domain Data Management for Sustainable Development

Paolo Diviacco
Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS), Italy

Adam Leadbetter
Marine Institute, Ireland

Helen Glaves
British Geological Survey, UK

A volume in the Advances in Environmental Engineering and Green Technologies (AEEGT) Book Series

IGI GLOBAL
DISSEMINATOR OF KNOWLEDGE
www.igi-global.com

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Diviacco, Paolo, editor. | Leadbetter, Adam, 1979- editor. | Glaves, Helen, 1966- editor.
Title: Oceanographic and marine cross-domain data management for sustainable development / Paolo Diviacco, Adam Leadbetter, and Helen Glaves, editors.
Description: Hershey, PA : Information Science Reference, 2017. | Series: Advances in environmental engineering and green technologies | Includes bibliographical references and index.
Identifiers: LCCN 2016023682| ISBN 9781522507000 (hardcover) | ISBN 9781522507017 (ebook)
Subjects: LCSH: Marine resources--Management--Data processing. | Marine resources conservation--Data processing. | Sustainable aquaculture--Data processing. | Sustainable marine
Classification: LCC GC1018.5 .O25 2017 | DDC 333.91/64--dc23 LC record available at https://lccn.loc.gov/2016023682

This book is published in the IGI Global book series Advances in Environmental Engineering and Green Technologies (AEEGT) (ISSN: 2326-9162; eISSN: 2326-9170)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 4
Linked Ocean Data 2.0

Adam Leadbetter
Marine Institute, Ireland

Adam Shepherd
Woods Hole Oceanographic Institution, USA

Michelle Cheatham
Wright State University, USA

Rob Thomas
British Oceanographic Data Centre, National Oceanography Centre, UK

ABSTRACT

Within the theme of sustainable development, it is not desirable to either have data siloed in one location where it cannot be reused for purposes beyond which it was originally collected, or in a state where it cannot be integrated into a holistic view of the marine environment. As such, the links between datasets should be formally documented and exploited as best as possible. Given this, the use of Semantic Web technology and information modelling patterns are explored in this chapter with reference to the marine domain. Further, new strategies for adding semantic annotation to data in real-time are discussed and prototyped.

INTRODUCTION

Within the domain of marine data management, there is a history, beginning in the 1980s, of using controlled vocabularies to define the content of metadata fields and data file channels (UNESCO, 1987; Lowry, 2003; Merati & Burger, 2004; Lawrence et al., 2009; Schaap & Lowry, 2010). The availability of such vocabularies as Semantic Web resources through such platforms as the Marine Metadata Interoperability Ontology Registry and Repository (Graybeal et al., 2012) and the Natural Environment Research Council Vocabulary Server (Leadbetter et al., 2014) has led to the use of Linked Data (Berners-Lee, 2006) patterns to publish metadata and data concerning the marine domain. The Linked Ocean Data concept has been shown to facilitate distributed eScience through validation of chained web processing services and dynamic discovery and aggregation of datasets alongside increased impact of datasets through formalised links to the acquisition methodologies and associated publications interpreting the data (Leadbetter, 2015).

When considering sustainable development of the oceans, it is most important to be able to combine data from a range of producers in a single application. For example, the Marine Renewable Energy Portal

DOI: 10.4018/978-1-5225-0700-0.ch004

delivered by the Sustainable Energy Authority of Ireland and the Irish Marine Institute allows potential energy developers access to live dashboards of current and forecast wave, wind and tidal conditions using data from the Marine Institute, the Commissioner of Irish Lights and the United States' National Oceanic and Atmospheric Administration. However, these data are not connected in a formal Linked Data sense and moving in this direction would better allow users to incorporate other data sources, generate their own dashboards, and enquire of the data in new ways to generate new information and knowledge. Further, these principles can be extended to the creation of new decision support tools for planning operations on board research vessels, or at remote bases when considering the maintenance of missions of deployed Autonomous Underwater Vehicles.

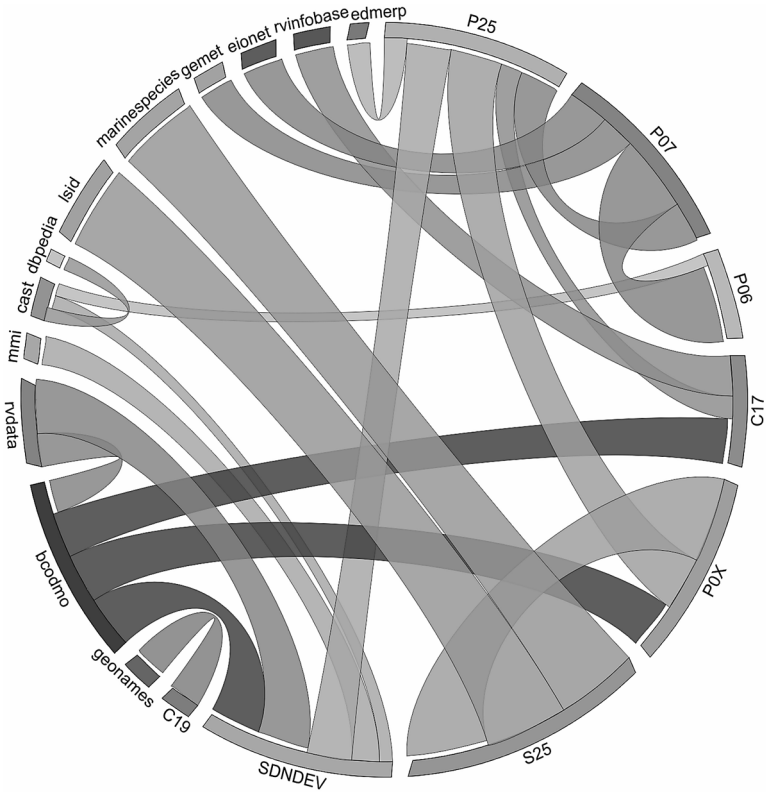
In this chapter we will expand on the existing Linked Ocean Data research to include the emerging concepts of heterogeneous data integration through the use of ontology design patterns, and publishing Linked Data from sources of real-time observations. This will include:

- Revisiting the current state of the Linked Ocean Data cloud to show how it has expanded since it was first proposed in 2013
- Conceptualising the Linked Ocean Data cloud using advances in interchange visualisation techniques (Zeng et al., 2013; Krzywinski et al, 2011) to give new insights into the interconnectedness of the data and information represented within the cloud
- Examining how the emergence of the concept of ontology design patterns has been successfully applied in the ocean science domain
- Considering the relationships which are forming between Linked Data and Big Data and how these have begun to be explored in marine science
- Why the Linked Ocean Data paradigm is important for a sustainable approach to the exploration and development of the ocean

THE STATE OF THE LINKED OCEAN DATA CLOUD

The Linked Ocean Data concept was introduced by Leadbetter *et al.* (2013) and was refined by Leadbetter (2015). The initial Linked Ocean Data cloud consisted of 18 nodes, and by the final publication listed above a further two nodes were incorporated. However, the visualizations used in these descriptions of the Linked Ocean Data cloud have not shown any quantitative information concerning the linkages between the nodes of the cloud, solely the qualitative fact that nodes are linked. However, prior to the introduction of the Linked Ocean Data cloud, Leadbetter and Lowry (2012) experiment with visualisations of connections between controlled vocabularies using circos plots (Krzywinski 2009) which were originally designed for use in visualizing genomics data, but have since been repurposed to show customer flow in the motor industry, volume of courier shipments, database schemas, and presidential debates. The 2012 mappings from the NERC Vocabulary Server to other external controlled vocabulary services is shown in Figure 1. While providing some qualitative information about both the number and directions of the connections in the space occupied by the Linked Ocean Data cloud, this approach is not intuitively interpreted by those users who are unfamiliar with the circos concept. More recently, Zeng et al. (2013) have developed the circos plot concepts to show interchange patterns in the movements of passengers through public transport networks. From the 2015 state of the Linked Ocean Data cloud, it could be seen that there are several nodes where an interchange of data links may occur. Therefore an

Figure 1. External mappings from the NERC Vocabulary Server to other Semantic Web resources (1 o'clock to 7 o'clock) and to the NERC Vocabulary Server (other sectors) as of December 2012
(after Leadbetter & Lowry, 2012). Ribbon width is representative of the number of mappings, on a logarithmic scale.



exploration of Zeng et al.'s interchange diagrams for the Linked Ocean Data space has been undertaken and is presented here.

The Vocabulary of Interlinked Datasets (VoID; Alexander & Hausenblas, 2009) provides a Semantic Web format for standardised descriptions of the connections from one Linked Data resource to another. The NERC Vocabulary Server (n.d.), which provides many controlled vocabularies used by Linked Ocean Data publishers, the Rolling Deck to Repository Project (n.d.) and the Biological and Chemical Data Management Office of the Woods Hole Oceanographic Institution (n.d.) all publish VoID documents making these data easy to collect and analyse. The state interlinkages between these Semantic Web resources as of 12th October 2015 is summarised in Table 1.

A further development of the circos visualisation by Krzywinski et al (2011) is the hive plot, which uses radial axes to collect related network nodes and the network edges are shown as arcs connecting these edges. This is useful, as the demonstration in Figure 2 shows that even with a small, but highly connected subset of the Linked Ocean Data cloud plotted in this way, there is still a difficulty in interpreting the information. As an example, the visualisation of the interchange paths between the Linked Ocean Data cloud nodes as shown in Figure 2 is also presented as a hive plot in Figure 3. One benefit of the hive plot approach over the other visualisation methods seen before is the hive panel view, which allows both the connections between the entire network of Linked Ocean Data Nodes and between given nodes in the system to be viewed easily in one plot group.

Table 1. The number of Linked Data relationships between nodes of the Linked Ocean Data cloud as of 20th October 2015, used to generate Figures 2 and 3

From Node	To Node	Number of Mappings	From Node	To Node	Number of Mappings
NVS-C19	geonames	126	BCODMO-Parameter	NVS-P03	4
NVS-L06	HeritageData	20	BCODMO-Parameter	NVS-P09	3
NVS-L06	MMISW-IOOS-Platform	26	BCODMO-Instrument	NVS-L05	93
NVS-L05	MMISW-TRDI-Glossary	2	BCODMO-Instrument	NVS-L22	33
NVS-L05	CAST	6	BCODMO-Instrument	NVS-C77	1
NVS-L22	MMISW-TRDI-Models	12	BCODMO-Platform	NVS-L06	168
NVS-C38	OrdSurv	205	R2R-Device	NVS-L05	90
NVS-P01	SISSVOC	58	R2R-Device	NVS-L22	24
NVS-P08	SISSVOC	1	R2R-Vessel	NVS-L06	1
NVS-P08	Programmes	10	R2R-Vessel	NVS-C19	42
NVS-P21	SISSVOC	3	R2R-Port	NVS-C38	258
NVS-P06	DBPEDIA	470	R2R-Holding	BCODMO-Deployment	362
NVS-P06	CAST	5	MMISW-CF	NVS-P07	2671
NVS-P25	CAST	1	MMISW-TRDI-Glossary	NVS-L05	26
NVS-S25	LSID	3050	MMISW-IOOS-Platform	NVS-L06	2
NVS-P07	EIONET	71	MMISW-TRDI-Models	NVS-L22	12
NVS-S27	CHEMDPlus	345	SISSVOC	QUDT	28
NVS-S27	OBO	269	SISSVOC	NVS-P01	84
BCODMO-Parameter	NVS-P01	101	SISSVOC	NVS-P02	22
BCODMO-Parameter	NVS-P02	5	SISSVOC	NVS-C38	1

Figure 2. A network interchange diagram (after Zeng et al., 2013) for four nodes of the Linked Ocean Data Cloud. Although there is now some clarity on the number and direction of mappings, the hub node of the NERC Vocabulary Server makes producing a coherent visualization of this type difficult

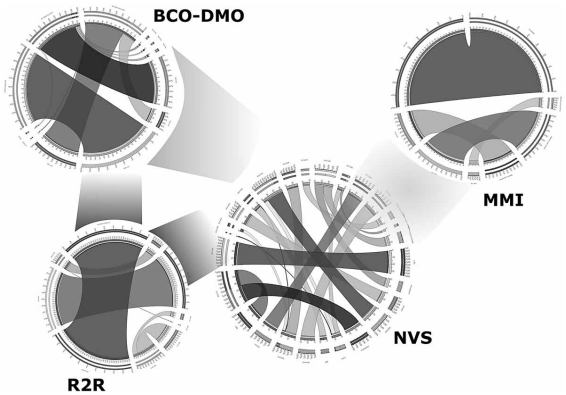
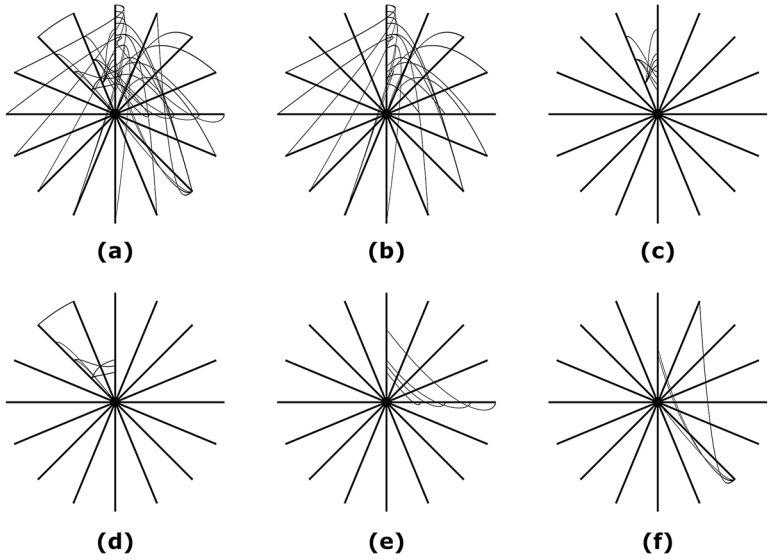


Figure 3. A hive panel showing the mappings between (a) all nodes of the Linked Ocean Data Cloud; and from (b) NERC Vocabulary Server; (c) the Biological and Chemical Oceanography Data Management Office (BCO-DMO); (d) Rolling Deck to Repository (R2R); (e) the Marine Metadata Interoperability (MMI) Ontology Register and Repository; and (f) CSIRO to other nodes on the Linked Ocean Data cloud. Clockwise from the top, the axes represent the NERC Vocabulary Server; NASA's Global Change Master Directory and Quantities, Units, Dimensions and Data Types; GeoNames; Heritage Data; MMI; Ordnance Survey; CSIRO; BBC; DBPEDIA; NERC's Chemical Analytical Services Thesaurus; Life Sciences ID; ChemDPlus; European Environment Agency; Open Biomedical Ontologies; R2R; BCO-DMO



Integration of new data sources to the Linked Ocean Data cloud may be expedited if new data sources can be modelled against patterns which have emerged from a combined computer science and marine domain effort. The development of these ontology design patterns are described below.

ONTOLOGY DESIGN PATTERNS

Many different fields that are part craft and part science have found a need to encapsulate "best practices" for dealing with problems that recur frequently in those fields. For instance, in the 1970s an architect named Christopher Alexander developed a set of architectural patterns for different situations, such as a vestibule that is light and airy but not too hot when exposed to direct sunlight (Alexander, Ishikawa & Silverstein, 1977). Nearly twenty years later, a group of computer scientists known as the Gang of Four published a book of patterns related to needs that frequently arise during object oriented programming, such as the decorator pattern, which allows a programmer to change the behaviour of a particular object without affecting the behaviour of other objects of the same class (Gamma *et al.*, 1994). A decade later, the success of design patterns in architecture and computer science eventually led to them being applied to the growing field of data science, as ontology design patterns (ODPs) (Gangemi, 2005). An ODP is a reusable solution to a data modelling problem that commonly occurs across many different domains

or within a wide variety of contexts within a single domain. For instance, concepts like Person or Event need to be represented in many different situations, while the concept of a Sample is key to many different geoscience datasets, from biology to chemistry to geology. When multiple scientific datasets need to be integrated, the join points tend to be precisely this type of core concept, which means that ODPs are very useful for data modelling and integration. This section will introduce some of the problems that arise when integrating scientific data, how ontology design patterns attempt to address these problems, and how they are being utilized in practice in the EarthCube GeoLink project.

Challenges of Scientific Data Integration

The complexity and scope of knowledge that man has gained about various aspects of the physical world has led to scientists to specialize further and further, to the point where some researchers now spend their entire careers studying the behaviour of one particular protein or a single species of plankton. This level of focus is unavoidable to some degree – no one person can be an expert in everything. However, it is sometimes easy to forget that the world around us is an interconnected system whose behaviour spans the artificial boundaries between traditional scientific disciplines, and often the greatest leaps forward in our understanding come at the intersection of different areas of study. These types of breakthroughs require the integration of data from many different scientific domains, and this integration must be done in such a way that the detail, uncertainty, and above all the context of the data are preserved.

The first challenge of scientific data integration is accessibility. Much of the data underpinning past and present scientific publications is not readily accessible – it exists only in isolated databases, as files on a grad student's computer, or in tables within PDF documents. The consequence of this is that it is often difficult to replicate published experimental results and to do new analyses on existing data. There is a monetary cost as well: if data is not stored and shared in an accessible manner then it must be collected independently by multiple researchers, using limited scientific funding that could better be spent elsewhere. Fortunately, many funding agencies have taken action on this issue and now require that data collected via funded programs be stored in official data repositories. This is a promising development, but many current data repositories do not make data integration easy. Traditional scientific data repositories are generally either relational databases or file servers containing spreadsheet, CSV, or irregularly structured text files. There can be various obstacles to retrieving this data, particularly due to a lack of consistency. For instance, some repositories might be accessible via websites or structured query mechanisms while others require a login and use of secure file transfer or copy protocols. Financial and legal concerns also inhibit data integration. Some data might be stored in proprietary database or file formats that require expensive software licenses to read, and licenses indicating what users are allowed to do with the data can be missing or restrictive, resulting in legal uncertainty.

One approach that has been proposed to address these accessibility issues is publishing data as "Linked Data." In a linked dataset, every entity is given a Uniform Resource Identifier (URI) that can be accessed via HTTP, similar to a standard web page. When this URI is dereferenced, there is structured data providing more information about the entity. This data is generally expressed using RDF, and it may include URIs to other, related, entities. Using these standards, it possible to make data available in a way that is both accessible and understandable.

An example to clarify this description seems warranted. Assume that Dr. Jane Doe, a scientist at State University, wants to publish a linked dataset containing information about the papers she has written, one of which is called "An Exploration of the Feasibility of Tenure." One way for Dr. Doe to do this is

to acquire ownership of a domain name and assign URIs in that namespace to each of the entities in her dataset. For instance, if the domain name is profdoe.edu, then she might use the URI profdoe.edu/Jane-Doe to represent herself and profdoe.edu/TenureFeasibilityExploration to represent the paper. Dr. Doe could then create files containing RDF statements about these entities and deploy them on a webserver. An RDF statement is a subject-predicate-object triple. For example, the following triple states that the paper's title is "An Exploration of the Feasibility of Tenure":

```
<profdoe.edu/TenureFeasibilityExploration>
<profdoe.edu/hasTitle>
”An Exploration of the Feasibility of Tenure”@en .
Similarly, this statement expresses that the paper was written by Dr. Doe:
<profdoe.edu/TenureFeasibilityExploration>
<profdoe.edu/hasAuthor>
<profdoe/JaneDoe> .
```

Using the linked data standards eliminates many of the problems described above, because the data is in a consistent format and accessible in a uniform manner rather than, for example, in proprietary databases behind firewalls. Problems related to licensing can be handled by linking to an appropriate license that has been encoded as linked data triples. For instance, the RDF triple below indicates that Professor Doe's linked dataset containing her publications (the dataset, not necessarily the publications themselves) is covered by version 3.0 of the creative commons "ShareAlike" license. Licensing information is important for those wishing to make use of the information available in the linked data cloud.

```
<profdoe.edu/publications.rdf>
cc:license
<http://creativecommons.org/licenses/by-sa/3.0/> .
```

Accessibility is only one requirement for semantic data integration, however. For data to be truly useful, scientists need to be able to interpret and use it after they acquire it. Doing this requires semantic context. In relational databases and spreadsheets, this is sometimes lacking because important information about what the various data fields mean and how they relate to one another is often implicit in the names of database tables and column headers, some of which are incomprehensible to anyone other than the dataset's creator. This problem is reduced to some degree if the linked data standards are used, because some of the meaning of these column names can be explicitly defined. For instance, the following RDF triples state that the "has Author" construct in our example relates a paper to a person and if Paper X "has Author" Person Y, then Person Y "wrote" Paper X.

```
<www.profdoe.edu/hasAuthor>
<rdfs:domain> <www.profdoe.edu/Paper>
<www.profdoe.edu/hasAuthor>
<rdfs:range> <www.profdoe.edu/Person>
<www.profdoe.edu/hasAuthor>
<owl:inverseOf> <www.profdoe.edu/wrote>
```

Using all of these RDF statements, a piece of software called a reasoner would be able to infer some semantic context for the data, such as that www.profdoe.edu/JaneDoe is a person and that she wrote the paper represented by www.profdoe.edu/TenureFeasibilityExploration, without the need for any natural language processing. These constraints enable a data provider to make the meaning of field names and relationships more precise, which in turn facilitates data integration, but there is still room for ambiguity. This potential for misunderstandings decreases as the links between datasets become more numerous, thereby further constraining potential interpretations of entities. For example, it is possible to express that the Jane Doe referred to using the URI www.profdoe.edu/JaneDoe is the same one that DBPedia, the linked data version of Wikipedia, refers to as dbpedia.org/page/Jane_Doe.

```
<www.profdoe.edu/JaneDoe>
<rdfs:seeAlso> <dbpedia.org/page/Jane_Doe>
```

Such links can be made at the schema level as well as the data (or instance) level. For instance, we can state that "hasAuthor" in this dataset is similar to "creator" in the Dublin Core terminology.

```
<www.profdoe.edu/hasAuthor>
<rdfs:seeAlso> <purl.org/dc/terms/creator>
```

While these examples are fictitious, Listing 1 contains a fully worked example RDF document describing a dataset published by the British Oceanographic Data Centre, and linked to many remote vocabulary resources. Leadbetter (2015) describes in detail the publication process for these RDF documents. Establishing these links can be very difficult, particularly if the datasets are large and complex, which is routinely the case in scientific domains. The fields of ontology alignment and co-reference resolution seek to develop tools and techniques to facilitate the identification of links between datasets (Euzenat & Shvaiko, 2007), however scientific datasets are particularly challenging for several reasons. Perhaps most obviously, such datasets can be extremely large. Consider climate data collected on a 1° grid – this results in over a petabyte of data, more than enough to swamp most existing data integration techniques. Additionally, scientific datasets generally have a spatiotemporal aspect, but current alignment algorithms struggle with finding relationships across this type of data because of the variety of ways to express it. For example, spatial regions can be represented by geopolitical entities (whose borders change over time), by the names of nearby points of interest, or by polygons whose points are given via latitude and longitude. Similarly, issues pertaining to measurement resolution, time zones, the international dateline, etc. can confuse the comparison of timestamps of data observations. Furthermore, scientific datasets frequently involve data of very different modalities, from audio recordings of dolphin calls to radar images of storms to spectroscopy of cellular organisms. Such data is also obviously collected at widely differing scales, from micrometres to kilometres. And oftentimes the data that needs to be integrated is from domains with only a small degree of semantic overlap, as is the case with, for example, one dataset containing information about NSF project awards and another with the salinity values for ocean water collected during oceanographic cruises, several of which were funded by NSF.

At this point we can clearly see that successfully integrating scientific datasets holds the promise of major advances in our knowledge of the world around us, but achieving this goal is likely to be exceedingly challenging. What is the best way to start? One proposal that is gaining traction is to focus on the key concepts that recur frequently across many subdomains. This idea of focusing on the few similari-

Listing 1. A fully worked example RDF document, as published by the British Oceanographic Data Centre within their Published Data Library. The RDF document describes the dataset, its authors, and how the dataset should be cited within the scientific literature.

```
@prefix dc1: <http://purl.org/dc/elements/1.1/> .
@prefix ns1: <http://www.opengis.net/> .

<https://www.bodc.ac.uk/data/published_data_library/catalogue/10.5285/0c9b8ad1-
ba02-0c0e-e053-6c86abc03d26/> dc1:Period "start=1915-01-01; end=2014-12-31;
scheme=W3C-DTF"@en;
    dc1:bibliographicCitation """
                Haigh I.D.; Wadey M.P.; Gallop S.L.; Loehr H.; Nicholls R.J.;
Horsburgh K.J.; Brown J.; Bradshaw E. (2015). A database of 100 years (1915-
2014) of coastal flooding in the UK. British Oceanographic Data Centre - Natu-
ral Environment Research Council, UK. doi:10/zcm."""@en;
    dc1:contributor "Bradshaw E. "@en,
        "Brown J. "@en,
        "Gallop S.L. "@en,
        "Horsburgh K.J. "@en,
        "Loehr H. "@en,
        "Nicholls R.J. "@en,
        "Wadey M.P. "@en;
    dc1:coverage <http://vocab.nerc.ac.uk/collection/C19/current/1/>,
        <http://vocab.nerc.ac.uk/collection/C19/current/1_2/>,
        <http://vocab.nerc.ac.uk/collection/C19/current/1_4/>,
        <http://vocab.nerc.ac.uk/collection/C19/current/1_5/>,
        <http://vocab.nerc.ac.uk/collection/C19/current/1_7/>;
    dc1:creator "Haigh I.D. "@en;
    dc1:description "This database, and the accompanying website called 'Surge-
Watch' (http://surgewatch.stg.rlp.io), provides a systematic UK-wide record of
high sea level and coastal flood events over the last 100 years (1915-2014).
Derived using records from the National Tide Gauge Network, a dataset of ex-
ceedence probabilities from the Environment Agency and meteorological fields
from the 20th Century Reanalysis, the database captures information of 96
storm events that generated the highest sea levels around the UK since 1915.
For each event, the database contains information about: (1) the storm that
generated that event; (2) the sea levels recorded around the UK during the
event; and (3) the occurrence and severity of coastal flooding as consequence
of the event. The data are presented to be easily assessable and understand-
able to a wide range of interested parties.  The database contains 100 files;
four CSV files and 96 PDF files. Two CSV files contain the meteorological and
sea level data for each of the 96 events. A third file contains the list of
the top 20 largest skew surges at each of the 40 study tide gauge site. In
the file containing the sea level and skew surge data, the tide gauge sites
```

continued on following page

Listing 1. Continued

```
are numbered 1 to 40. A fourth accompanying CSV file lists, for reference, the
site name and location (longitude and latitude). A description of the param-
eters in each of the four CSV files is given in the table below. There are
also 96 separate PDF files containing the event commentaries. For each event
these contain a concise narrative of the meteorological and sea level condi-
tions experienced during the event, and a succinct description of the evidence
available in support of coastal flooding, with a brief account of the recorded
consequences to people and property. In addition, these contain graphical rep-
resentation of the storm track and mean sea level pressure and wind fields at
the time of maximum high water, the return period and skew surge magnitudes at
sites around the UK, and a table of the date and time, offset return period,
water level, predicted tide and skew surge for each site where the 1 in 5 year
threshold was reached or exceeded for each event.  A detailed description of
how the database was created is given in Haigh et al. (2015).  Coastal flooding
caused by extreme sea levels can be devastating, with long-lasting and diverse
consequences. The UK has a long history of severe coastal flooding. The recent
2013-14 winter in particular, produced a sequence of some of the worst coastal
flooding the UK has experienced in the last 100 years. At present 2.5 mil-
lion properties and £150 billion of assets are potentially exposed to coastal
flooding. Yet despite these concerns, there is no formal, national framework
in the UK to record flood severity and consequences and thus benefit an under-
standing of coastal flooding mechanisms and consequences. Without a systematic
record of flood events, assessment of coastal flooding around the UK coast
is limited.   The database was created at the School of Ocean and Earth Sci-
ence, National Oceanography Centre, University of Southampton with help from
the Faculty of Engineering and the Environment, University of Southampton, the
National Oceanography Centre and the British Oceanographic Data Centre. Col-
lation of the database and the development of the website was funded through
a Natural Environment Research Council (NERC) impact acceleration grant. The
database contributes to the objectives of UK Engineering and Physical Sciences
Research Council (EPSRC) consortium project FLOOD Memory (EP/K013513/1).”@en;
    dc1:format <http://vocab.nerc.ac.uk/collection/M01/current/DEL/>,
        <http://vocab.nerc.ac.uk/collection/M01/current/DOC/>;
    dc1:identifier <http://dx.doi.org/10.5285/0c9b8ad1-ba02-0c0e-e053-
6c86abc03d26>,
                <http://dx.doi.org/10/zcm>;
    dc1:language <urn:ietf:params:language:en-GB>;
    dc1:publisher <http://www.bodc.ac.uk/>;
    dc1:relation <https://www.bodc.ac.uk/data/information_and_inventories/ed-
med/report/6120/>;
    dc1:subject <http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_TopicCategoryCode_elevation,>,  <http://www.isotc211.org/2005/resourc-
es/Codelist/gmxCodelists.xml#MD_TopicCategoryCode_oceans>;
```

continued on following page

Listing 1. Continued

```
    dc1:title "A database of 100 years (1915-2014) of coastal flooding in the
UK."@en;
    ns1:gmlid <http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_TopicCategoryCode_elevation,>,      <http://www.isotc211.org/2005/re-
sources/Codelist/gmxCodelists.xml#MD_TopicCategoryCode_oceans>;
```

ties that exist amidst the many differences inherent in the datasets is at the core of the ontology design pattern approach to data modelling and integration.

Fundamentals of ODPs

While an ODP differs in important ways from an ontology, the basic components of both are the same: classes, instances, and properties. A class represents a grouping of objects with similar characteristics. Classes are often arranged in a hierarchy using subclass relationships. For instance, Submarine may be a subclass of Vessel. An instance (or individual, the terms are often used interchangeably) is a particular object. An instance has a type that is some class within the ontology. For example, the Anorep 1 is an instance of type Submarine. This is somewhat analogous to classes and instances of those classes in object-oriented programming languages, such as Java. Relationships between instances, such as captainOf and hasName, are called properties. All properties are directed binary relations that map an instance with a type from the domain to something in the range. Properties that map an instance to another instance (e.g. captainOf, which may map an instance of type Person to an instance of type Vessel) are object properties, whereas properties that map an instance to a literal value (e.g. hasName, which maps an instance of type Person to a string value) are datatype properties. Common data types include integers, doubles, strings, and dateTime. Both object properties and data properties must involve an instance. A third type of property, called an annotation property, can be used to describe relationships between any types of entities (i.e. instances, classes or other properties). All of this information: classes, properties, and any restrictions on them, such as cardinality, disjointness, etc., are called the schema, or T-box (for terminology), of the ontology. Conversely, the instance data, or A-box (for assertions), contains assertions about individuals using data from the T-box. Both T-box and A-box statements are generally expressed using the Web Ontology Language (OWL). A more formal and extensive treatment of these topics can be found in Hitzler, Krotzsch, and Rudolph (2011).

While ODPs are made up of the same components as full ontologies they differ in important ways, the biggest of which is that an ODP focuses on only one generic notion. The OWL axiomatization of the ODP is carefully formulated such that no overly strong (i.e., application-specific) ontological commitment is made by the pattern, in the same way that an architectural pattern for a vestibule avoids making constraints on the rest of the building. In comparison to a monolithic upper ontology, an ODP can be seen as a snippet that defines only one particular notion without the excessive semantic constraints that an upper ontology may entail. An analogy may be helpful here. An ODP differs from an ontology in the same way that a paragraph is differs from a project report. The key difference is that a good paragraph

(and a good ODP) very tightly describes a single idea. You likely can re-use a single paragraph you write about some aspect of your work in many different project reports, but it's unlikely that you can re-use a full project report in many different situations, simply because the complete report says too much to be fully reusable in different situations. This is true even if the project report is extremely well written, by people who are established experts in their field. However, if one writes a project report on the same general subject as a different report that she wrote previously, she may very well cite that other report in the current one, referring to it when it makes sense. This is the relationship between data integrated by stitching together ODPs, and existing full domain ontologies – it is generally not possible to use the full domain ontologies directly, but one can, and should, refer to them by creating the appropriate links.

Figure 4 shows the core of an ODP for representing an oceanographic cruise, which is defined as an expedition undertaken by a vessel in the ocean or other navigable water body in order to conduct oceanographic research activities. Cruises hold a critical role in the ocean sciences because most field observations, data acquisition, and scientific experiments can only be accomplished when researchers are on a scientific expedition. From a data integration perspective, the notion of a cruise is important because it acts as a type of "glue" connecting all of the data and research results resulting from activities carried out during an oceanographic study.

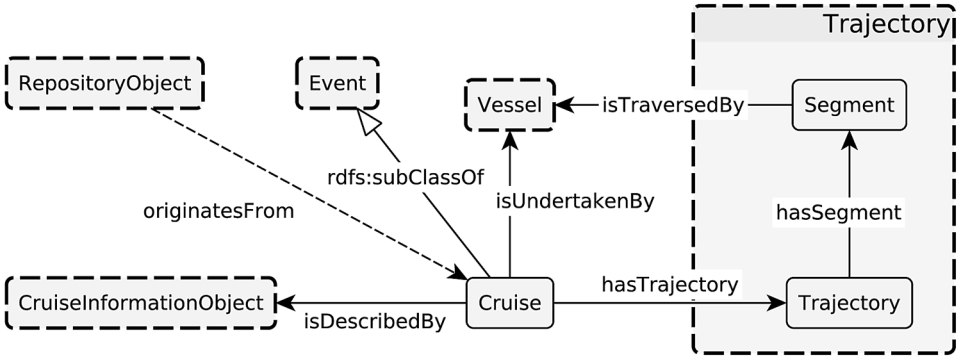
ODPs can leverage one another through axioms that formalize horizontal links between them. This allows applications to formally reason over the data and draw new inferences. For example, a key component of a cruise is the path the vessel takes on its voyage. An ODP to represent a trajectory already existed (Hu, *et al.* 2013). When the cruise ODP was created, axioms to connect the relevant concepts within that pattern to the aforementioned trajectory pattern were specified. For example, axiom 1 shows that a cruise has exactly one trajectory and is undertaken by exactly one vessel.

$$\text{Cruise} \sqsubseteq (=1\ \text{hasTrajectory.Trajectory}) \sqcap (=1\ \text{isUndertakenBy.Vessel}) \tag{1}$$

ODPs can also extend one another in order to represent different levels of abstraction. For instance, an oceanographic cruise is a specialization of a more general event pattern. This is shown intuitively in Figure 5 and specified formally in axioms 2 through 6a,b.

$$\text{Cruise} \sqsubseteq \text{Event} \tag{2}$$

Figure 4. An oceanographic cruise ODP



$$\text{CruiseRoleType} \sqsubseteq \text{RoleType} \tag{3}$$

$$\text{CruiseRoleType}(x) \text{ for every role type } x \text{ defined for the pattern} \tag{4}$$

$$\text{RCruise} \circ \text{owl:topObjectProperty} \circ \text{RCruiseRoleType} \sqsubseteq \text{providesRoleType} \tag{5}$$

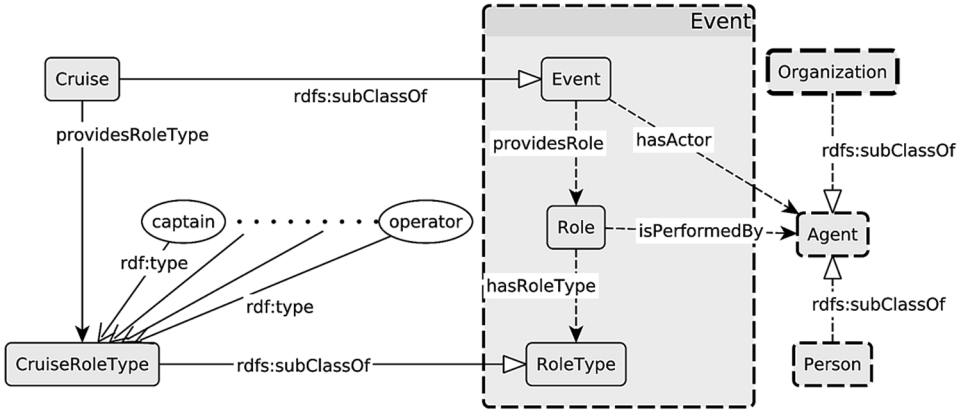
$$\text{Cruise} \equiv \exists \text{RCruise.Self}, \text{CruiseRoleType} \equiv \exists \text{RCruiseRoleType.Self} \tag{6a,b}$$

ODPs are useful for more than just data integration. Another important benefit is their usability by domain scientists unfamiliar with knowledge representation and reasoning technologies. Because ODPs represent key concepts in a domain that recur frequently across many datasets, they can serve as semantic "anchors" that allow someone looking at a dataset for the first time to get conceptually oriented. The ability of ODPs to specialize other patterns in order to represent data at a variety of abstraction levels also allows individuals with different levels of understanding of the domain, such as students, to work with the data at a level that matches their expertise.

GeoLink: ODPs in Action

As discussed previously, understanding past and present data related to the world around us to the point where it is possible to actually predict aspects of its future state is extremely challenging. The earth is a complex and interconnected system that no one scientist, research group, or even field of study can hope to understand. Instead, geoscientists of all different stripes must work together to make progress: geologists, meteorologists, climatologists, ecologists, archaeologists, and so on. The National Science Foundation (NSF) has recognized this need, and in 2011 it launched the EarthCube initiative. The goal of this effort is to galvanize a community-driven approach to collaboration across traditional geoscience domains. Rather than impose standards and infrastructure from the top down, the NSF is providing

Figure 5. The cruise ODP as a specialization of the event ODP



funding to various interdisciplinary teams to explore alternatives, identify the best ones, and disseminate those findings across the community. While this is obviously not something that can be done quickly or easily, the NSF has committed to seeing the effort through until at least 2022.

One of the projects funded under the EarthCube initiative is "GeoLink – Leveraging Semantics and Linked Data for Data Sharing and Discovery in the Geosciences". The goal of GeoLink is to lower barriers to cross-repository data discovery and access, while respecting and preserving repository autonomy and heterogeneity, by applying ODPs to link the repositories' holdings. Spanning ocean, earth and polar sciences, GeoLink's data providers are Rolling Deck to Repository (R2R), the Biological and Chemical Oceanographic Data Management Office (BCO-DMO), Integrated Earth Data Applications (IEDA), the Long-Term Ecological Research Network (LTER), DataONE, and the International Ocean Discovery Program (IODP). If a researcher could query across all of these repositories, tasks such as the following become feasible:

- "Find abundance data of *Euphausia pacifica* sampled with the MOCNESS instrument."
- "Find all NSF awards and publications related to the location called 'Station Papa'."
- "Find data of all measurements of salinity taken at 'Station Papa' over the last 40 years."
- "Find people who collected biodiversity data along the Gulf of Mexico."

These tasks may seem simple on first glance, but consider a couple of features of the queries necessary to accomplish them. First, each of these queries involves multiple facets of the data, including measurements, instruments, scientists, research themes, and time and place. Put in the language of the database community, there are many 'joins' involved, and these joins are over very different data types. Second, while many of these facets appear in multiple datasets, their role, prominence, and meaning within those datasets varies greatly. This is because each repository has a unique point of view on the domain. For instance, some repositories may see a notion such as 'instrument type' as central to their holdings, some as data that is potentially interesting but not required, and others may see instrument type as a detail that is unimportant to the community they serve. Performing queries involving facets that are ascribed such different roles in different datasets is a difficult challenge, and the main goal of the GeoLink effort.

As a sample case study, let us consider a vessel operator that wants to determine their organization's impact on research related to a particular type of phytoplankton. In this case, the organization may want to issue a query such as "Find me all fluorescence data collected during research cruises on vessels operated by Organization A." Such a query involves data stored in at least two repositories: R2R and BCO-DMO. R2R's data schema revolves around the concept of a research cruise – what vessel sailed, the track of the cruise, who provided funding for the cruise, etc. Information about the data the vessel's instrumentation collected is included in the R2R repository, but it is not the central focus. Conversely, BCO-DMO is centred on the data gathered by researchers and the instruments used to collect it. Because R2R gets its data from the vessel's operating organization, when a group of researchers brings their own instrumentation aboard a vessel, it does not always show up in the R2R repository. The result is that, while the same cruises are typically represented in both R2R and BCO-DMO, these are two separate representations each with a different, but complementary, focus.

Completing the task described in this case study requires two subtasks: (1) finding all of the research cruises operated by Organization A (using the R2R repository) and all the cruises in BCO-DMO that involved the collection of data related to fluorescence and then (2) finding the intersection of these two

sets. Prior to GeoLink, this was surprisingly difficult. One challenge is that a 'cruise' in R2R is not called a 'cruise' in BCO-DMO; rather it is a 'deployment' with a 'platform' of type 'vessel'. A deployment is defined as the activity of data collection made possible by some platform, such as a vessel, laboratory, aircraft, or satellite. The platform is the base from which instruments are deployed to collect data. The implication is that we must query R2R and BCO-DMO differently for the same concepts, due to slight variations in their vocabularies. Once we have the cruises from R2R involving Organization A and the deployments of type vessel with datasets involving fluorescence from BCO-DMO, we need to find the overlap between these two sets. This leads us to our second challenge: determining when a particular Cruise X in R2R was the same as Cruise Y in BCO-DMO. Prior to GeoLink, this was done via ad-hoc scripts and manual input from the owners of both ontologies.

This use case illustrates that even relatively simple queries require an alignment of both the vocabularies and the instances in the relevant repositories. Arriving at the realization that a cruise in R2R is equivalent to a deployment on a platform of type vessel in BCO-DMO required the expertise and in-depth collaboration of the repository custodians. The same is true for error-free confirmation that Cruise X in R2R is the same as Cruise Y in BCO-DMO. This is not scalable, particularly when queries require data in more than two repositories.

GeoLink address this issue through the use of ontology design patterns as a mediating semantic layer focused on the areas of conceptual overlap between ontologies in the domain. Once the vocabularies of individual ontologies are aligned to these patterns (subtask 1 from the use case), automated co-reference resolution becomes much more accurate (subtask 2), due to the larger volume of data available about each item. For instance, if we know from R2R that a person with the name Michael Smith served as Chief Scientist on Cruise X which was funded by Organization Y, and from BCO-DMO that Cruise X studied ocean salinity, and from a third repository that a paper with "ocean salinity" as a keyword that acknowledges Organization Y as a funding source was written by M. Smith, it is highly likely that this author is the same Michael Smith mentioned in R2R. In this way, the GeoLink project is showing that ontology design patterns facilitate the alignment of vocabularies and instance data across repositories, even when the repositories' views of the domain vary widely.

One potential criticism of GeoLink is that the ODP-based approach "reinvents the wheel" rather than leveraging the many carefully developed domain ontologies that are relevant, such as OBOE (for ecological observational data) (Madin, *et al.* 2007), OBO Foundry (a collection of open biomedical ontologies) (Smith, *et al.* 2007), O&M (for observations and measurements) (Probst, 2006), and ENVO (for environments) (Buttigieg, *et al.* 2013). The GeoLink approach is not meant to imply a negative view on the utility or correctness of those ontologies, but rather to merely acknowledge that wide-ranging domain ontologies, while useful in many situations, are quite difficult to use when integrating large and heterogeneous datasets such as those found in the geosciences. For instance in the OBOE ontology, if an observation occurred within a certain context (i.e. environment), then the entity that was observed is considered to also be in that context. This might not be a safe assumption for some datasets.

Of course, many individuals and applications would greatly benefit from being able to query the GeoLink data using terms found in established domain ontologies. For instance, existing applications often automatically query for instances of foaf:Person when looking for the individuals described in a dataset. GeoLink can respond appropriately to such queries by establishing the relevant links between its internal schema entities, such as gl:Person, and related concepts in external ontologies like foaf:Person. Care must be taken when encoding these relationships, however.

One of the powerful aspects of linked data and ontologies is the ability to make inferences based on the available facts. For instance, assume that the knowledge base contains the fact that any Chief Scientist must have a university degree in a scientific discipline, and that Jane Doe was the Chief Scientist on Cruise123. Even if there is no information about Jane Doe's degree in the knowledge base, it can be inferred that she must have one. Therefore a query to the knowledge base for any people with a scientific degree would include Jane Doe among the results. Inferencing can also help to recognize when data has been entered incorrectly. One of the axioms in the Cruise ODP is that a cruise involves exactly one vessel. If a single cruise has two vessels associated with it, those vessels must either actually be the same ship or there is a mistake in the data.

Inferences are made through a piece of software called a reasoner, and they represent one of the biggest benefits of the semantic web – a knowledge base that you can't reason over is essentially just an "RDFized" database, to which the word "semantic" cannot be reasonably applied. However, if the relationships between the GeoLink design patterns and domain ontologies are specified using strong logical commitments such as owl:equivalentClass or rdf:subClassOf, then reasoners may behave in undesired ways when they are applied to the dataset. To see this, consider that a Program in GeoLink can be considered similar to a Project in the project management ontology PROMONT (Abels, *et al.* 2007). Furthermore, the GeoLink entities Person and PrincipleInvestigatorRole are similar to PROMONT's Employee and Task. One way to represent these relationships is through the following triples:

<gl:Program> <owl:equivalentClass> <PROMONT:Project> .
<gl:Person> <owl:equivalentClass> <PROMONT:Employee> .
<gl:PrincipleInvestigatorRole> <owl:equivalentClass> <PROMONT:Task> .

Assume that Jane Doe is the Principal Investigator on both Program123 and Program456. The start and end dates of these programs are such that there is some overlap. Everything may work fine at first, but PROMONT is designed for precise project management applications. What if the developers of PROMONT decide to add an axiom to their ontology indicating that it is not possible for an employee to be working on two tasks simultaneously? An action beyond GeoLink's control has now caused a logical inconsistency in the knowledge base, and a reasoner will fail. In order to avoid problems such as this, GeoLink plans to establish links to domain ontologies using the gl:matches predicate, as shown in the example below.

<gl:Program> <gl:matches> <PROMONT:Project> .

The gl:matches entity is defined very broadly to mean that two entities linked with this predicate are "sufficiently similar that they can be used interchangeably in some information retrieval applications." Note that this is the same definition as the one for skos:closeMatch (Alexander, 1977). The difference is that, unlike skos:closeMatch, the gl:matches property is not defined to be associative.

With this setup, users and applications can query GeoLink for instances of PROMONT:Project by finding all instances of any GeoLink class that gl:matches PROMONT:Project. This level of indirection makes it clear to the consumer of the data that it may not be appropriate to use the GeoLink instances in all situations in which a PROMONT:Project is expected. This approach allows GeoLink to be queried using terminology familiar to domain experts while still supporting reasoning.

Unlike traditional mission-driven science funding initiatives, the National Science Foundation funds hypothesis-driven research. This difference produces research quality data with varying types of heterogeneity in size and scope. Because of this variability, integrating datasets across NSF data repositories is a difficult task to achieve and scale. For the data repositories participating in the GeoLink project, Ontology Design Patterns, as a mediating semantic layer, have helped to integrate these heterogeneous data holding for producing new information and knowledge. Now that this is possible, with improved accuracy of the asserted shared connections and intersection points, this new knowledge can be put in the hands of researchers and policy makers affecting decisions related to the Earth's ecosystems.

STREAMING LINKED DATA

It has now been seen that Linked Ocean Data can be created within design patterns which give a structure to data made available online. Traditionally, oceanographic data has been published in a *post-facto* delayed mode which improves the base archives used to develop climatologies, extend time series, and contribute to studies of climate change, among other activities (JCOMM, 2009). However, there has been a more recent push to deliver marine science data in real-time (Fredericks, 2015). Operational data aggregation and assembly from distributed data sources will be essential to the ability to adequately describe and predict the physical, chemical and biological state of the ocean. These activities demand a trustworthy and consistent quality description for every observation distributed as part of the global ocean observing system.

At the same time, the "Big Data" paradigm has been increasingly prevalent in the global consciousness. Big Data recognises that data are being generated at an increased volume, at a greater speed, covering a greater range of parameters than ever before. Activities such as the International Argo Programme and the Everyone's Gliding Observatories have added to the instrumentation of the ocean which has vastly increased the volume of data being captured, and the platforms deployed by these programmes often report in real-time increasing the data velocity. Novel sensors, such as the biogeochemical sensors developed within the SenseOCEAN project, are increasing the variety of the data these platforms collect meaning the marine science community is beginning to move into a Big Ocean Data paradigm. This paradigm allows the use of emerging software architecture models for the streaming of data to be leveraged within the ocean sciences domain. Big Data is not solely linked to the volume of data being processed, but also has characteristics of: "velocity", i.e. Big Data is often available in real-time; and "variety", i.e. Big Data is often complex with a number of dimensions (De Mauro, Greco, and Grimaldi 2015). Other characteristics which are often applied to the concept of Big Data are its "veracity" which indicates that the quality of the source data must be considered and is vital to its effective processing; and the "complexity" of the data as the sources may be many and varied such that the data must be linked, connected, and correlated so users can grasp the information the data is supposed to convey (Hilbert 2013). Bearing in mind both the "velocity" and the "complexity" aspects of Big Data, it can be seen that there is value in the exploration of streaming Linked Data from oceanographic instrumentation.

The Header-Dictionary-Triples (RDF-HDT) data structure and serialization format offers a compact structure for storing triples and a binary serialization of the RDF model which allows RDF datasets to be compressed while maintaining search and browse capabilities (Fernández, Martínez-Prieto, Gutiérrez, Polleres, & Arias 2013). Initially, this would seem to be the approach to follow for streaming Linked Ocean Data from instruments to data centres and then to the World Wide Web. However, the experiences

of the SenseOCEAN project have shown that there is a practical issue with attempting to deploy any extra software layers, such as the RDF-HDT layer, on to instruments to be deployed remotely in the marine environment as the bandwidth between sensor, logger and transmitter is highly limited - particularly between sensor and logger. Also the deployment of extra software for processing on autonomous platforms, such as Argo floats and Autonomous Underwater Vehicles which are where much of the extra volume of observational data in the marine sciences are collected, significantly increases the battery usage on those platforms and therefore reduces the length of time for which they can be deployed. In order to build Linked Ocean Data into a Big Ocean Data paradigm, an alternative solution must therefore be sought.

As the demand for data to be delivered in real-time from a range of internet applications, new architectures for software processing such as the *lambda*- and *kappa*-architectures, have been developed. In *lambda*-architecture, an immutable sequence of records is captured and fed into a batch system and a stream processing system in parallel (Figure 6). The transformation logic is, however, implemented twice, once in the batch system and once in the stream processing system. Results from both systems are stitched together at query time to produce a complete answer (Marz and Warren, 2015). However, particularly in ocean science scenarios, the desired workflow is to process the data in some rapid manner as close to the time of collection as possible using the *a priori* knowledge of the dataset, and then re-process once the *post hoc* knowledge base is increased. This fits well within the basis of the *kappa*-architecture, which was proposed by Kreps (2014) as an alternative to the *lambda*-architecture (Figure 7). The lambda architecture concentrates on the ability to reprocess the full data stream at a later date through storing the full data message queue in a message queue which allows for multiple subscribers (for example, Apache Kafka). When reprocessing is required, new processing job code is introduced to the stream processing system which can be run against the entire message queue to generate an *n+1* version of the output.

Figure 6. The lambda-architecture. Example stream processing systems in the second stage include the Apache projects Hadoop, Samza and Storm.

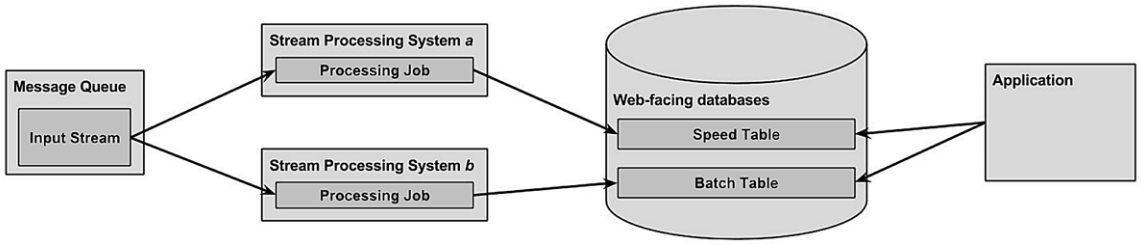
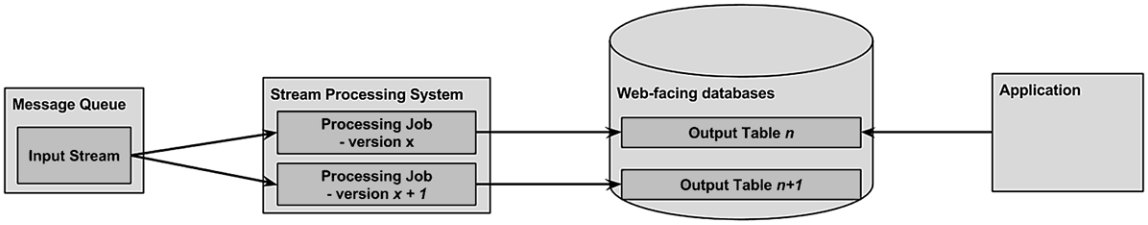
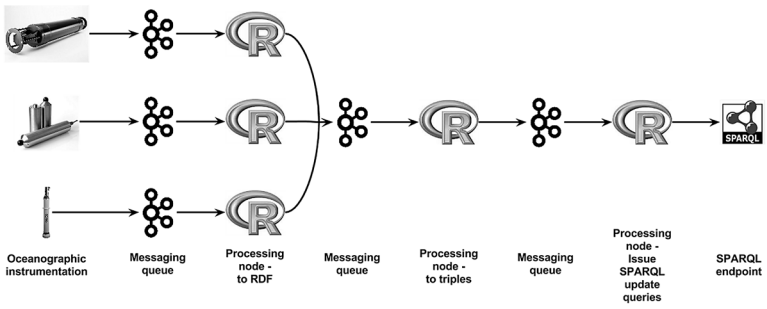


Figure 7. The kappa-architecture



In terms of streaming Linked Data within an architecture like this, a possible approach is outlined below. The initial message received from the instrumentation is placed into a message queue, a processing logic node converts this into a full RDF document and posts this to a new message queue. The RDF document is read from its message queue and parsed into its component triples, each of which are added to the next message queue. A final processing node reads this message queue and passes the triples to a SPARQL endpoint using an UPDATE command. This combination of messaging queues and nodes containing processing logic which subscribe to these queues is known as a "topology". Specifically, a topology is a graph of computation in which the nodes represent individual functions enacted on data and the edges denote the pathways by which data may move between nodes. A demonstration topology has been created using Apache Kafka as the messaging queue and the R statistical programming language as the stream processing system. R has been demonstrated as an appropriate platform for processing Semantic Web data streams by Willighagen (2014), and the support for interfacing with Kafka from R is very good, using a simple Application Programming Interface to push messages to and read messages from a queue. The topology is outlined in Figure 8 and detailed in Listing 2.

Figure 8. An example Big Ocean Data topology for streaming Linked Ocean Data



Listing 2. An R statistical programming language encoding of the topology in Figure 8

```
library(rkafka)
library(elastic)
library(rrdf)
library(jsonlite)
message2kafka <- function(kafkaIP, kafkaTopic, message) {
  if (missing(kafkaIP)) {
    stop('Missing input argument: kafkaIP')
  } else if (missing(kafkaTopic)) {
    stop('Missing input argument: kafkaTopic')
  } else if (missing(message)) {
    stop('Missing input argument: message')
  } else {
    messageProducer <- rkafka.createProducer(kafkaIP)
    rkafka.send(messageProducer, kafkaTopic, kafkaIP, message)
    rkafka.closeProducer(messageProducer)
```

continued on following page

Listing 2. Continued

```
  }
}
kafka2r <-function(zookeeperIP, kafkaTopic) {
  rdfConsumer <- rkafka.createConsumer(zookeeperIP,
                                       kafkaTopic, consumerTimeoutMs = "10000")
  msg <- rkafka.read(rdfConsumer)
  rkafka.closeConsumer(rdfConsumer)
  msg
}
rdfmultiplex <- function(zookeeperIP, kafkaTopic) {
  tripleInAsJSON <- ".."
  while(tripleInAsJSON != "") {
    tripleInAsJSON  <-  kafka2r(zookeeperIP, kafkaTopic)  # Try Catch Loop needed here
    tripleIn <- fromJSON(tripleInAsJSON)
    print(tripleIn)
  }
}
rdfxmlfile2kafka <- function(fileName, kafkaIP, kafkaTopic) {
  if (file.exists(fileName)) {
    rawRdfXml  <-  readChar(fileName, file.info(fileName)$size)
    message2kafka(kafkaIP, kafkaTopic, rawRdfXml)
  } else {
    stop('Specified input file does not exist')
  }
}
rdf2triples <-function(formatRDF, zookeeperIP,
                       kafkaInTopic, kafkaIP, kafkaOutTopic) {
  rdfXmlStr <- kafka2r(zookeeperIP, kafkaInTopic)
  tripleStore <- new.rdf(ontology = FALSE)
  fromString.rdf(toString(rdfXmlStr), formatRDF, appendTo = tripleStore)
  triples <- sparql.rdf(tripleStore, "select * where {?a ?b ?c}")
  for (i in 1:nrow(triples)) {
    jsonTriples <- (paste0('{"s": "', unname(triples[i, 1]), '", "p": "',
                unname(triples[i, 2]), '", "o": "',
                unname(triples[i, 3]), '"}'))
    message2kafka(kafkaIP,kafkaOutTopic,jsonTriples)
  }
}
# read an XML file into Kafka
rdfxmlfile2kafka('~/rdf/oa_overlay.xml', 'localhost:9092', 'rawrdfxml')
# Read the RDF/XML from Kafka and create a series of RDF triples
rdf2triples('RDF/XML', 'localhost:2181', 'rawrdfxml', 'localhost:9092', 'rdftriples')
# Multiplex the RDF Triples
rdfmultiplex('localhost:2181','rdftriples')
```

The topology shown in Figure 8 takes raw oceanographic instrument output, normally as text over a serial port, as its inputs (to the far left of the image within Figure 8). A script, for example written in Python or R, listens to the port over which the instrument is streaming its outputs and these are pushed to the Kafka messaging queue. The R topology reads the raw data from the message queues for each instrument and converts the full reading to an RDF XML document which is pushed onto a combined message queue for each instrument. The R topology then reads each of these documents and breaks them down into the component triples, each of which are passed on to a new message queue. The final node of the R topology takes the triples in turn and issues SPARQL update queries to push them on to a SPARQL endpoint. This topology has been successfully tested on local desktop machines in batch mode in order to prove the validity of the concept as far as the input to the SPARQL endpoint, but it remains to be proven in a production environment outside of a batch processing mode. However, the careful preparation of the data into a standardised structure for use in data intensive science supports the concept of High Performance Data as proposed by Evans *et al.* (2015).

One such standardised structure is the Observations & Measurements (O&M) model (Cox, 2006) which is both an International Organization for Standardization (ISO) and Open Geospatial Consortium (OGC) standard which defines a conceptual schema encoding for observations, and for features involved in sampling when making observations. The OGC make specific use of O&M in their Sensor Web Enablement standards suite, where it provides the response model for the Sensor Observation Service. In this case, Sensor Web Enablement and Sensor Observation Services are of particular interest as they are intended, respectively, to make sensors discoverable, accessible and usable and to query both real-time and time-series sensor data via the Web. However, one drawback has been the reliance on the service response documents being delivered via eXstensible Markup Language (XML) documents, which are heavy-weight for use in real-time systems and are not favoured by web developers looking to build responsive user interfaces. One option is to use JavaScript Object Notation as an alternative to XML as it is a lighter-weight document meaning quicker response times from applications and easier development of interfaces. As Tim Bray noted in 2013 "… these days, if you want to interchange tuples or tables of tuples or numbers and strings, you have JSON. If you want to do nontrivial publishing automation, use XML. If you want to interchange smart bitmaps of page images, there's PDF. I personally think we're probably done with inventing low-level textual interchange formats" (Bray, 2013).

One barrier to this has been the lack of either a formal schema for JSON to allow the validation of a JSON document and an Observations & Measurements encoding for JSON. The first issue has been addressed by the emergence and coming to maturity of JSON Schema (Galiegue, Zyp, & Court, 2013) specifies a JSON-based format to define the structure of JSON data for validation, documentation, and interaction control. The second barrier is overcome by the publication of a JSON Schema for O&M (OM-JSON; Cox & Taylor, 2015). This allows for a JSON object to be created containing the full O&M data model, and to be validated using the appropriate JSON Schema. This is of particular interest in the Linked Ocean Data space, as OM-JSON specifies that such properties as: the Observed Property; the Procedure; the Feature of Interest; and the Units of Measure are all defined by HTTP URIs. Therefore OM-JSON is immediately Linked Data at one level or another, all the more so if common controlled vocabularies, such as the NERC Vocabulary Server, are used to provide the URIs to such resources (see Listing 3).

The example OM-JSON encoding shown in Listing 3 is from an experimental Sensor Observation Service instance which takes JSON feeds from instruments deployed within the marine environment, and maps these data onto the O&M data model (Irish Marine Institute, n.d.). A further step towards

Listing 3. An example of schema compliant OM-JSON created from a lightweight, experimental Sensor Observation Service.

```
{
  "foi": {
    "href": "http://example.marine.ie/feature/galwayBayCableObservatory"
  },
  "id": "http://example.marine.ie/ctd/Idronaut/3137/b457-a9c72a392d90",
  "member": [ {
    "id": "Temp",
    "observedProperty": {
      "href": "http://vocab.nerc.ac.uk/collection/P01/current/TEMPPR01/"
    },
    "procedure": {
      "href": "http://vocab.nerc.ac.uk/collection/L22/current/TOOL0861/"
    },
    "result": {
      "uom": "http://vocab.nerc.ac.uk/collection/P06/current/UPAA/",
      "value": 13.988
    },
    "resultTime": "2015-10-12T15:33:44Z",
    "type": "Measurement"
  }, {
    "id": "Sal",
    "observedProperty": {
      "href": "http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/"
    },
    "procedure": {
      "href": "http://vocab.nerc.ac.uk/collection/L22/current/TOOL0861/"
    },
    "result": {
      "uom": "http://vocab.nerc.ac.uk/collection/P06/current/UUUU/",
      "value": 34.775
    },
    "resultTime": "2015-10-12T15:33:44Z",
    "type": "Measurement"
  }, {
    "id": "Press",
    "observedProperty": {
      "href": "http://vocab.nerc.ac.uk/collection/P07/current/CFSN0330/"
    },
    "procedure": {
      "href": "http://vocab.nerc.ac.uk/collection/L22/current/TOOL0861/"
    },
```

continued on following page

Listing 3. Continued

```
    "result": {
      "uom": "http://vocab.nerc.ac.uk/collection/P06/current/UPDB/",
      "value": 27.7
    },
    "resultTime": "2015-10-12T15:33:44Z",
    "type": "Measurement"
  }, {
    "id": "SoundV",
    "observedProperty": {
      "href": "http://vocab.nerc.ac.uk/collection/P01/current/SVELCV01/"
    },
    "procedure": {
      "href": "http://vocab.nerc.ac.uk/collection/L22/current/TOOL0861/"
    },
    "result": {
      "uom": "http://vocab.nerc.ac.uk/collection/P06/current/UVAA/",
      "value": 1503.6205
    },
    "resultTime": "2015-10-12T15:33:44Z",
    "type": "Measurement"
  }, {
    "id": "Press",
    "observedProperty": {
      "href": "http://vocab.nerc.ac.uk/collection/P01/current/CNDCST01/"
    },
    "procedure": {
      "href": "http://vocab.nerc.ac.uk/collection/L22/current/TOOL0861/"
    },
    "result": {
      "uom": "http://vocab.nerc.ac.uk/collection/P06/current/UECA/",
      "value": 41.694
    },
    "resultTime": "2015-10-12T15:33:44Z",
    "type": "Measurement"
  } ],
  "phenomenonTime": {
    "instant": "2015-10-12T15:33:44Z"
  }
}
```

making such an OM-JSON file 5-star Linked Data would be to overlay JSON-LD patterns on the OM-JSON. JSON-LD is a method of transporting Linked Data using JSON. JSON-LD is designed around the concept of a "context" to provide additional mappings from JSON to an RDF model. The context links object properties in a JSON document to concepts in an ontology. In order to map the JSON-LD syntax to RDF, JSON-LD allows values to be coerced to a specified type or to be tagged with a language. A context can be embedded directly in a JSON-LD document or put into a separate file and referenced from different documents (from traditional JSON documents via an HTTP Link header). As an ontology for O&M already exists in OWL (Cox, 2013) it is possible to do this for OM-JSON by inserting URIs from the O&M ontology into the OM-JSON context. One strategy for creating an OM-JSON context is given in Listing 4. In order to show the knowledge encapsulated by combing Listing 3 and Listing 4, Listing 5 shows the RDF graph inferred through overlaying the JSON-LD context of Listing 4 on to the OM-JSON of Listing 3.

FUTURE RESEARCH DIRECTIONS

The solutions outlined above to "Born Semantic" data are as yet only prototypes and have neither been used in a production environment nor, in keeping with the Big Data paradigm, have they been shown to scale. The next steps in their development should be to remove the reliance on an intermediate platform

Listing 4. A possible strategy for creating a truly Linked Data Sensor Observation Service involves the output in Listing 3 overlain with a JSON-LD Context document of the type shown here.

```
"@context": {
    "observedProperty": "http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#observedProperty",
    "href": "@id",
    "foi": "http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#featureOfInterest",
    "id": "@id",
    "phenomenonTime": "http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#phenomenonTime",
    "instant": "@value",
    "member": "http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#
relatedObservation",
    "procedure": "http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#Process",
    "resultTime": "http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#resultTime",
    "result": "http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#result",
    "value": "@value"
  }
```

Listing 5. The RDF graph inferred by overlaying the OM-JSON of Listing 3 with the JSON-LD context of Listing 4.

```
<http://example.marine.ie/ctd/Idronaut/3137/b457-a9c72a392d90> <http://def.
seegrid.csiro.au/isotc211/iso19156/2011/observation#featureOfInterest> <http://
example.marine.ie/feature/galwayBayCableObservatory> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#phenomenonTime> "2015-10-12T15:33:44Z" ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#relatedObservation> < http://example.marine.ie/ctd/Idronaut/3137/
b457-a9c72a392d90/Press>, < http://example.marine.ie/ctd/Idronaut/3137/
b457-a9c72a392d90/Sal>, < http://example.marine.ie/ctd/Idronaut/3137/b457-
a9c72a392d90/SoundV>, < http://example.marine.ie/ctd/Idronaut/3137/b457-
a9c72a392d90/Temp> .

< http://example.marine.ie/ctd/Idronaut/3137/b457-a9c72a392d90/Press> <http://
def.seegrid.csiro.au/isotc211/iso19156/2011/observation#Process> <http://vo-
cab.nerc.ac.uk/collection/L22/current/TOOL0861/> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#observedProperty> <http://vocab.nerc.ac.uk/collection/P01/current/
CNDCST01/>, <http://vocab.nerc.ac.uk/collection/P07/current/CFSN0330/> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#result> "2.77E1"^^<http://www.w3.org/2001/XMLSchema#double>,
"4.1694E1"^^<http://www.w3.org/2001/XMLSchema#double> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#resultTime> "2015-10-12T15:33:44Z" .

< http://example.marine.ie/ctd/Idronaut/3137/b457-a9c72a392d90/Sal> <http://
def.seegrid.csiro.au/isotc211/iso19156/2011/observation#Process> <http://vo-
cab.nerc.ac.uk/collection/L22/current/TOOL0861/> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#observedProperty> <http://vocab.nerc.ac.uk/collection/P01/current/
PSALCU01/> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#result> "3.4775E1"^^<http://www.w3.org/2001/XMLSchema#double> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#resultTime> "2015-10-12T15:33:44Z" .

< http://example.marine.ie/ctd/Idronaut/3137/b457-a9c72a392d90/SoundV> <http://
def.seegrid.csiro.au/isotc211/iso19156/2011/observation#Process> <http://vo-
cab.nerc.ac.uk/collection/L22/current/TOOL0861/> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#observedProperty> <http://vocab.nerc.ac.uk/collection/P01/current/
SVELCV01/> ;
```

continued on following page

Listing 5. Continued

```
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#result> "1.5036205E3"^^<http://www.w3.org/2001/XMLSchema#double> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#resultTime> "2015-10-12T15:33:44Z" .

< http://example.marine.ie/ctd/Idronaut/3137/b457-a9c72a392d90/Temp> <http://
def.seegrid.csiro.au/isotc211/iso19156/2011/observation#Process> <http://vo-
cab.nerc.ac.uk/collection/L22/current/TOOL0861/> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#observedProperty> <http://vocab.nerc.ac.uk/collection/P01/current/
TEMPPR01/> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#result> "1.3988E1"^^<http://www.w3.org/2001/XMLSchema#double> ;
        <http://def.seegrid.csiro.au/isotc211/iso19156/2011/
observation#resultTime> "2015-10-12T15:33:44Z" .
```

such as R for the encoding of the nodes of the topology, and the code should be moved out to either a more formal programming language (such as Java or Go) or to a dedicated stream processing platform such either Spark or Storm from the Apache Software Foundation.

Similarly, the High Performance Datasets described above are of low-volume, but only of a medium complexity. As the creation of terabyte scale ocean datasets becomes much more prevalent, the question of converting 'Big Data' sets that comprise thousands of individual heterogeneous files (e.g., bathymetry data sets) into 'High Performance Data' (HPD) sets that can be accessed in High Performance Computing environments becomes much more relevant. This is a known problem where both climate and oceans modellers want access to national scale, calibrated higher resolution bathymetry data sets, as reported by the Australian National University to the September 2015 Ocean Data Interoperability Platform workshop. A related Big Data issue relevant to sustainable development of the coastline is the merging the high resolution LiDAR data sets (in LAS formats) with shallow water bathymetry (in CARIS, ASCII, ESRI Grid, or possibly a well-managed netCDF flavour) to create high resolution coastal elevation data sets for accurate tsunami and storm surge modelling. These remain open research issues.

Further, the extension of Observations and Measurements through application schema is a well-known through domain specialization (Cox, 2013). However, in the past this has mainly been through an informal use of common feature-type catalogues; sensor registers; parameter dictionaries; and result formats for a given domain. As shown by Diviacco and Leadbetter (2016, see chapter X this volume) this does not lead to a sustainable development paradigm. However, the lightweight formalisation of the JSON schema approach as used in OM-JSON is already being extended by the Ocean Acidification community in a more formalised way, and future research will determine if this presents a better route for the extension of these core data models.

CONCLUSION

In this chapter, we have seen how the concept of Linked Ocean Data is exploiting the potential of third generation World Wide Web technology, specifically the Semantic Web and Linked Data, to provide interconnections between datasets published online. For sustainable development of the marine environment, where a holistic view of the marine system is required, Linked Data approaches allow for the easier integration of new data sources, particularly where relevant Ontology Design Patterns have been published and Linked Data can be created to those patterns. Finally, the emergence of Big Data streaming systems with small chunks of processing logic placed between fully re-analysable message queues presents a solution to the question of data being "Born Semantic" or "Born Linked" allowing new observations of the marine environment to be connected to the Linked Ocean Data cloud from the moment they are captured. Through the emergence of OM-JSON, the Born Semantic data can also be compliant with standards for accessing sensor observations through Web interfaces.

REFERENCES

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: towns, buildings, construction* (Vol. 2). Oxford, UK: Oxford University Press.

Alexander, K., & Hausenblas, M. (2009). Describing linked datasets-on the design and usage of void, the vocabulary of interlinked datasets. In *Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09).*

Berners-Lee, T. (2006). *Linked Data - Design Issues*. World Wide Web Consortium. Retrieved May 29, 2015 from http://www.w3.org/DesignIssues/LinkedData.html

Bray, T. (2013). *XML's 15th Birthday.* Retrieved February 2, 2016 from http://www.tbray.org/ongoing/When/201x/2013/02/10/XML-at-15

Cox, S. (2006). *Observations and measurements. Open Geospatial Consortium Best Practices Document*. Open Geospatial Consortium.

Cox, S. (2013, October). An explicit OWL representation of ISO/OGC Observations and Measurements. In SSN@ ISWC (pp. 1-18).

Cox, S. (2013, December). *Observations to information*. Abstract IN42A-01 presented at 2013, Fall Meeting, AGU, San Francisco, CA.

Cox, S., & Taylor, P. (2015). OM-JSON - a JSON implementation of O&M. Presentation to the Open Geospatial Consortium Technical and Planning Committee Sensor Web Enablement Domain Working Group, Nottingham, UK.

De Mauro, A., Greco, M., & Grimaldi, M. (2015, February). What is big data? A consensual definition and a review of key research topics. In AIP Conference Proceedings (Vol. 1644, pp. 97-104).

Diviacco, P., & Leadbetter, A. (2016). Balancing Formalization and Representation in cross-domain data management for sustainable development. In P. Diviacco, A. Leadbetter, & H. Glaves (Eds.), *Oceanographic and Marine Cross-Domain Data Management for Sustainable Development*. Hershey, PA: IGI Global.

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching*. Heidelberg, Germany: Springer.

Evans, B., Wyborn, L., Pugh, T., Allen, C., Antony, J., Gohar, K., . . . Bell, G. (2015). The NCI High Performance Computing and High Performance Data Platform to Support the Analysis of Petascale Environmental Data Collections. In Environmental Software Systems. Infrastructures, Services and Applications (pp. 569-577). Springer International Publishing. doi:10.1007/978-3-319-15994-2_58

Fernández, J. D., Martínez-Prieto, M. A., Gutiérrez, C., Polleres, A., & Arias, M. (2013). Binary RDF representation for publication and exchange (HDT). *Web Semantics: Science, Services, and Agents on the World Wide Web*, *19*, 22–41. doi:10.1016/j.websem.2013.01.002

Fredericks, J. (2015). Persistence of knowledge across layered architectures. In P. Diviacco, P. Fox, C. Pshenichny, & A. Leadbetter (Eds.), *Collaborative Knowledge in Scientific Research Networks*. Hershey, PA: IGI Global. doi:10.4018/978-1-4666-6567-5.ch013

Galiegue, F., Zyp, K., & Court, G. (2013). *JSON Schema: core definitions and terminology*. Retrieved October 12, 2015 from http://tools.ietf.org/html/draft-zyp-json-schema-04

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns: elements of reusable object-oriented software*. New York, NY: Pearson Education.

Gangemi, A. (2005). Ontology design patterns for semantic web content. In *The Semantic Web–ISWC 2005* (pp. 262–276). Berlin: Springer Berlin Heidelberg. doi:10.1007/11574620_21

Graybeal, J., Isenor, A., & Rueda, C. (2012). Semantic mediation of vocabularies for ocean observing systems. *Computers & Geosciences*, *40*, 120–131. doi:10.1016/j.cageo.2011.08.002

Guizzardi, G. (2006, July). The role of foundational ontologies for conceptual modeling and domain ontology representation. In *Databases and Information Systems, 2006 7th International Baltic Conference on* (pp. 17-25). IEEE. doi:10.1109/DBIS.2006.1678468

Hilbert, M. (2013). *Big data for development: From information-to knowledge societies*. Available at SSRN 2205145.

Hitzler, P., Krotzsch, M., & Rudolph, S. (2009). *Foundations of semantic web technologies*. CRC Press.

Hu, Y., Janowicz, K., Carral, D., Scheider, S., Kuhn, W., Berg-Cross, G., & Kolas, D. (2013). A geo-ontology design pattern for semantic trajectories. In *Spatial Information Theory* (pp. 438–456). Springer International Publishing.

Irish Marine Institute. (n.d.). *Sensor Observation Service*. Retrieved from: https://github.com/IrishMarineInstitute/sensor-observation-service

JCOMM. (2009). *Cookbook for Submitting Data in Real Time and In Delayed Mode.* Joint World Meteorological Organisation-Intergovernmental Oceanographic Commision Technical Commission for Oceanography and Marine Meteorology. Retrieved May 29, 2015 from http://www.jcomm.info/index.php?option=com_content&view=article&id=37

Kreps, J. (2014) *Questioning the lambda architecture*. O'Reilly Radar. Retrieved August 28, 2015 from http://radar.oreilly.com/2014/07/questioning-the-lambda-architecture.html

Krzywinski, M., Birol, I., Jones, S. J., & Marra, M. A. (2012). Hive plots—rational approach to visualizing networks. *Briefings in Bioinformatics*, *13*(5), 627–644. doi:10.1093/bib/bbr069 PMID:22155641

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., & Marra, M. A. et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. doi:10.1101/gr.092759.109 PMID:19541911

Lawrence, B., Lowry, R., Miller, P., Snaith, H., & Woolf, A. (2009). Information in environmental data grids. *Philosophical Transactions of the Royal Society of London A: Mathematical*, *Physical and Engineering Sciences*, *367*(1890), 1003–1014. doi:10.1098/rsta.2008.0237

Leadbetter, A. (2015). Linked ocean data. In T. Narock & P. Fox (Eds.), The Semantic Web in Earth and Space Science: Current Status and Future Directions. Amsterdam: IOS Press.

Leadbetter, A., Arko, B., Chandler, C., Shepherd, A., & Lowry, R. (2013). Linked Data: An Oceanographic Perspective. *Journal of Ocean Technology*, *8*(3), 7–12.

Leadbetter, A., & Lowry, R. (2012). *The NERC Vocabulary Server: Version 2.0*. Abstract IN51D-1709 presented at 2012, Fall Meeting, AGU, San Francisco, CA.

Leadbetter, A., Lowry, R., & Clements, D. (2014). Putting meaning into NETMAR - the open service network for marine environmental science. *International Journal of Digital Earth*, *7*(10), 811–828. doi:10.1080/17538947.2013.781243

Lowry, R. (2003). *EnParDis - Enabling parameter discovery.* Paper presented at 2003 Global Organisation for Earth System Science Portals. Central Laboratory of the Research Councils, Daresbury Laboratory.

Marz, N., & Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co.

Merati, N., & Burger, E. (2004). *Implementing Marine XML for NOAA Observing Data*. Paper presented at 2004 User Conference, Esri, Redlands, CA.

Rolling Deck. (n.d.). *Dataset Description for Rolling Deck to Repository (R2R) Linked Data*. Retrieved from: http://linked.rvdata.us/.well-known/void

Schaap, D., & Lowry, R. (2010). SeaDataNet - Pan-European infrastructure for marine and ocean data management: Unified access to distributed data sets. *International Journal of Digital Earth*, *3*(sup-1Supplement 1), 50–69. doi:10.1080/17538941003660974

UNESCO. (1987). GF3: A General Formatting System for geo-referenced data. Volume 2: technical description of the GF3 format and code tables. Paris: Intergovernmental Oceanographic Commission.

Vocab. (n.d.). Retrieved from: http://vocab.nerc.ac.uk/.well-known/void

Willighagen, E. (2014). Accessing biological data in R with semantic web technologies (No. e185v3). *PeerJ PrePrints*. Retrieved from https://dx.doi.org/10.7287/peerj.preprints.185v3

Woods Hole. (n.d.). Retrieved from: http://www.bco-dmo.org/.well-known/void

Zeng, W., Fu, C.-W., Müller Arisona, S., & Qu, H. (2013). Visualizing Interchange Patterns in Massive Movement Data. *Computer Graphics Forum*, *32*(3), 271–280. doi:10.1111/cgf.12114

ADDITIONAL READING

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, 205-227.

Kleppmann, M. (2015). *Apache Kafka, Samza, and the UNIX Philosophy of Distributed Data*. Retrieved October 12, 2015 from http://www.confluent.io/blog/apache-kafka-samza-and-the-unix-philosophy-of-distributed-data

Morrison, J. P. (2010). *Flow-Based Programming: A new approach to application development*. CreateSpace.

Narock, T., & Fox, P. (Eds.). (2015). *The Semantic Web in Earth and Space Science. Current Status and Future Directions*. IOS Press.

KEY TERMS AND DEFINITIONS

Big Data: Big Data can be simply defined as "having more data than I had yesterday, and not knowing what to do with it". Key aspects of Big Data are the data volume; the speed of data production and transfer (its velocity); and the wide variety of data types introduced.

Big Ocean Data: An application of the ideas of Big Data specifically to the marine science domain.

Controlled Vocabulary: A controlled vocabulary provides a way to organise knowledge for subsequent retrieval. Vocabularies are used in subject indexing schemes, subject headings, and their content can be organised hierarchically to create thesauri, taxonomies and other forms of knowledge organization systems. Controlled vocabulary schemes mandate the use of predefined, authorised terms that have been preselected by the individual or group governing the vocabulary, in contrast to natural language vocabularies where no such restriction is put in place.

Linked Data: Linked Data is a technique that uses the Word Wide Web to connect related data that was not previously linked, or uses the Web to lower the barriers to linking data which are currently linked using other methods. In practice, Linked Data often uses the Resource Description Framework model and Web addresses (as Uniform Resource Locators, or URLs) to achieve these linkages.

Linked Ocean Data: A subset of Linked Data, with a specific focus on the marine science domain.

Ontology: In computer science, an ontology formally represents knowledge as a hierarchy of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts.

Ontology Design Pattern: An Ontology Design Pattern is a reusable solution to a data modelling problem that commonly occurs across many different domains or within a wide variety of contexts within a single domain.

Resource Description Framework (RDF): RDF is a standard model for the exchange of data over the World Wide Web. The RDF data model is based upon the idea of making statements about resources in the form of subject-predicate-object expressions, known as triples. A classic example of a subject-predicate-object triple is “sky”-”has colour”-”blue”.

Semantic Web: The Semantic Web is the World Wide Web of data built on the Resource Description Framework model as a foundation for publishing and linking data online.

Streaming Data: Streaming data is an analytic computing paradigm that is focused on speed of throughput of data.

Topology: In the context of processing streams of data, a topology is a graph of computation. Each node in the graph contains the processing logic and the graph’s edges (or connections between the nodes) indicate the pathways of data between the nodes.