The eReefs data brokering layer for hydrological and environmental data

Jonathan Yu, Benjamin Leighton, Nicholas Car, Shane Seaton
and Jonathan Hodge

ABSTRACT

The environmental sciences are witnessing a data revolution as large amounts of data are being made available at an increasing rate. Many datasets are being published through operational monitoring programs, research activities and global earth observation virtual laboratories. An important aspect is the ability to query relevant metadata which can potentially provide useful information to discover, access and interpret environmental datasets, information about the data providers themselves, data services, data encodings, observation and measurement properties and data service endpoints. However, support for producing and accessing metadata descriptions in a flexible, extensible, easily integrated and easily discovered manner is lacking as current methods require interpreting multiple standards and formalisms. In this paper, we propose components to streamline discovery and access of hydrological and environmental data: a Data Provider Node ontology (DPN-O) which allows precise descriptions to be captured about datasets, data services and their interfaces; and a Data Brokering Layer which provides an Application Programming Interface (API) for registering metadata for discovery and query of registered DPN datasets. We discuss this work in the context of the eReefs project which is developing an integrated information platform for discovery and visualization of observational and modelled data of the Great Barrier Reef.

Key words | data brokering, data discovery, data integration, environmental data, Great Barrier Reef, ontologies

Jonathan Yu (corresponding author)
Benjamin Leighton
Land and Water, CSIRO,
Highett,
Victoria,
Australia
E-mail: *Jonathan.Yu@csiro.au*

Nicholas Car
Land and Water, CSIRO,
Dutton Park,
Queensland,
Australia

Shane Seaton
Land and Water, CSIRO,
Black Mountain,
ACT,
Australia

Jonathan Hodge
Oceans and Atmosphere, CSIRO,
Dutton Park,
Queensland,
Australia

INTRODUCTION

The eReefs project is developing an interoperable coastal information platform (IP) for the conservation of the Great Barrier Reef (Car 2013; Car & Hodge 2013) to meet a number of use cases specified in Car & Murray (2013) (see Table 1). This IP aims to provide integrated discovery, query and visualization across a number of datasets access via multiple scientific repositories in multiple agencies working across multiple domains. Besides implementation of the IP, best practice methodologies will be adopted by streamlining data acquisition and integration into the eReefs IP. End users, such as researchers and client applications, will be able to access data and metadata through the platform in the same manner regardless of data source while minimal customization of existing infrastructure will be required of data providers.

A harmonized Water Quality vocabulary has been developed and published at http://environment.data.gov.au/def/ (Simons *et al.* 2013a, 2013b; Cox *et al.* 2014a). This vocabulary harmonized a number of controlled vocabularies in the Australian context with the Chemical Entities of Biological Interest (CHEBI) ontology and Quantities, Units, Dimensions and Types (QUDT) ontology and is encoded as a Simple

doi: 10.2166/hydro.2015.165

Table 1 | Use cases listed according to categories recommended in Car & Murray (2013). Initial business analytics processes collected many use cases, all of which were able to be classified according to the categories in this table. Categories similar to these have been used to classify use cases for previous distributed systems

Title	Category
Discover data	1. End User
Access data	2. End User
Add new data to a service	3. Data and Functionality Provision
Add a new service	4. Data and Functionality Provision
List datasets not compliant with the eReefs IP data model	5. Enablement and Governance
Link vocabulary terms to external vocabularies	6. Cross-business Domain Integration
List all services not responding in a timely manner	7. System Maintenance

Knowledge Organization System (SKOS) vocabulary. It has been developed to support the eReefs IP and is published using the Spatial Information Services Stack Vocabulary Service (SISSVoc) (Cox *et al.* 2014b). The SISSVoc vocabulary search tool is used to help data providers and end users discover appropriate vocabulary definitions (see http://sissvoc.ereefs.info/search/).

In Yu *et al.* (2014), a methodology was proposed for enhancing data services using the water quality vocabularies. The methodology outlines how existing data services publishing netCDF-CF (http://cfconventions.org/, last accessed 31 March 2015) and WaterML2.0 (Taylor 2012; Taylor *et al.* 2013) can be enhanced with links to the water quality vocabulary using a 'Linked Data' approach and standardized web vocabularies. An example of the use of this methodology for enhancing netCDF metadata with links to the water quality vocabulary is given in the listing below. In the example below, the original netCDF metadata definition 'Nap_MIM', i.e. the first six lines, is annotated with links to related water quality vocabulary definitions in the subsequent five lines, which specify links to specific definitions regarding the observed scaled quantity, the unit of measure, the substance or taxon being observed, the medium of the observation and the procedure used to obtain the observation. Each of these links can then be resolved to semantic descriptions for those definitions which give more precise detail and links to other concepts. This provides a flexible and lightweight annotation method and resolution to more semantics than the original label in the netCDF header 'TSS, MIM SVDC on Rrs'.

```
float Nap_MIM(time, latitude, longitude) ;
Nap_MIM:_FillValue = -999.f ;
Nap_MIM:long_name = "TSS, MIM SVDC on Rrs" ;
Nap_MIM:units = "mg/L" ;
Nap_MIM:valid_min = 0.01209607f ;
Nap_MIM:valid_max = 226.9626f ;
Nap_MIM:scaledQuantityKind_id
  ="http://environment.data.gov.au/water/quality/def/property/
  solids-total_suspended" ;
Nap_MIM:unit_id = "http://environment.data.gov.au/water/
  quality/def/unit/MilliGramsPerLitre" ;
Nap_MIM:substanceOrTaxon_id = "http://environment.data.gov.
  au/water/quality/def/object/solids";
Nap_MIM:medium_id = "http://environment.data.gov.au/water/
  quality/def/object/ocean"
Nap_MIM:procedure_id = "http://data.ereefs.org.au/ocean-colour/
  MIM_SVDC_RRS" ;
```

Using domain definitions in the standardized web vocabularies, they allow existing datasets to be enhanced with consistent semantics. This in turn allows data to be more easily discovered, accessed, integrated and analysed. This approach is similar to general 'Linked Data' methodologies where resources are linked to other resources using the Resource Description Framework (RDF) and where both resource and resource relationship types are found in standardized web vocabularies. It is not currently possible to include RDF metadata in netCDF headers due to netCDF formatting restrictions however, netCDF-LD – an extension to the current netCDF standard – has also been proposed which will further allow the semantics of netCDF (www.unidata.ucar.edu/software/netcdf/, last accessed 31 March 2015) metadata to be linked with domain semantics and other web resources using semantic web technologies such as RDF (Yu *et al.* 2015). More work is required to implement it through the netCDF standard formats.

The eReefs Data Brokering Layer (DBL) (previously introduced in Car *et al.* 2014), uses the brokering pattern to provide an Application Programming Interface (API) for querying, filtering and facilitating access to relevant eReefs data provider node (DPN) descriptions, service implementations, related datasets and resolvable data endpoints via searching over harvested metadata and semantic definitions. The DBL is a key middleware component in the eReefs IP and provides the means for data providers to register their data services. This is implemented via simple RDF descriptions of both their data and the services that deliver them. The DBL also provides a RESTful API for client applications to query and filter relevant information about data provider nodes, their advertised service endpoints and their available datasets, however, specific details of the DBL implemented were not presented previously.

Providing middleware interfaces and data brokering as the eReefs DBL does add considerably to the flexibility of the eReefs system as a whole. eReefs follows several other distributed, environmental, data systems in implementing a brokering approach, such as Australia's NEII (Bureau of Meteorology 2014) and Global Earth Observation System of Systems (GEOSS) (Nativi *et al.* 2015). eReefs also adopt the general principles that Nativi *et al.* (2013) give regarding the brokering approach:

- Autonomy: keep existing data infrastructures as autonomous as possible;
- Subsidiarity: supplementing existing infrastructure and governance arrangements with mediation;
- Interconnection: connecting existing infrastructure with tooling and approaches;
- Low entry barrier: minimising resources needed to participate;
- Flexibility: accommodate existing and future systems and technologies; and
- Effectiveness: deliver the information needs of the users.

In this paper, we present details of the current eReefs implementation of the DBL and its API. We also show how the various components, such as the water quality vocabularies and the Observable Properties ontology, and methodologies, such as the eReefs conventions for metadata annotation of netCDF headers, are leveraged by the DBL to enhance discovery of datasets and how this approach differs from other brokering approaches. In the next section we present the Data Provider Node concept and its implementation as an Web Ontology Language (OWL) ontology for describing and annotating the semantics around data providers, the data provider nodes they are responsible for, their respective datasets and their web-resolvable endpoints. The following section presents the eReefs DBL, its design, implementation as a RESTful API and its use in client applications and front-ends. This is followed by a section where we present some implications from the DBL and then other related work. Finally, we conclude and consider future work in the final section.

DPN CONCEPT AND THE DPN ONTOLOGY

Environmental datasets can often exist in a multiplicity of representations. In a service-oriented architecture, datasets may be replicated across different service implementations via identical service interfaces, for example, a coverage dataset of chlorophyll-a concentration delivered via two instances of the Web Map Service (WMS), one implemented using Geoserver and the other using MapServer. In practice, due to efficiencies and preferences for certain implementations, datasets may also be delivered in a range of formats via different sets of service implementation and service interfaces, for example, the same coverage dataset of chlorophyll-a concentration may be delivered as netCDF via THREDDS (www.unidata.ucar.edu/software/thredds/current/tds/, last accessed 31 March 2015) and as JavaScript Object Notation (JSON) (www.json.org/, last accessed 31 March 2015) data via RESTful APIs. In both cases, the abstract notion of a single dataset persists beyond these concrete representations.

In Car *et al.* (2014), a notion of a conceptual dataset was introduced and mechanisms for capturing the semantics of these conceptual datasets were raised as requirements in order for linkages between the various realizations of the conceptual datasets as data views. Figure 1 depicts the conceptual dataset and its access through a range of web services.

Furthermore, in Car *et al.* (2014), the authors propose a nodal structure whereby web service owners establish a DPN, which catalogs Datasets in order to manage the governance of multiple data services and data owners participating in a distributed and federated IP. The idea is that DPNs would be able to deliver a number of data services and register the relevant metadata about datasets as well as the data services in a consistent, extensible, and

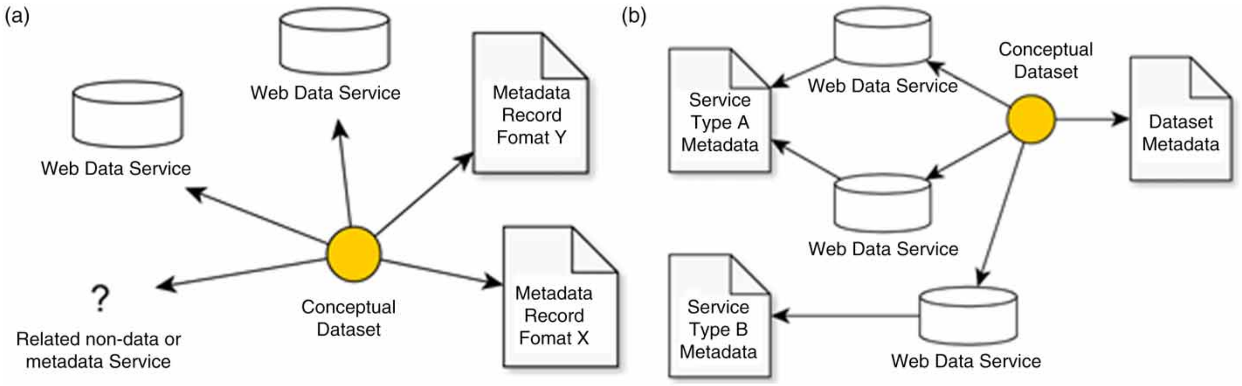


Figure 1 | (a) Conceptual dataset with a series of data, metadata and other views and (b) conceptual dataset linking to both dataset metadata and data services metadata.

machine-readable fashion. This would enable a central node, such as the eReefs IP central node, to index and harvest relevant information and make the respective datasets, data services, and information about the data owners discoverable and facilitate access to data assets, as shown in Figure 2.

Metadata catalogs may be used to facilitate the registration and harvesting of web services metadata. However, there are currently limitations for using metadata catalogs' entries to enable the discovery and harvesting of data services and their capabilities. Examples are Catalog Services for the Web (CSW) as standard interfaces (Senkler 2007). Implementations include GeoNetwork (http://geonetwork-opensource.org, last accessed 31 March 2015) and pyCSW (http://pycsw.org, last accessed 31 March 2015). Comprehensive Knowledge Archive Network (CKAN) (http://ckan.org, last accessed 31 March 2015) has CSW extensions available. These rely on strongly constrained data model-based methods. The current metadata catalog implementations are dataset-centric and limited in their ability to catalog data services as first-class objects.

On the other hand, RDF- and OWL-based ontologies can be used to describe metadata about dataset, data services and dataset relations, in a form that then allows this

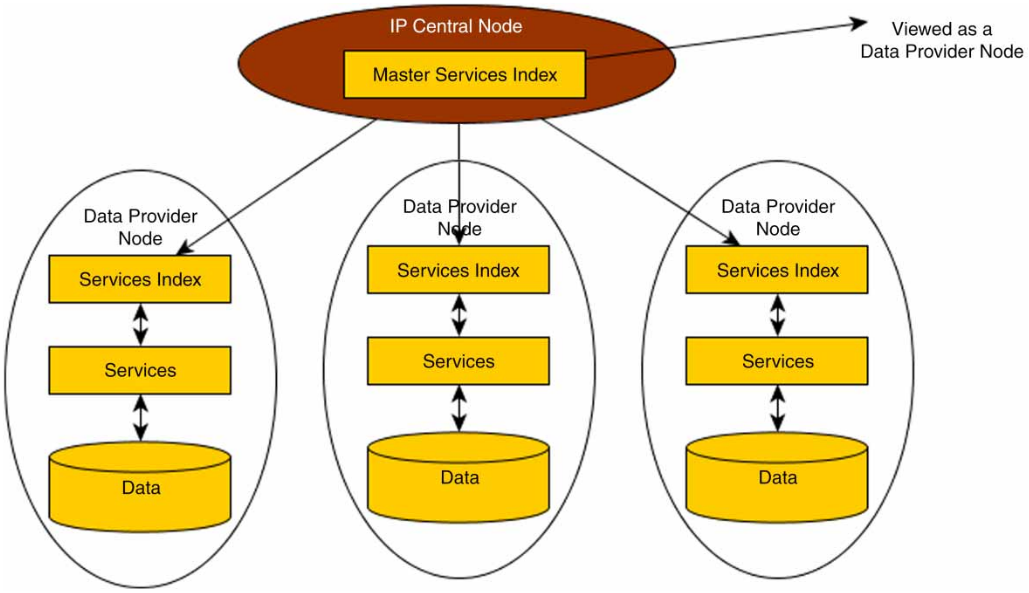


Figure 2 | A generic, nodal, IP architecture showing service indexing.

metadata to be published as Linked Data (Bizer *et al.* 2009). RDF provides a flexible platform for descriptions of resources, which are identified using Uniform Resource Identifiers (URIs), and linked internally and externally to form a graph of knowledge.

There have been a number of ontologies developed for describing datasets and services (see Table 2). The services ontologies shown in Table 2 below were designed with the use cases of service orchestration and choreography, which focus on representing the various bespoke service interfaces, inputs and outputs, formats for marshalling and unmarshalling objects and enterprise messaging automation. These include Web Service Modeling Ontology (WSMO) (Roman *et al.* 2005), OWL-S (Martin 2004), and SAWSDL (Kopecký *et al.* 2007) and reflects the lineage of these ontologies from the web services domain of Web Services Description Language (WSDL) and Simple Object Access Protocol. Other service ontologies provide lightweight descriptions of RESTful APIs which include Web Application Description Language (WADL) (Hadley 2009), and hRests (Kopecký *et al.* 2008) as shown in Table 2. Verborgh *et al.* (2014) provide a more in-depth survey of various kinds of approaches for semantic descriptions of REST APIs comparing lightweight approaches, logics-based and JSON-based approaches. Table 2 also presents a number of ontologies for describing datasets. In many cases, datasets can be subsets of larger datasets, for example a data cube consisting of a few dimensions taken from a multi-dimensional dataset. This is largely supported by the dataset ontologies shown in Table 2, however, the ontologies tend to differ in the way the dimensions of datasets are actually described. The RDF Data Cube (Cyganiak & Reynolds 2014) allows observed values to be defined natively as an RDF Dataset and allows the multi-dimensionality of the datasets to be represented easily. VoID (Alexander *et al.* 2011) defines a Dataset as 'a set of RDF triples that are published, maintained or aggregated by a single provider' capturing the governance of the data. However, most environmental datasets are provided in encodings other than RDF. Data Catalog Vocabulary (DCAT) describes a

Table 2 | Survey of dataset and service ontologies

Ontology	Dataset/service	Complexity	Main purpose
WSMO and WSMO lite	Service	Medium	The WSMO semantics that distinguishes four semantic aspects of services: function, behaviour, information model, and non-functional properties, which together form a basis for semantic automation and service orchestration
OWL-S	Service	Medium	Supports representation of web service descriptions using RDF and its behaviour for service orchestration
SAWSDL	Service	Medium	Enriches WSDL with semantic annotations in RDF for classifying, discovering, matching, composing, and invoking Web services
hRESTS	Service	Low	Modelling Web API details relevant for invocation support
SSWAP	Dataset/Service	Low	Uses RESTful architecture concepts and provides simple OWL ontology for describing Providers, Resources, Graph, Subject, and Object. Allows for ontology reasoning to support semantic search and service matchmaking
WADL	Service	Low	Provides XML-based descriptions of HTTP-based Web applications – web resources available, associations, methods and data format MIME types available
VOID	Dataset	Low	Supports descriptions of RDF dataset metadata. Assumes a dataset is natively encoded as RDF. Allows linking between datasets via void:LinkSet
RDF Data cube	Dataset	Medium	Provides an ontology to describe datasets as a set of observed values organized along a group of dimensions, together with associated metadata. There is a strong alignment with statistical dataset use cases and is based on the SDMX approach for statistical data exchange
DCAT	Dataset	Low	An RDF vocabulary for supporting interoperability between web accessible data catalogs. Defines a concepts for Dataset, Distribution, Catalog, Catalog record and reuses the FOAF concepts for Organisation and Person

dataset as 'a collection of data, published or curated by a single agent, and available for access or download in one or more formats' (W3C 2014). DCAT allows the dataset's metadata and its distribution means to be captured in RDF and does not natively represent the dataset itself using RDF. The DCAT Distribution class is used to allow various forms of a dataset to be defined, such as a download URL or an API.

From the brief survey of the current dataset and service ontologies, there are some well-covered areas but also some gaps regarding representing datasets in ways that capture the services used to deliver them, governance aspects of multiple data services and dataset owners participating in a distributed and federated IP. In particular, the DCAT ontology appears to align with the notion of a conceptual dataset that is decoupled from the data format, catalog record and distribution means. DCAT does not, however, allow the data services aspect to be characterized, although one may extend the definition of Distribution to define a Web Service class.

The services ontologies detail technical aspects of web service orchestration such as messaging protocols and formats, yet there are a number of established, non-ontology-based nevertheless standardized, data web services, some with well-known implementations. For example, the international spatial data standards body the Open Geospatial Consortium (OGC) has services such as the Web Feature Service (WFS) (Vretanos 2010) with implementations such as GeoServer (http://geoserver.org/, last accessed 31 March 2015). The academic and research collaborative program Unidata (www.unidata.ucar.edu/about/, last accessed 31 March 2015) has THREDDS (www.unidata.ucar.edu/software/thredds/current/tds/, last accessed 31 March 2015). These data web services are well known and widely used, supported by many software libraries, but are not based on either the WS* stack or RESTful practices, which did not exist when they were originally developed after their development, so are incompatible with the service ontologies listed above. Therefore, the gap which a services ontology would fill is the ability to describe metadata about the available web services – service interfaces, implementation and their relevant web endpoints.

To our knowledge, there is currently no ontology that can relate multiple services to a notion of a conceptual dataset, provide social and governance metadata about such a dataset and also provide lightweight descriptions of known services.

The DPN ontology

In light of the limitations in the offerings of the current dataset and service ontologies available, we have developed the DPN Ontology (DPN-O) (http://purl.org/dpn, last accessed 31 March 2015). The DPN-O allows semantics about a DPN to be captured in terms of its organization, its datasets and their associated service implementations or interfaces (see Figure 3).

The DPN-O contains classes for DPNs (institutions or sub-institutional groups) that contain Services which have various Interfaces. The list of described service types can be added to allowing IPs to deliver a growing range of services. The use of an ontology, implemented in RDF, allows the conceptual dataset entities to act as an index to which services are related. The RDF Individuals instantiated from the respective classes allow identity (a URI) to be minted for each of them.

In DPN-O, the definition of the dpn:Dataset and dpn:ServiceInterface classes are stubs, as they are expected to be substituted by classes from existing and well-accepted ontologies, such as DCAT for downloadable datasets and WSMO-lite service interfaces for service orchestration. The design of the DPN-O supports the descriptions of relationships between DPNs and the respective services, datasets, and organisations. Therefore this allows the representation of actual governance arrangements, datasets and their hosting arrangements, and technical deployments of data services.

Figure 4 shows a conceptual diagram of an instance of the DPN-O for the CSIRO Ocean Colour group who provide a THREDDS data service. The organization, its delivery node and associated data services (in this example only one) are shown, as is the link from the data service description to the actual THREDDS data service endpoint. Also shown is a link from the service description to a DBL. The DBL is described in the next section. The graph shown in Figure 4 is given in Figure 5 as an RDF encoding using the Turtle notation (www.w3.org/TR/turtle/, last accessed 31 March 2015).

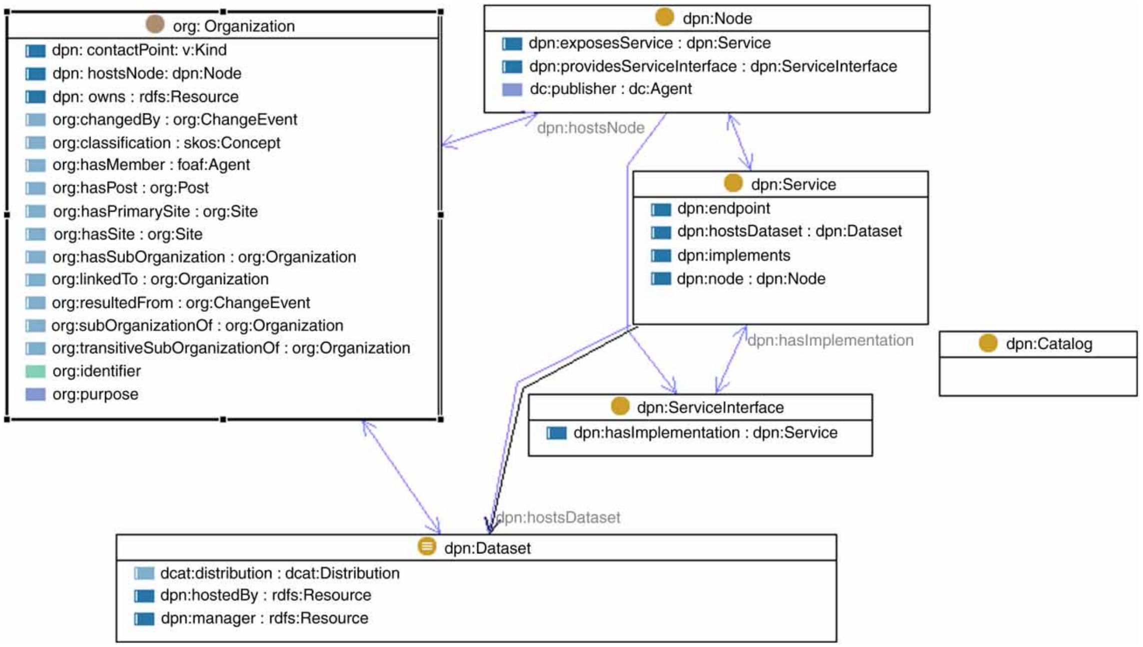

org: Organization
dpn: contactPoint: v:Kind
dpn: hostsNode: dpn:Node
dpn: owns : rdfs:Resource
org:changedBy : org:ChangeEvent
org:classification : skos:Concept
org:hasMember : foaf:Agent
org:hasPost : org:Post
org:hasPrimarySite : org:Site
org:hasSite : org:Site
org:hasSubOrganization : org:Organization
org:linkedTo : org:Organization
org:resultedFrom : org:ChangeEvent
org:subOrganizationOf : org:Organization
org:transitiveSubOrganizationOf : org:Organization
org:identifier
org:purpose
dpn:Node
dpn:exposesService : dpn:Service
dpn:providesServiceInterface : dpn:ServiceInterface
dc:publisher : dc:Agent
dpn:hostsNode
dpn:Service
dpn:endpoint
dpn:hostsDataset : dpn:Dataset
dpn:implements
dpn:node : dpn:Node
dpn:hasImplementation
dpn:Catalog
dpn:ServiceInterface
dpn:hasImplementation : dpn:Service
dpn:hostsDataset
dpn:Dataset
dcat:distribution : dcat:Distribution
dpn:hostedBy : rdfs:Resource
dpn:manager : rdfs:Resource

Figure 3 | The DPN ontology core elements.

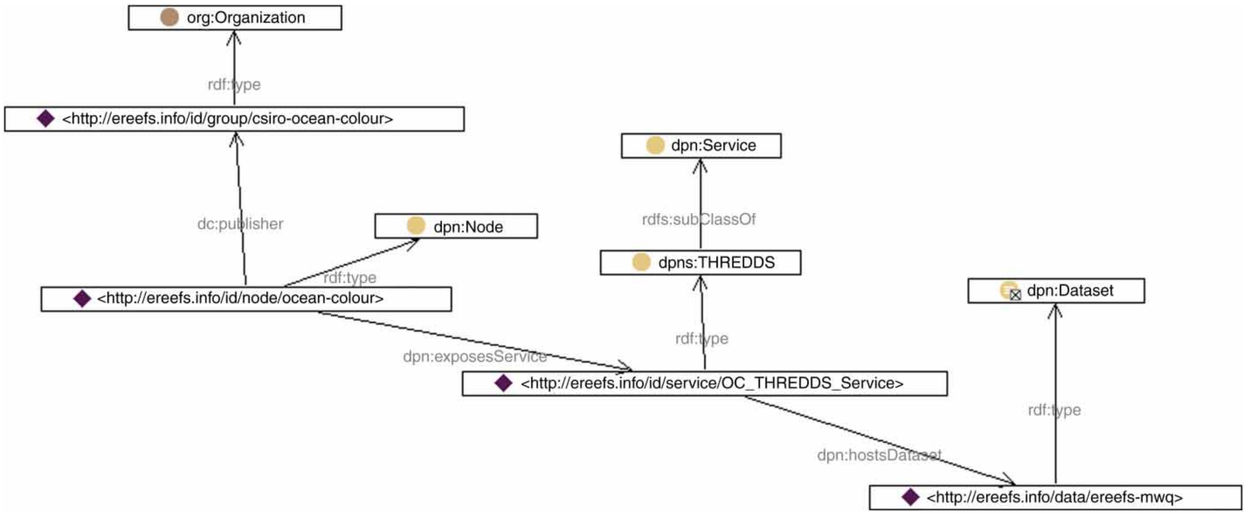
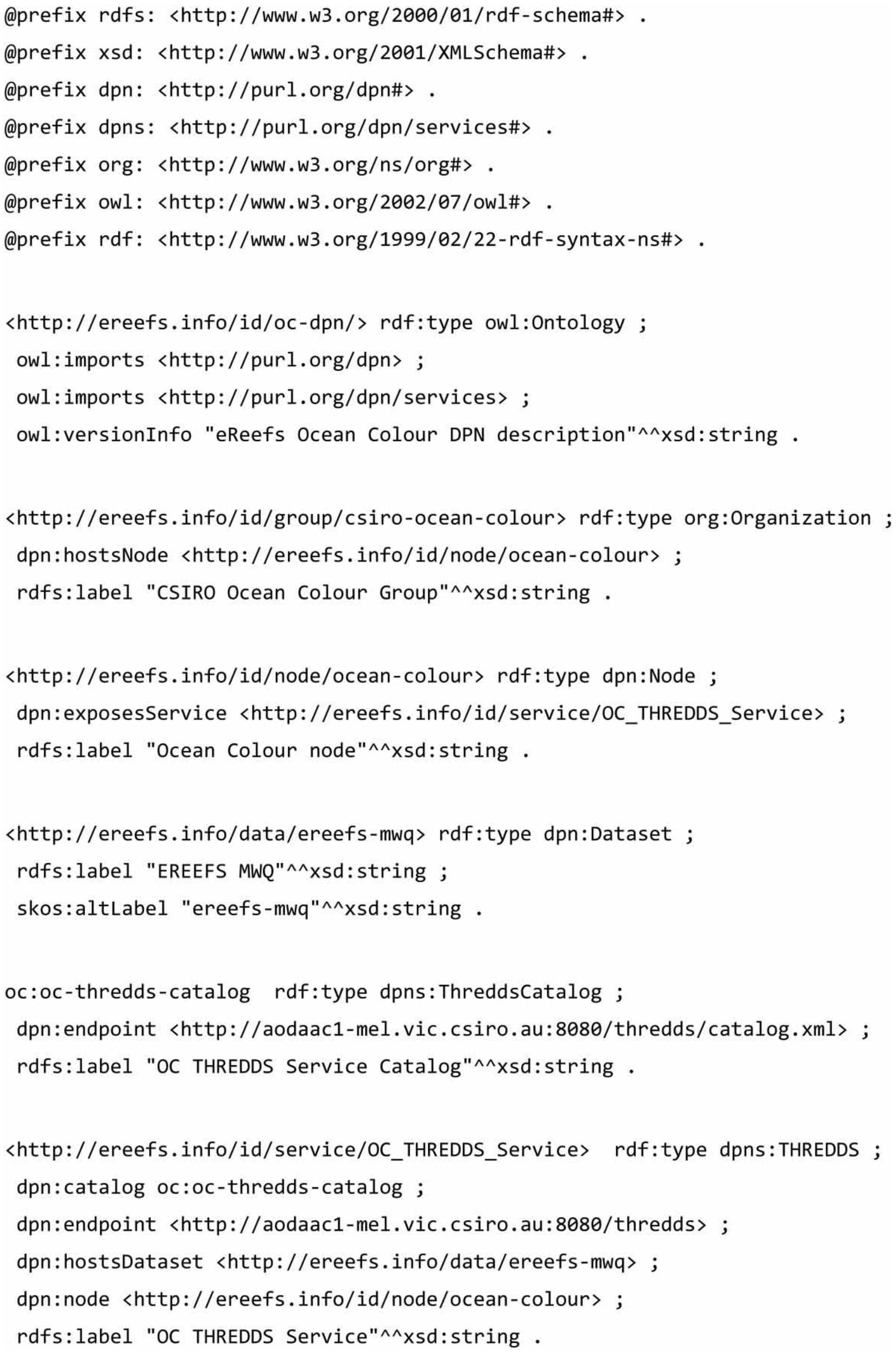

org:Organization
rdf:type
<http://ereefs.info/id/group/csiro-ocean-colour>
dpn:Service
rdfs:subClassOf
dc:publisher
dpn:Node
dpns:THREDDS
rdf:type
dpn:Dataset
<http://ereefs.info/id/node/ocean-colour>
rdf:type
dpn:exposesService
<http://ereefs.info/id/service/OC_THREDDS_Service>
rdf:type
dpn:hostsDataset
<http://ereefs.info/data/ereefs-mwq>

Figure 4 | Conceptual diagram for the CSIRO Ocean Colour DPN.

EREEFS DBL

The eReefs DBL allows mediation between client applications, APIs, widgets and the services and datasets to which DPNs provide access. A Linked Data approach is implemented where people and automated agents can use a 'Follow-Your-Nose' method to incrementally discover more data and metadata about things by following typed links from a starting point. The Hypertext Transfer Protocol (HTTP) and HTML and RDF are the mechanisms and data formalisms used for web link resolutions and machine readability. URIs provide identity to elements in



```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dpn: <http://purl.org/dpn#> .
@prefix dpns: <http://purl.org/dpn/services#> .
@prefix org: <http://www.w3.org/ns/org#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://ereefs.info/id/oc-dpn/> rdf:type owl:Ontology ;
 owl:imports <http://purl.org/dpn> ;
 owl:imports <http://purl.org/dpn/services> ;
 owl:versionInfo "eReefs Ocean Colour DPN description"^^xsd:string .

<http://ereefs.info/id/group/csiro-ocean-colour> rdf:type org:Organization ;
 dpn:hostsNode <http://ereefs.info/id/node/ocean-colour> ;
 rdfs:label "CSIRO Ocean Colour Group"^^xsd:string .

<http://ereefs.info/id/node/ocean-colour> rdf:type dpn:Node ;
 dpn:exposesService <http://ereefs.info/id/service/OC_THREDDS_Service> ;
 rdfs:label "Ocean Colour node"^^xsd:string .

<http://ereefs.info/data/ereefs-mwq> rdf:type dpn:Dataset ;
 rdfs:label "EREEFS MWQ"^^xsd:string ;
 skos:altLabel "ereefs-mwq"^^xsd:string .

oc:oc-thredds-catalog  rdf:type dpns:ThreddsCatalog ;
 dpn:endpoint <http://aodaac1-mel.vic.csiro.au:8080/thredds/catalog.xml> ;
 rdfs:label "OC THREDDS Service Catalog"^^xsd:string .

<http://ereefs.info/id/service/OC_THREDDS_Service>  rdf:type dpns:THREDDS ;
 dpn:catalog oc:oc-thredds-catalog ;
 dpn:endpoint <http://aodaac1-mel.vic.csiro.au:8080/thredds> ;
 dpn:hostsDataset <http://ereefs.info/data/ereefs-mwq> ;
 dpn:node <http://ereefs.info/id/node/ocean-colour> ;
 rdfs:label "OC THREDDS Service"^^xsd:string .
```

Figure 5 | RDF encoding of the CSIRO Ocean Colour DPN.

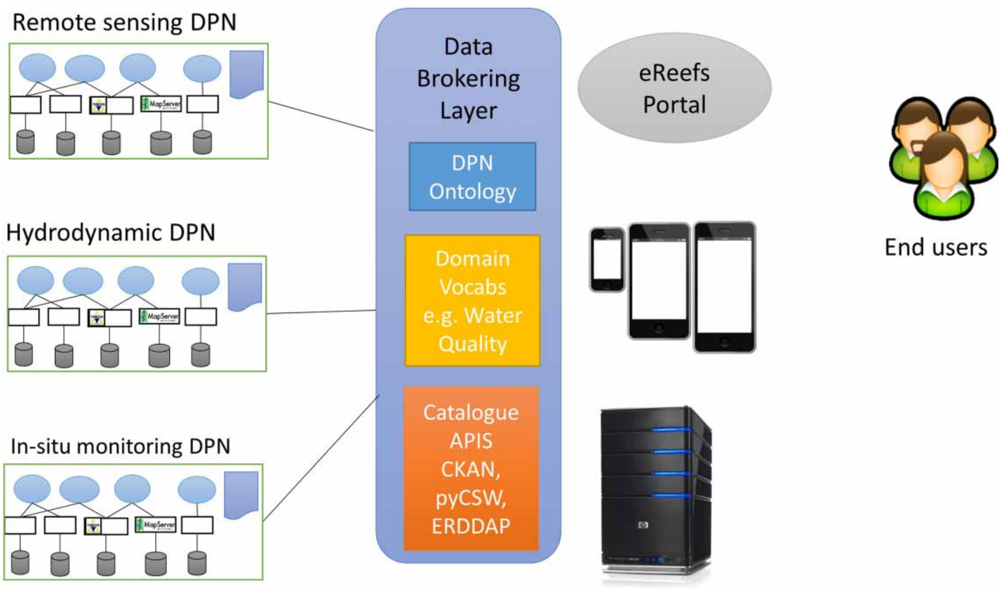


Figure 6 | Data brokering using the DBL. DPN descriptions are registered with the DBL. DBL provides APIs to facilitate data and service discovery.

the system – datasets, observed properties, services, and data providers – and HTTP the protocol for access to them. Ontologies, implemented in RDF, provide semantics about those identified elements, in a web-compatible form, to allow clients to traverse knowledge domains.

Organisations describe their data and service offerings as instances of the DPN-O. Those instances are then registered in a DBL which indexes their content, indexes and caches content to which they refer (such as data domain vocabularies and ontologies) and provides an API for leveraging the information. A DBL API then creates an IP consisting of multiple DPN and cached domain metadata. Figure 6 shows some eReefs DPNs, the eReefs DBL and example clients of the DBL API.

The DBL's RESTful API facilitates data and service discovery by client applications. It includes methods which align with the DPN ontology core elements as well as search functionality. The API's responses are encoded in JSON-LD which is a profile of JSON used for Linked Data as well as a simpler JSON serialisation for less verbose data descriptions (see Table 3). This allows a decoupling of the client applications from the services and more flexible data integration into the applications.

The eReefs DBL API instance endpoints given in Table 3 are available via the eReefs DBL URL: http://ereefs.org.au/dbl. The /dpn endpoint gives a list of DPNs registered with the DBL, which users can query further to get lists of services, datasets etc. The /catalog endpoint provides a list of particular implementations of services, for example the THREDDS data service catalog (refer to http://ereefs.org.au/dbl/catalog).

Table 3 | The DBL REST API

ID	REST API endpoint	HTTP method	Parameters	Response content type
1	/dpn/	GET	N/A	application/json
2	/catalog/	GET	N/A	application/json
3	/service/ /service/?_format = json-ld	GET	 _format	application/json application/ld + json
4	/dataset/	GET	N/A	application/json
5	/search	GET	term/uri offset limit prefLabel	application/json

The /service/ endpoint acts as a register of the DBL's services. HTTP Query String Arguments may be added to the /service/ URI to filter results. They provide a view of the semantics of the DPN and its available services, hosted datasets and service endpoints. Similarly, the /dataset/ endpoint return metadata on the available datasets.

Dataset classification

A key aspect to integrating data into the eReefs IP is adopting the eReefs conventions for annotating data with specific classification metadata using methodology given in Yu *et al.* (2014). This includes metadata about the substance or taxon of interest, scaled quantity kinds, units of measure and medium that the observations forming the data relate to. The choice of these metadata concepts is informed by the Observations & Measurement standard (Cox 2011) and formalized in the Observed Properties ontology (Cox *et al.* 2014a). Specific values for the various concepts are taken from vocabularies relevant to a particular IP. Figures 6 and 7 both show the eReefs DBL instance using a ‘Water Quality’ domain vocabulary for instances of concepts such as substance or taxon of interest. Navigable URIs uniquely identify concepts in the vocabulary and form the basis of the linked data approach to eReefs metadata.

For each data provider, a process for binding vocabulary URIs to data is required. A business analysis process takes place to establish the association between vocabulary concepts and data provider data. Typically, this occurs as a series of conversations between an eReefs vocabulary expert and a data provider administrator who has knowledge of the relevant science domain. The outcome of this process is an informal document describing data mappings between URIs, which uniquely identify vocabulary concepts, and data variables.

Metadata integration for eReefs data delivered using the netCDF format via THREDDS services sees the data provider adding *substanceOrTaxon*, *scaledQuantityKind*, *unit*, and *medium* attributes to the metadata header section of the data file and specifying values as navigable URIs to their definitions, which are taken from appropriate formalized vocabularies. For example, observations of chlorophyll concentrations are specified in the netCDF header such that the variable ‘Chl_MIM’ is annotated with a field for ‘substanceOrTaxon_id’ with the value set to the URI ‘http://environment.data.gov.au/water/quality/def/object/chlorophyll_a’, which associates the variable and the dataset with the ‘chlorophyll a’ concept in the vocabulary. A metadata indexing process harvests the metadata in the netCDF header and associates the http://environment.data.gov.au/water/quality/def/object/chlorophyll_a URI with all instances of its occurrence in metadata of datasets and variables across all eReefs data providers. In this way, heterogenous datasets that refer to common concepts but use different nomenclature (e.g. variable names) are linked. Figure 7 shows example URIs for the various attributes relating to *chlorophyll observation*.

By associating defined concepts with defined relationships and data, the eReefs infrastructure can effectively index data across many different DPN and present users with many axes of classification on which to find and judge data the DPN’s data is fit for particular tasks.

The DBL/search endpoint allows clients to query for data against declared semantics harvested from the dataset metadata which cross-references domain vocabulary terms. The search endpoint accepts either keywords or the specific URI for a term and returns datasets using that term or URI for example, for the dataset shown in Figure 7, ‘chlorophyll-a concentration’ would be found by using the search endpoint with keyword thus: http://ereefs.org.au/dbl/search?term=chlorophyll or with URI thus: http://ereefs.org.au/dbl/search?uri=http://environment.data.gov.au/def/object/chlorophyll. Other parameters allow paging of results and other delivery conveniences.

The DBL API presented in this paper is simple but provides a component for establishing a registry system for data providers and their hosted services and datasets. The DPN-O descriptions provide the necessary metadata about the DPNs and their service and datasets. Registry of these descriptions into a DBL instance allows data discovery about DPNs, services and datasets and facilitates querying and processing of content in a flexible way while preserving relevant contextual information.

The eReefs Viz Portal client demonstrator

A client application has been developed as a data discovery demonstrator using the DBL, called the eReefs Viz Portal (http://vizportal.meteor.com/). The Viz Portal is a lightweight client application which leverages the rich semantics available via the DBL. Human-readable labels annotated in the vocabulary concept descriptions and harvested metadata from the datasets and data services retrieved via the DBL are rendered in the Viz Portal. Figure 8 shows a workflow for searching using the term ‘chlorophyll’

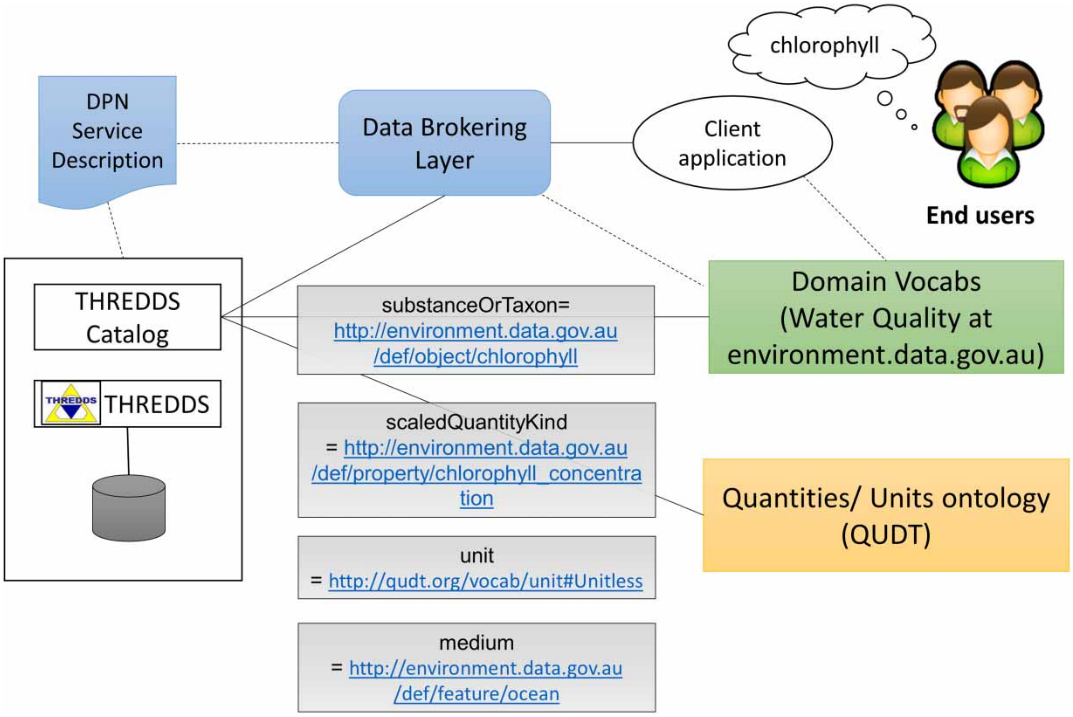

chlorophyll
DPN Service Description
Data Brokering Layer
Client application
End users
THREDDS Catalog
THREDDS
substanceOrTaxon= http://environment.data.gov.au/def/object/chlorophyll
Domain Vocabs (Water Quality at environment.data.gov.au)
scaledQuantityKind = http://environment.data.gov.au/def/property/chlorophyll_concentration
Quantities/ Units ontology (QUDT)
unit = http://qudt.org/vocab/unit#Unitless
medium = http://environment.data.gov.au/def/feature/ocean

Figure 7 | DBL harvests metadata from services and facilitating client applications discover data.

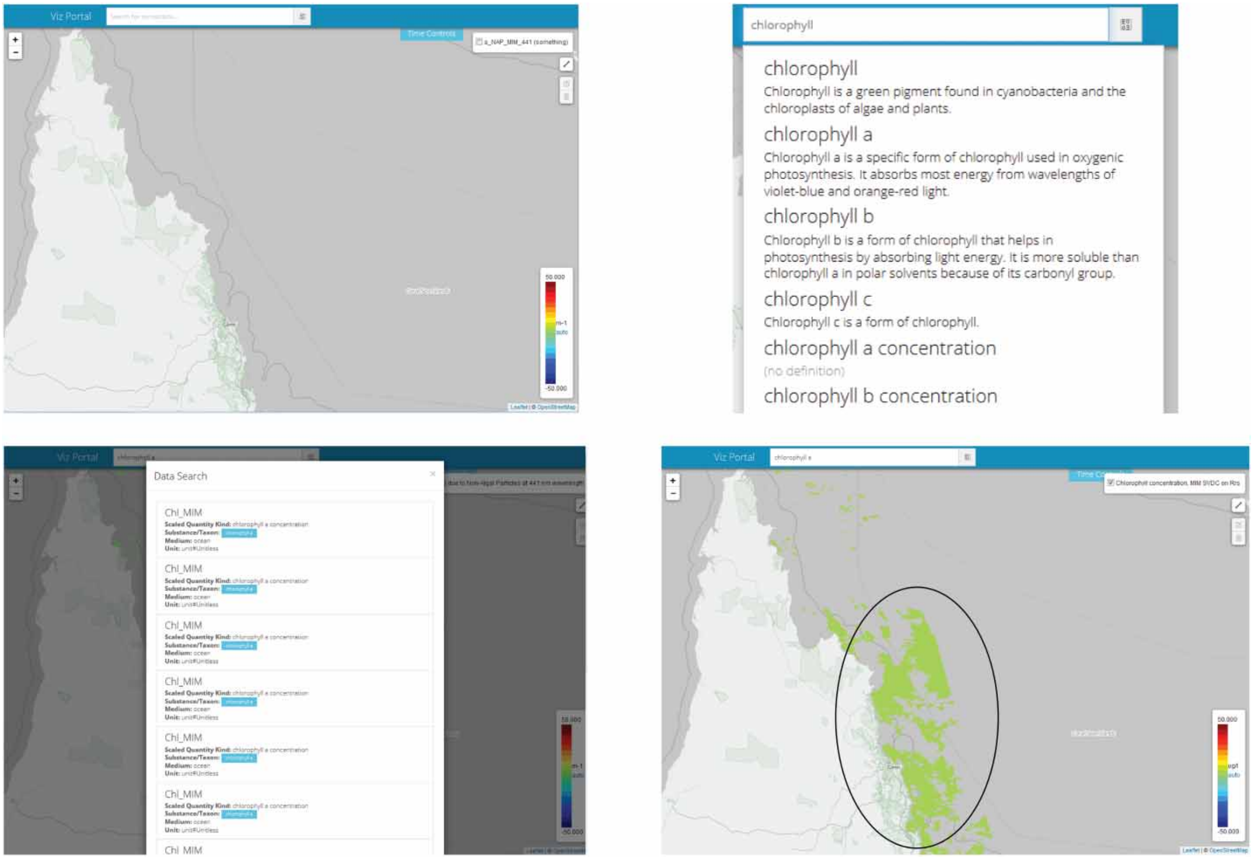
allowing a user to drill down to the specific vocabulary definition, retrieve datasets related to the selected vocabulary definition and visualize them on the portal map interface.

Other possible clients for the DBL include map visualization widgets, faceted data search, and automated generation of environmental report cards. Without a mediating layer to the data services and datasets, each client would require software libraries to handle binding and access to the data individually. The DBL provides an API to allow these clients to be developed in a light-weight fashion where the functionality of data query and access is already taken care of.

DISCUSSION

The DBL provides a mechanism for registering DPN-O instances and enables rich data and services discovery. In this paper, we have presented an implementation of the DBL in which DPNs use well-known data web services such as THREDDS and harvests relevant metadata from them. A cache of DPN metadata is built and indexed within the DBL which allows powerful querying, filtering and access to data across potentially very heterogeneous data sources which may yet be compared on many axes of classification. As the DBL is a web service API, many client applications and user interfaces can be made to utilize it to meet a wide range of applications such as map visualisations, faceted data search, and automated generation of environmental report cards. The DBL also allows a client application to be decoupled from the data services themselves which allows for dynamic discovery of datasets, e.g. a client application that binds its data queries via a DBL for gridded data for ‘chlorophyll concentration’ datasets will receive all related datasets via a vocabulary-based search across all registered DPNs. As new DPNs are added and registered into a DBL, there is no added work required for such client applications to receive updated datasets.

The current process for defining a DPN-o instance and the binding of vocabulary definitions for a given DPN is manually performed on behalf of the data providers. The intention of the current phase of eReefs was to demonstrate its application for data brokering and integration. The cost



chlorophyll

chlorophyll
Chlorophyll is a green pigment found in cyanobacteria and the chloroplasts of algae and plants.

chlorophyll a
Chlorophyll a is a specific form of chlorophyll used in oxygenic photosynthesis. It absorbs most energy from wavelengths of violet-blue and orange-red light.

chlorophyll b
Chlorophyll b is a form of chlorophyll that helps in photosynthesis by absorbing light energy. It is more soluble than chlorophyll a in polar solvents because of its carbonyl group.

chlorophyll c
Chlorophyll c is a form of chlorophyll.

chlorophyll a concentration
(no definition)

chlorophyll b concentration

Data Search

Chl_MIM
Scaled Quantity Kind: chlorophyll a concentration
Substance/Taxon: [illegible]
Medium: ocean
Unit: unit#Unitless

Chl_MIM
Scaled Quantity Kind: chlorophyll a concentration
Substance/Taxon: [illegible]
Medium: ocean
Unit: unit#Unitless

Chl_MIM
Scaled Quantity Kind: chlorophyll a concentration
Substance/Taxon: [illegible]
Medium: ocean
Unit: unit#Unitless

Chl_MIM
Scaled Quantity Kind: chlorophyll a concentration
Substance/Taxon: [illegible]
Medium: ocean
Unit: unit#Unitless

Chl_MIM
Scaled Quantity Kind: chlorophyll a concentration
Substance/Taxon: [illegible]
Medium: ocean
Unit: unit#Unitless

Chl_MIM
Scaled Quantity Kind: chlorophyll a concentration
Substance/Taxon: [illegible]
Medium: ocean
Unit: unit#Unitless

Chl_MIM

Figure 8 | (a) Top left: Viz Portal screenshot; (b) Top-right: User searches for dataset by keyword against labels in the vocabulary service; (c) Bottom-left: Matching datasets returned from DBL; (d) Bottom-right: Data visualisation of Chlorophyll observations.

of implementing the set of DPN-o instances for the data providers at the moment is small as there are five DPNs, although as an increasing number of DPNs are required to be integrated, the manual approach is not efficient.

The data brokering approach presented in this paper, which uses the DBL, the DPN-O and the respective environmental vocabulary definitions, provides a number of benefits for the eReefs project. It provides the project with added flexibility and extensibility in being able to develop multiple client applications and in registering additional or deregistering data services on-demand as it decouples the binding of data assets from the client applications.

Despite the benefits of the DBL and the brokering architecture, there is a cost in specifying both the DPN-O instance and the bindings to the vocabulary definitions. Therefore, tools to automate the generation of DPN-o instances are required to facilitate DPNs to be registered on-demand, e.g. web-based user interfaces as well as REST Web APIs. Similarly, additional tooling is required to guide users in selecting appropriate vocabularies for binding to vocabulary definitions for configuring netCDF headers. An evaluation of these tools and their effectiveness in reliably and simply developing the DPN-o instances and mappings is also required. These tools are being developed as items of future work. Since the DBL indexes service endpoints, it can be used at the access point for service conformance (both structural and performance) testing in order to ensure services meet certain IP expectations or contracts. Currently, test frameworks bind to static definitions of service endpoints and test classes are hard-coded for each type of service implementation. The ability to define tests based on service types and expected outputs from inputs queried by the DBL would allow a flexible way of testing conformance for the

DPN services. Further work is required to develop a test framework which leverages the DBL.

Despite the limitations of current metadata catalog implementations, e.g. Catalog Service for the Web, CKAN, and ERDDAP, they provide functionality to discover and harvest metadata describing datasets metadata. Thus, additional work is required to interface them so that metadata from these catalog implementations will be able to be harvested and handled by the DBL.

RELATED WORK

A number of prior and current efforts are underway to develop integrated data platforms especially in the environmental domain. The DataOne (Allard 2012) initiative has developed an integrated data platform bringing together multiple disciplines and organisations to provide tooling and data to support biological, ecological, and environmental science and research. The DataOne platform provides a means for sharing data, tooling and findings. Thus DataOne's scope is much broader than the eReefs use cases and includes collaborative tool development and a wider stakeholder engagement. DataOne does recognize the need for a nodal approach with the concept of a DataOne Member Node, which may include existing or new repositories. The work presented in this paper provides an ontology for describing these nodes as DPNs and tooling for precise semantic definitions of the datasets as well as support service endpoint discovery in a lightweight fashion.

Catalog services have been recognized as useful components in integrated data platforms. The USGS Geo Data Portal integration (Blodgett *et al.* 2012) is a platform for brokering access to geospatial vector and gridded datasets via data services with the use of catalog metadata to facilitate users to interpret and select aggregated datasets via a portal. In contrast to the USGS Geo Data Portal, the approach in the eReefs project is to enable richer dataset and service metadata capture using semantic web technologies so that the utility of metadata can be broadened beyond users to allow machines to facilitate richer data discovery.

The EVOp project (Vitolo *et al.* 2015) seeks to also establish an integrated data platform and a portal and has focused on the ability to allow model selection for datasets (e.g. between TOPMODEL and FUSE) as well as extracting data from a wide range of data sources via *ad-hoc* methods (e.g. web-scraping) and service-based real-time data endpoints. The authors recognize that 'self-describing data formats would be a better solution to store and transfer environmental data as they could integrate metadata information and standardised definitions of domain-specific variables and uncertainties', which is the approach the eReefs project has taken.

The DPN-O introduced in this paper has defined the dpn:Dataset, dpn:Service and dpn:ServiceInterface classes as stubs, which provides the option of further specialising them with definitions from existing and well-accepted ontologies. The reason for this is to ensure minimal ontological commitment and opportunities for greater reuse of the DPN-O than if implementation details were defined. This also allows the opportunity for multiple formalisms to be used with the DPN-O as required in different contexts and implementations. Therefore, this design accommodates a wide range of metadata formalisms. Relevant ISO standards may be used as specialisations of these classes – namely ISO 19119 for describing services metadata; and ISO 19115 for describing metadata for geographic datasets. These ISO standards have been developed under the governance of the Technical Committee ISO/TC 211 for Geographic Information. Both ISO standards are done using platform-neutral Universal Modeling Language (UML) modelling. The current practice is to map these to XML-based implementations, however, mapping the conceptual models to OWL/RDF-based implementations would enable these standards to be used as options to align or specialise the above DPN-O classes. OWL-S has been used in previous work to align the ISO standards with OWL/RDF ontologies (Yue *et al.* 2007), however, as noted above under the section 'DPN concept and the DPN ontology', in some cases, OWL-S provides a more heavyweight approach than the proposed approach of using DPN-O. The use of the DPN-O approach provides a lightweight alternative to OWL-S by providing the option to specify a small amount of information about the service and the dataset for client to be discoverable and useful. The DPN-O can also be specialised further to describe services and datasets in much more detail if necessary, for example, with the full suite of the above service and dataset metadata standards. Future work is required to demonstrate the use of such services and dataset ontologies and models, e.g. ISO 19115 and 19119, for encoding metadata about relevant services and datasets using the DPN-O.

eReefs shares the aim of the GEOSS project (Nativi *et al.* 2015) in an attempt to reduce the effort required by data contributors by brokering access to service-delivered data. Unlike GEOSS' Ranking Algorithm approach presented in Nativi *et al.* (2015), the eReefs DBL uses a separation between a dataset's domain metadata (metadata of the form usually encountered in data catalogs) and service metadata, which indicates how a dataset is made available. This allows discovery of data independently of data access and does not impose a judgement call on data access via a rank. It does have the overhead of requiring service type (or service interface) definitions with type vocabularies, as per the DPN-O. When service definitions and description vocabulary terms are used, dataset access can be related unambiguously to users, rather than proxied with a rank. This more open-ended approach was taken to ensure that dataset access evolution could be catered for and to give choice to potential dataset users. The rise in popularity of non-standardised data services such as InfluxDB (http://influxdb.com/) for timeseries data in the commercial domain in place of standards such as the Sensor Observation Service, make it clear that, if possible, new or less well-established services need to be catered for in the environmental domain. Our brokering approach only requires semantic descriptions of a service' endpoints for this to occur and will allow users to judge the service's utility.

CONCLUSION AND FUTURE WORK

In this paper we have presented the DPN concepts and ontology formulation for describing data owners, datasets and data services. We have also presented the DBL which provides a general purpose API for querying, filtering and facilitating access to sets of DPNs and their data holdings via harvesting their metadata and linking them to definitional datasets such as domain vocabularies. We have presented an instance of the DBL as a key middleware component in the eReefs IP. It provides the means for registering metadata about a data provider, its set of data services as a DPN and the datasets via a simple RDF description using the DPN ontology. We discussed how the various components, such as the water quality vocabularies, the Observable Properties ontology and the eReefs conventions for data annotation, are leveraged by the DBL to enhance discovery of datasets. We also provided example uses of the DBL in prototype client application called the eReefs Viz Portal.

This paper has identified gaps in the existing dataset and service ontologies and has presented an ontology developed to capture a DPN concept and the idea of a conceptual dataset and related services. The DPN-O description allows semantic representation of organizations, groups, DPN, datasets, service types, service implementations and their web endpoints using semantic web technologies. The data delivered via DPNs can be annotated with links to domain concepts, for example, water quality observable properties like chlorophyll-a concentration. DPN-O is an improvement on existing ontologies by proposing simple constructs for the use cases outlined in this paper. DPN-O descriptions are lightweight and extensible, easily allowing any number of existing and new service types to be linked to DPN's datasets. Since semantic web technologies adopt the open-world assumption, descriptions can be simple or highly detailed in line with DPN owners' wishes and expectations of metadata utility.

Further work is required to develop tools to automate the process of creating DPN-O instances as well as binding vocabulary definitions to the respective data services. Also, future work will be undertaken to develop additional client applications to meet other eReefs use cases including mobile applications and generating reports as well as conformance checking of data assets registered in eReefs through the DBL.

ACKNOWLEDGEMENTS

This research was funded through the eReefs research project by CSIRO, the Science and Industry Endowment Fund (SIEF) and the Great Barrier Reef Foundation (GBRF).

REFERENCES

Alexander, K., Cyganiak, R., Hausenblas, M. & Zhao, J. 2011 *Describing Linked Datasets with the VoID Vocabulary*. W3C

Interes. Gr. Note. Available from: www.w3.org/TR/void/; last accessed 31 March 2015.

Allard, S. 2012 DataONE: facilitating eScience through collaboration. *J. eScience Librariansh.* **1**, 4–17.

Bizer, C., Heath, T. & Berners-Lee, T. 2009 Linked data – the story so far. *Int. J. Semant. Web Inf. Syst.* **5**, 1–22.

Blodgett, D., Booth, N., Kunicki, T., Walker, J. & Lucido, J. 2012 Description of the U.S. Geological Survey Geo Data Portal Data Integration Framework. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **5**, 1687–1691.

Bureau of Meteorology 2014 *National Environmental Information Infrastructure: Reference Architecture*, Environmental Information Programme Publication Series, document no. 4, Bureau of Meteorology, Canberra, Australia.

Car, N. J. 2013 The eReefs Information Architecture. In: *Proceedings of the 20th International Congress on Modelling and Simulation* (J. Piantadosi, R. S. Anderssen & J. Boland, eds). Modelling and Simulation Society of Australia and New Zealand, Adelaide, Australia, pp. 831–837.

Car, N. J. & Hodge, J. 2013 eReefs: distributed data, a unified picture. In: *Proceedings of the 7th eResearch Australasia Conference. eResearch 2013 Conference Secretariat, Brisbane, Australia*, pp. 1–4.

Car, N. J. & Murray, N. 2013 Implementing and adapting the WRON-RM Use Case categories for eReefs: aiming for Interoperable Systems' requirements analysis best practice. In: *Proceedings of the 20th International Congress on Modelling and Simulation. Adelaide, Australia*, pp. 1–6.

Car, N. J., Yu, J., Cox, S. J. D., Stenson, M. P., Atkinson, R. & Fitch, P. 2014 A services framework and support services for environmental information communities. In: *Proceedings of the 11th International Conference of Hydroinformatics (HIC). New York, NY, USA*.

Cox, S. J. (ed.) 2011 *ISO 19156:2011 – Geographic information – observations and measurements*. Available from: www.opengeospatial.org/standards/om; last accessed 31 March 2015.

Cox, S. J. D., Simons, B. A. & Yu, J. 2014a A harmonized vocabulary for water quality. In: *Proceedings of the 11th International Conference of Hydroinformatics (HIC)*. IWA Publishing, New York, NY, USA.

Cox, S., Yu, J. & Rankine, T. 2014b SISSVoc: A linked data API for access to SKOS vocabularies. *Semant. Web* (in press). Doi:10.3233/SW-140166

Cyganiak, R. & Reynolds, D. 2014 *The RDF Data Cube Vocabulary. W3C Recomm*. www.w3.org/TR/vocab-data-cube/ (accessed 31 March 2015).

Hadley, M. 2009 *Web Application Description Language. W3C Memb. Submiss*. www.w3.org/Submission/wadl/ (accessed 31 March 2015).

Kopecký, J., Vitvar, T., Bournez, C. & Farrell, J. 2007 SAWSDL: Semantic annotations for WSDL and XML schema. *IEEE Internet. Comput.* **11**, 60–67.

Kopecký, J., Gomadam, K. & Vitvar, T. 2008 hRESTS: An HTML microformat for describing RESTful Web services. In: *Proceedings International Conference on Web Intelligence*. IEEE, Sydney, Australia, pp. 619–625.

Martin, D. (ed.) 2004 *OWL-S: Semantic Markup for Web Services. W3C Memb. Submiss*. www.w3.org/Submission/OWL-S/ (accessed 31 March 2015).

Nativi, S., Craglia, M. & Pearlman, J. 2013 Earth science infrastructures interoperability: the brokering approach. *IEEE J. Selected Topics Appl. Earth Observ. Remote Sens.* **6** (3), 1118–1129.

Nativi, S., Mazzetti, P., Santoro, M., Papeschi, F., Craglia, M. & Ochiai, O. 2015 Big data challenges in building the Global Earth Observation System of Systems. *Environ. Model. Softw.* **68**, 1–26.

Roman, D., Keller, U., Lausen, H., de Bruijn, J., Lara, R., Stollberg, M., Polleres, A., Feier, C., Bussler, C. & Fensel, D. 2005 Web service modeling ontology. *Appl. Ontol.* **1**, 77–106.

Senkler, K. 2007 *OpenGIS Catalog Services Specification 2.0.2 – ISO Metadata Application Profile. 07-006r1*. www.opengeospatial.org/standards/cat (accessed 12 October 2015).

Simons, B. A., Yu, J. & Cox, S. J. D. 2013a Defining a water quality vocabulary using QUDT and ChEBI. In: *Proceedings of the 20th International Congress on Modelling and Simulation* (J. Piantadosi, R.S. Anderssen & J. Boland, eds). Modelling and Simulation Society of Australia and New Zealand, Adelaide, Australia, pp. 2548–2554.

Simons, B. A., Yu, J. & Cox, S. J. D. 2013b Water quality vocabulary development and deployment. In: *Proceedings of the American Geophysical Union Fall Meeting*, Abstract IN53D-1586. San Francisco, USA.

Taylor, P. 2012 *WaterML 2.0: Part 1- Timeseries. OGC Implemented Standard 10-126r3*. www.opengeospatial.org/standards/waterml (accessed 31 March 2015).

Taylor, P., Cox, S., Walker, G., Valentine, D. & Sheahan, P. 2013 WaterML2.0: development of an open standard for hydrological time-series data exchange. *J. Hydroinform.* **16** (2), 425–446.

Verborgh, R., Harth, A., Maleshkova, M., Stadtmüller, S., Steiner, T., Taheriyan, M. & Van de Walle, R. 2014 REST: advanced research topics and practical applications. In: *REST: Advanced Research Topics and Practical Applications* (C. Pautasso, E. Wilde & R. Alarcon, eds). Springer, New York, NY, pp. 69–89.

Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C. J. A. & Buytaert, W. 2015 Web technologies for environmental Big Data. *Environ. Model. Softw.* **63**, 185–198.

Vretanos, P. A. 2010 *OpenGIS Web Feature Service 2.0 Interface Standard*. http://portal.opengeospatial.org/files/?artifact_id=39967 (accessed 31 March 2015).

W3C 2014 *Data Catalog Vocabulary (DCAT)*. www.w3.org/TR/vocab-dcat (accessed 31 March 2015).

Yu, J., Simons, B. A., Car, N. & Cox, S. J. D. 2014 Enhancing water quality data service discovery and access using standard vocabularies. In: *Proceedings of the 11th International Conference of Hydroinformatics (HIC)*. IWA Publishing, New York, NY, USA.

Yu, J., Car, N. J., Leadbetter, A., Simons, B. A. & Cox, S. J. D. 2015 Towards linked data conventions for delivery of environmental data using netCDF. *Environ. Softw. Syst. Infrastruct. Serv. Appl.* **448**, 102–112.

Yue, P., Di, L., Yang, W., Yu, G. & Zhao, P. 2007 Semantics-based automatic composition of geospatial Web service chains. *Comput. Geosci.* **33**, 649–665. doi:10.1016/j.cageo.2006.09.003.

First received 31 March 2015; accepted in revised form 22 August 2015. Available online 28 October 2015