Earth Science Informatics (2020) 13:537–553
https://doi.org/10.1007/s12145-020-00445-w

METHODOLOGY ARTICLE

A modular approach to cataloguing marine science data

Adam Leadbetter[1] · Will Meaney[1] · Elizabeth Tray[2] · Andrew Conway[1] · Sarah Flynn[1] · Tara Keena[1] · Caoimhín Kelly[1] · Rob Thomas[1]

Received: 16 August 2019 / Accepted: 21 January 2020 / Published online: 8 February 2020

Abstract

The ability to access and search metadata for marine science data is both a key requirement for answering fundamental principles of data management (making data Findable, Accessible, Interoperable and Reusable) and also in meeting domain-specific, community defined standards and legislative requirements placed on data publishers. This paper describes a modular data model to answer the functional requirements developed from these drivers and illustrates how this data model can be operationalised. The ability of this solution to meet the FAIR principles is then assessed.

Keywords Data management · Data catalogue · Marine science data · FAIR principles · General data protection regulation

Introduction

In 2016, Wilkinson et al. introduced the FAIR principles of data management - that research data should be Findable, Accessible, Interoperable and Reusable. In order to meet the requirements of the Findable aspect of FAIR data, a dataset must be described by rich metadata in a searchable resource and the dataset must be assigned a clearly labelled persistent, unique identifier. The metadata describing the data resource should be released with a clear data usage license, detailed data provenance and ensure that the metadata meet domain-relevant community standards.

Online metadata catalogues for environmental data have been documented since early in the history of the World Wide Web (Günther et al., 1996). This early approach was focussed on helping users to discover, or find, data relevant for a given topic and to access it quickly in a user-friendly manner. It was also beneficial in meeting legislative requirements, such as the European Access to Information on the Environment Regulations (European Parliament, 2003). Even at this stage the issue of semantic interoperability of metadata and data was recognised as being important. As new paradigms of data have come to the fore, this conclusion has further grown in importance (Hilbring & Usländer, 2006; Proctor et al., 2010; Tanhua et al., 2019).

In the context of environmental Big Data, Vitolo et al. (2015) call for the use of data catalogues to allow the discovery of data services and their functionality. However, they point out that semantic heterogeneity is a hurdle which must be overcome in searching through catalogue services. Leadbetter et al. (2014) and Leadbetter & Vodden (2016) demonstrate how interoperable, homogenous semantics can provide improved knowledge-building and cross-disciplinary data integration in environmental data catalogues.

One such cross-disciplinary activity is Marine Spatial Planning (MSP) which is concerned with the management of the distribution of human activities in space and time in and around seas and oceans to achieve ecological, economic and societal objectives and outcomes (Ehler et al., 2019). Nylén et al., 2019 include as one of their steps in the MSP data process the establishment of a metadata catalogue for the data to be used in the process. Within their framework, the catalogue should be able to differentiate between the original versions of existing spatial data and newly created data products derived from one or more original datasets. This differentiation should also include the processing steps taken to generate the new data products. The data catalogue should also be able to handle both observed and modelled data, and for modelled data to provide information on the input parameters to the model and the methods employed by the model. Flynn et al. (2019) conclude that a data cataloguing system for MSP can allow the availability and suitability of data for the MSP process to be assessed at regular review cycles. Friddell et al.,

✉ Adam Leadbetter
adam.leadbetter@marine.ie

[1] Marine Institute, Rinville, Oranmore Co Galway, Ireland

[2] Galway-Mayo Institute of Technology, Galway, Ireland

2014 demonstrate that in other cross-disciplinary topics, in their case polar research, modularity is required in order to represent datasets, projects or programmes and other polar data resources within the catalogue system.

Marine Spatial Planning is also a European legislative requirement (European Parliament, 2014), as are other data integration programmes including the Marine Strategy Framework Directive (European Parliament, 2008) and the INSPIRE Spatial Data Infrastructure. A data catalogue should recognise these targets and look to meet the technical requirements that they set as well as highlighting which datasets may be relevant to them. These include, for example, the delivery of ISO19115/19139 standard metadata to comply with the INSPIRE Spatial Data Infrastructure (Craglia & Annoni, 2007). In addition to legislative requirements, community standards should also be adhered to, such as the European Directory of Marine Environmental Datasets (Schaap & Lowry, 2010) and the Marine Community Profile (Proctor et al., 2010).

Therefore, in the sphere of marine science data management, the need for a modular approach to data cataloguing which is designed to meet a number of requirements highlighted above (see Table 1) can be clearly seen. In this paper we describe a data cataloguing system developed at and in use at the Marine Institute, Ireland and will expand on the data model used in developing the catalogue; discuss the approach taken to implementing the catalogue; and discuss our findings and future work.

Data model

The data model used within this modular catalogue is focused on a number of high-level concepts and their inter-relationships, illustrated in Fig. 1. These concepts are modularly developed as classes within the data model and are described below. Examples of instances of the classes are given in the text and are also summarised in Table 2.

Dataset

First is the high-level Dataset class (Fig. 2). It may combine many different parameters, collected at multiple times and locations, using different instruments. A Dataset is linked to its storage and retention information and the classification, including licensing, associated with the Dataset under a machine actionable data policy. This machine actionable data policy is derived from a set of business rules associated with the data classifications laid out in the institutional data policy (such as Marine Institute, 2017). Therefore, a Dataset which is marked as containing personal data, as defined by the European General Data Protection Regulation (Voigt & Von dem Bussche, 2017) or business sensitive data will not be made publicly available. Examples of a Dataset include an institution's entire research vessel Conductivity-Temperature-Depth profile archive; or a spatial dataset such as the distribution and abundance of cetacean species within an exclusive economic zone.

Table 1 Functional requirements for a marine data cataloguing system

Requirement	Reference(s)
Provide an accessible, searchable metadata repository	Günther et al., 1996 Wilkinson et al., 2016
Assign datasets a clearly labelled, unique persistent identifier	Wilkinson et al., 2016
Data usage licenses are clear to users	Wilkinson et al., 2016
Detailed data provenance is provided, including data processing and identifiers linking raw and processed datasets and associated documentation	Wilkinson et al., 2016 Nylén et al., 2019
Domain-relevant metadata standards are met	Wilkinson et al., 2016
Legislative and technical metadata standards are met	Günther et al., 1996 Craglia & Annoni, 2007 Wilkinson et al., 2016
Semantic homogeneity and interoperability between metadata records is facilitated	Günther et al., 1996 Vitolo et al., 2015
Metadata must describe data services as well as datasets	Craglia & Annoni, 2007 Vitolo et al., 2015
Both observed and modelled datasets should be handled	Nylén et al., 2019
Modules or classes should be provided to represent metadata on a number of entity types including datasets and funding programmes or projects	Friddell et al., 2014

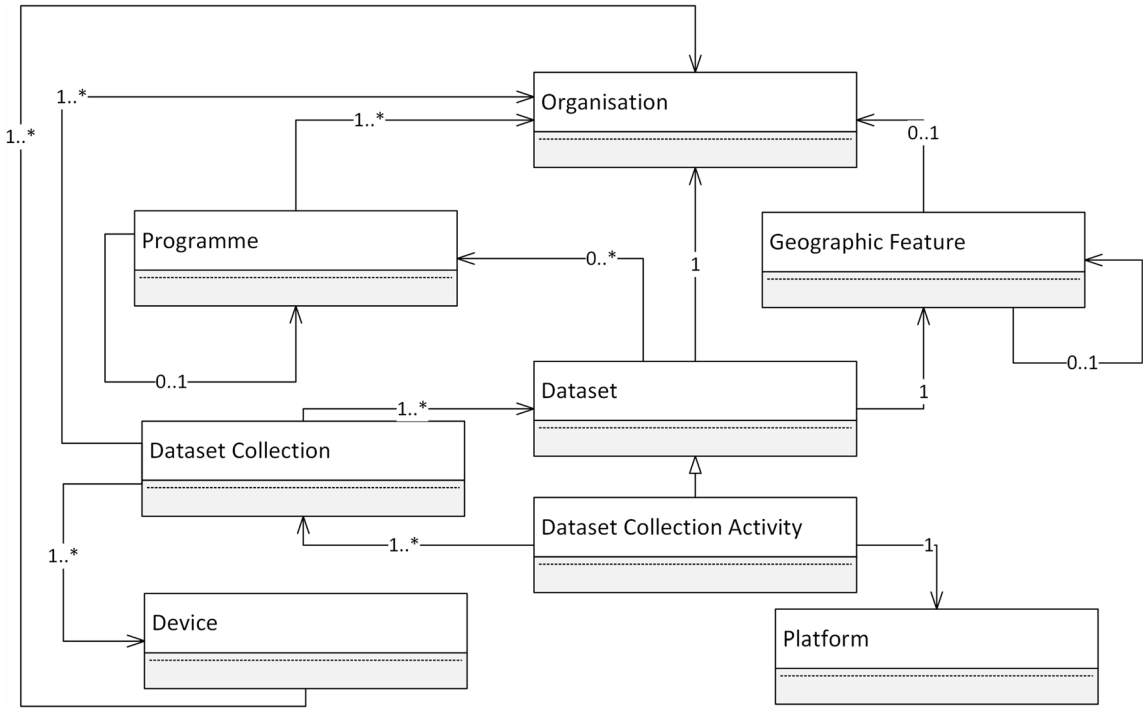

Fig. 1 A high-level overview of the data model used in the modular data catalogue approach. The overall class structure is shown in the Unified Modelling Language

Dataset Collection Activity

Related to a Dataset is a Dataset Collection Activity (Fig. 2). This class specialises the Dataset in that it has a mandatory end date and also a mandatory platform element, which indicates the vehicles, structures or organisms capable of bearing instruments or tools for the collection of physical, chemical, geological or biological samples or data. Examples of a Dataset Collection Activity include a research vessel survey or cruise; or the deployment of a moored buoy at a specific location for a given time period.

Platform

Within the INSPIRE spatial data infrastructure, the Environmental Monitoring Facilities component describes the environmental monitoring facility (a research vessel, a satellite) as a spatial object in the context of INSPIRE and observations and measurements linked to the environmental monitoring facility. (INSPIRE TWG EMF, 2013). The Platform class (Fig. 3) of this catalogue system seeks to carry the attributes required to complete an Environmental Monitoring Facilities instance when combined with details

Table 2 Examples of instances of the classes in the Data Catalogue data model

Class	Example(s)
Dataset	• An institution's entire research vessel Conductivity-Temperature-Depth profile archive
	• Distribution and abundance of cetacean species within an exclusive economic zone
Dataset Collection	• The Conductivity-Temperature-Depth profiles taken on a research vessel survey
	• A time series of atmospheric weather conditions recorded during the deployment of a sea-surface monitoring buoy
Dataset Collection Activity	• A research vessel survey or cruise
	• The deployment of a moored buoy at a specific location for a given time period
Geographic Feature	• A sampling location
	• A research vessel survey track
	• A polygon defining a lake or river catchment area
Platform	• A research vessel, such as the *RV Celtic Explorer*
	• An individual Argo programme drifting profiling float

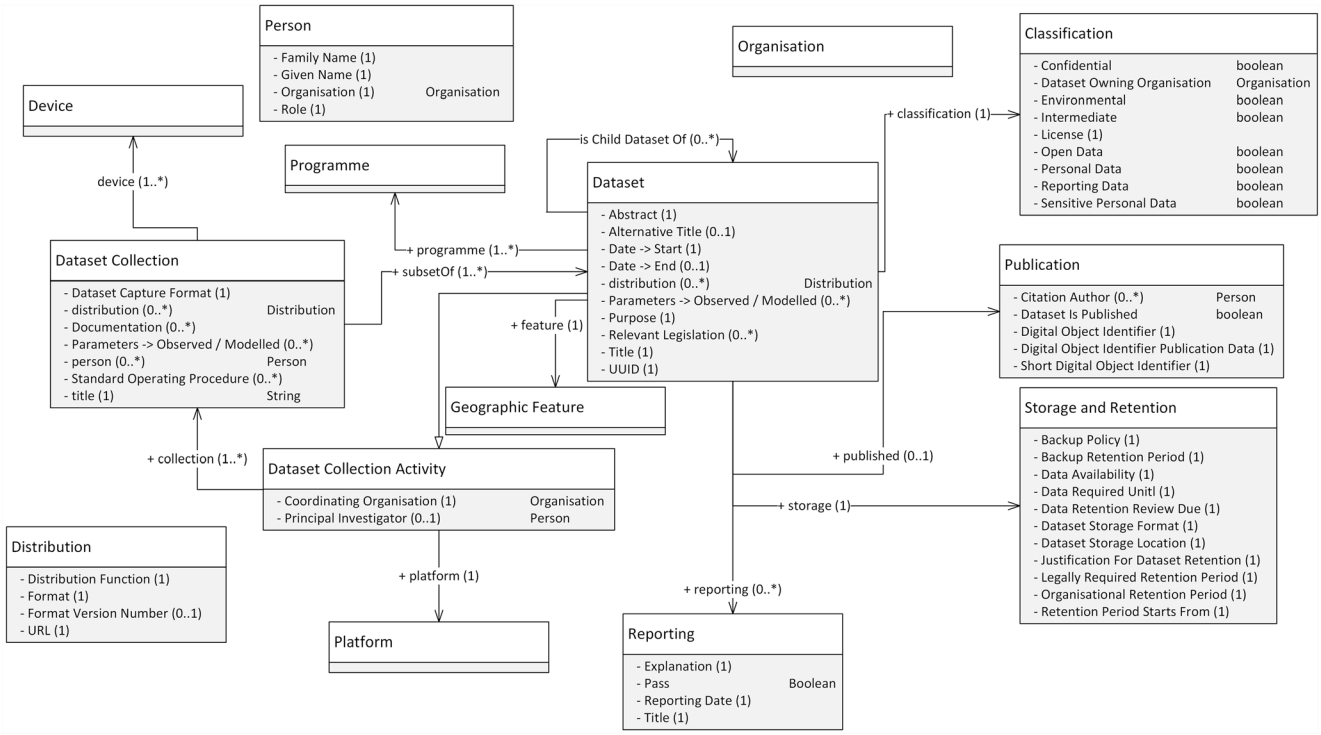

Fig. 2 A more detailed UML view of the Dataset, Dataset Collection Activity and Dataset Collection classes

from the Dataset Collection Activity class. It is also synonymous with the GeoLink class Platform which describes a "physical object of significance enabling observations resulting in a Dataset" (Krisnadhi et al., 2015). To this end a Platform instance is attributed with: its platform type; whether or not it is a mobile platform; which environmental regime it operates in; its operational start date, and if applicable, end date; and which Organisation is responsible for the platform. Where available, the International Council for the Exploration of the Seas platform code is also attributed to the Platform. Example instances of the Platform class include a research vessel, such as the *RV Celtic Explorer*, or an individual Argo programme drifting profiling float.

Dataset Collection

The Dataset Collection class (Fig. 2) is used to provide a link between a Dataset Collection Activity (e.g. a research vessel based survey; a deployment of a mooring) and a Dataset. As such, the Dataset Collection may be a subset of both the data collected by the Dataset Collection Activity (a limited set of the full parameters from that Activity) and the Dataset (possibly limited in time and/or parameter space). The Dataset Collection is linked to both a Dataset Collection Activity and a Dataset; and to the Device(s) used to sample the environment for a given range of parameters. An example of a Dataset Collection may be the Conductivity-Temperature-Depth profiles taken on a research vessel survey allowing the individual sensors to be connected to the activity and the calibration of those sensors to be connected with the associated measurements. A further example could be the time series of atmospheric weather conditions recorded during the deployment of a sea-surface monitoring buoy which allows for the change of sensors at service intervals of the buoy to be properly tracked within the catalogue.

Geographic Feature

A Geographic Feature (Fig. 3) is a mandatory attribute of a Dataset Collection Activity, and a recommended attribute of a Dataset. The Geographic Feature within this data catalogue model is closely related to the Open Geospatial Consortium and International Organisation for Standardisation's Simple Feature Access model (Herring, 2011). To this extent, the Geographic Feature class stores the geographic coordinates of points, lines, and polygons and the feature type for both the Simple Feature Access model and the European Commission's INSPIRE spatial data infrastructure. An instance of the Geographic Feature class may be attributed as a child of another Geographic Feature in order to build hierarchical networks of Geographical Features, such as river catchments and sea areas. Further attributes of a Feature within this model are the Coordinate Reference System used to define the latitude and longitude of the point, line or polygon; a URL to a definition of the Geographical Feature; and an organisation responsible for the Geographical Feature.

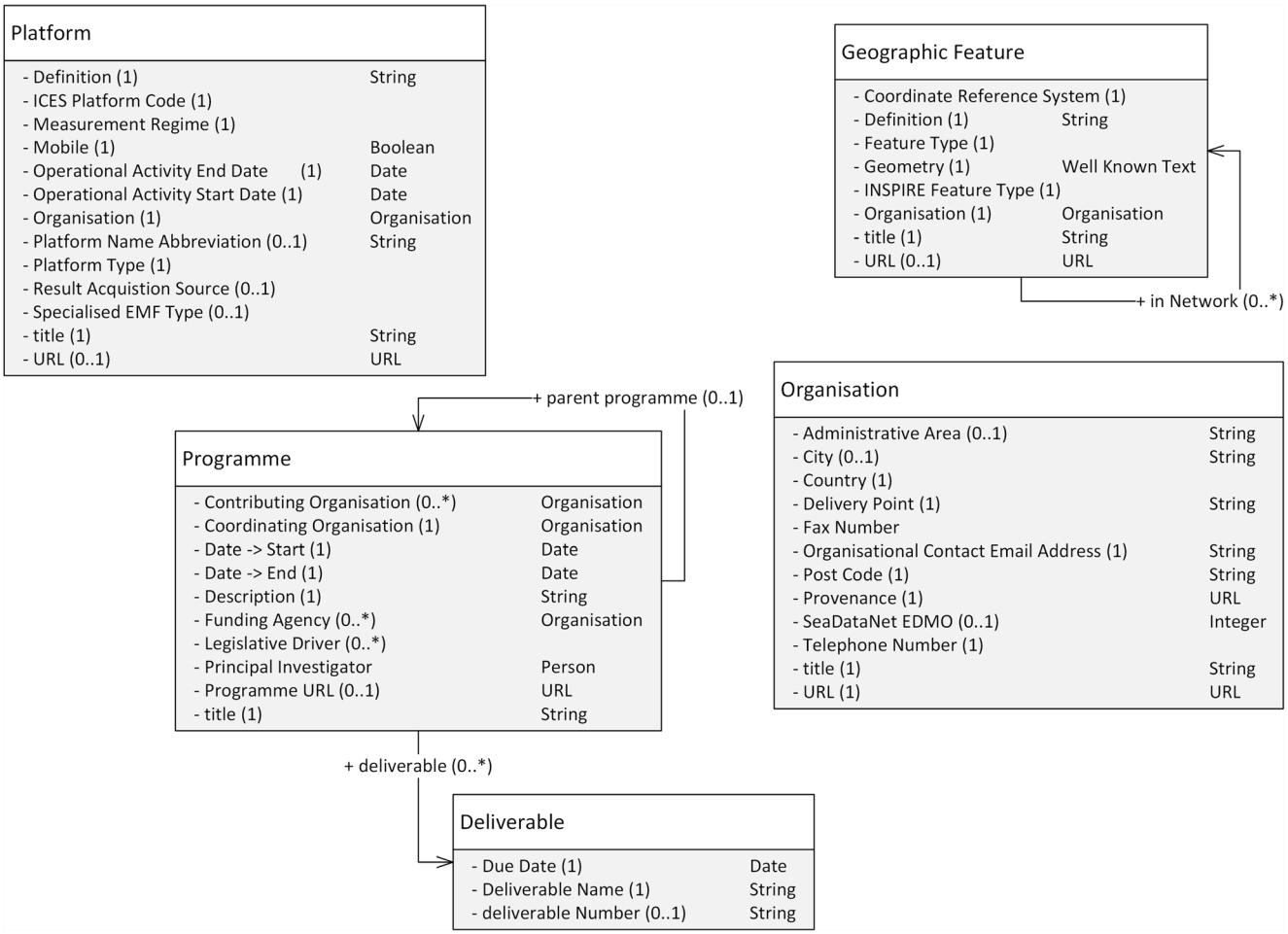

Fig. 3 A more detailed UML view of the Geographic Feature, Organisation Platform and Programme classes

Example instances of the class are a sampling location; a research vessel survey track; or a polygon defining a lake or river catchment area.

Programme

The Programme class (Fig. 3) is similar in scope to the EarthCube GeoLink ontology's Program class in that instances represent a "formally recognized scientific effort receiving significant funding, requiring large scale coordination" (Krisnadhi et al., 2015). An instance of the Programme class may have a coordinating organisation, and a number of contributing and funding organisations as well as the name of an individual who is the principal investigator of the Programme. A Programme is time bound by a start date and an optional end date, and may have a URL link to a website describing the Programme. A Programme may have a number of deliverables associated with it. An instance of the Programme class may also be the child of another instance of the same class.

Device

As stated above, Dataset Collection Activity takes place via a Platform and is linked to a Dataset through a Dataset Collection which describes the deployment of a Device on a Platform. The Device class (Fig. 4) is designed to allow a SensorML (Botts and Robin, 2007) record to be constructed for a given Device instance. As such, an instance of the Device class carries the input and output parameters of the Device, its measurement units, its manufacturer, operating organisation and start and end dates. It also carries links to the documentation regarding the calibration history of the Device. The Device class is more detailed than the similar GeoLink class of Instrument as it holds the Device's serial number as well as the instrument type from a controlled vocabulary.

Organisation

The Organisation class (Fig. 3) is designed to capture the details of research institutes, data holding centres, monitoring agencies, governmental and private organisations that are in

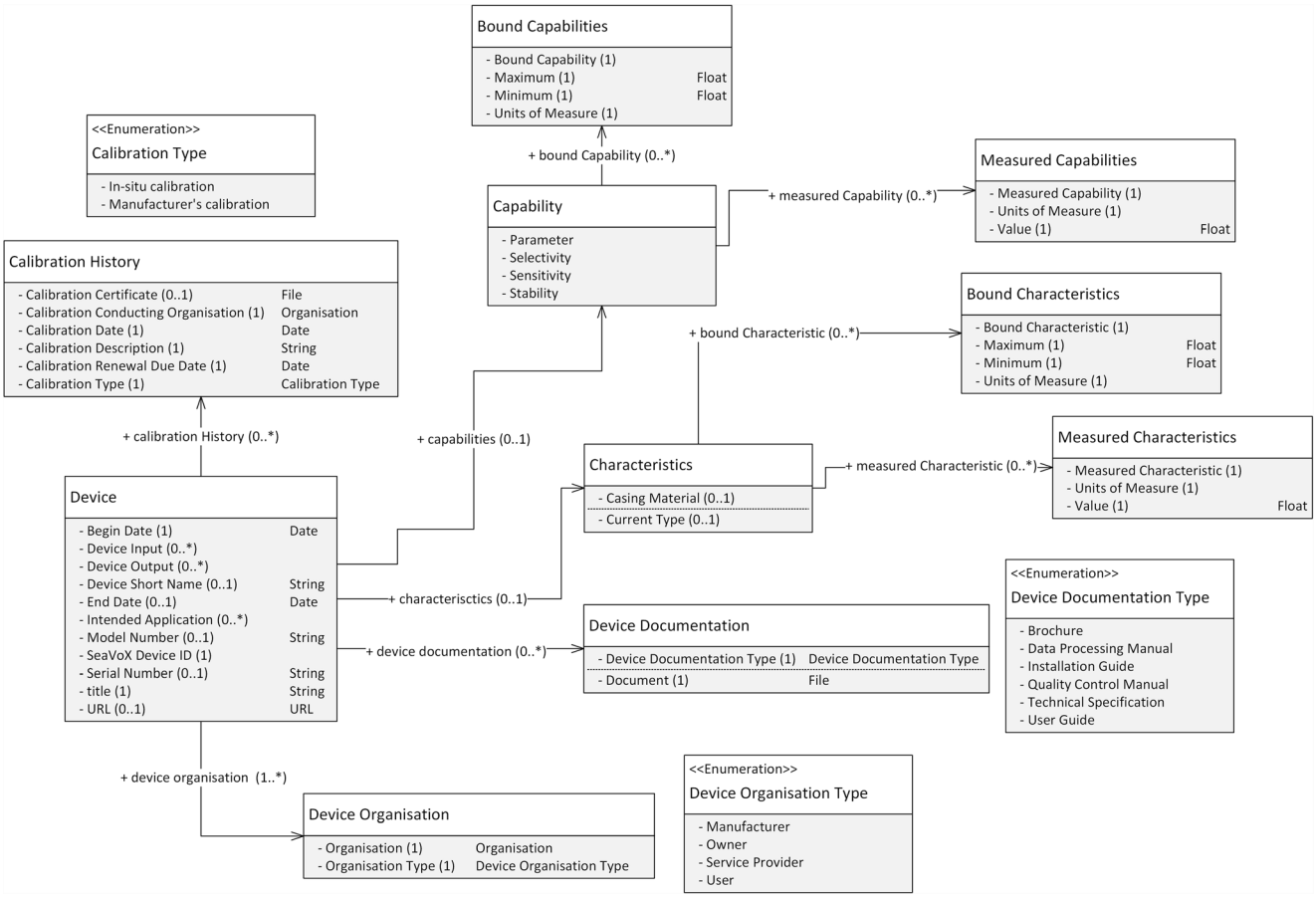

Fig. 4 A more detailed UML view of the Device class

one way or another engaged in oceanographic and marine research activities, data & information management and/or data acquisition activities. It is synonymous with the GeoLink Organisation class, but is more detailed in its attribution. Attributes include the full postal address of the organisation and institutional contact details (email, telephone, fax number, web site) which are used instead of personal contact details in any publically available metadata in order to comply with the European General Data Protection Regulation. A link to the page where the information was collected from is maintained. Where an organisation has an entry in the European Directory of Marine Organisations (Schaap & Lowry, 2010) the unique identifier from that directory is also assigned to the Organisation record here.

Re-use of community-managed controlled vocabulary terms

Many attributes of the classes in the data model are constrained against well-managed, community governed controlled vocabularies, which addresses one of the Interoperability aspects of the FAIR principles. These are highlighted in Table 3. Controlled vocabularies provide consistency in the labelling of metadata and, when published online, allow for interoperability through accessing labels and definitions through web services (Schaap and Lowry, 2010). Controlled vocabularies which have a hierarchy of terms published, that is a "thesaurus" (McGuinness, 2002), allow the more coarse grained terminology which is often used as a data discovery vector to be inferred from fine grained terminology which is important in usage metadata (see Fig. 5). Rather than storing a local copy of the full hierarchy of the vocabulary terms, the data catalogue solution presented here only tags its entities with the finest-grained vocabulary terms, and when coarser-grained terms are required to be attributed to the dataset for discovery purposes, these are inferred from queries to web services at the vocabulary service host organisations. Listing 1 shows an example SPARQL (the query language for semantic databases) query which builds up the hierarchy for a parameter usage vocabulary term which is illustrated in Fig. 5.

Implementation

In order to implement the data model described above, the architecture described below and illustrated in Fig. 6 has been adopted.

Table 3 The use of community governed controlled vocabularies to constrain varies properties within the data model

Class	Attribute	Controlled Vocabulary
Dataset / Dataset Collection	Parameter - > Observed / Modelled	British Oceanographic Data Centre (BODC) Parameter Usage Vocabulary (http://vocab.nerc.ac.uk/collection/P01)
Device	Device Input	BODC Parameter Usage Vocabulary
Device	Device Output	BODC Parameter Usage Vocabulary
Device	SeaVoX Device ID	SeaDataNet and MarineXML Vocabulary Content Governance Group (SeaVoX) Device Catalogue (http://vocab.nerc.ac.uk/collection/L22)
Device Bound Capabilities	Units of measure	BODC Data Storage Units (http://vocab.nerc.ac.uk/collection/P06)
Device Bound Characteristics	Units of measure	BODC Data Storage Units
Device Measured Capabilities	Units of measure	BODC Data Storage Units
Device Measured Characteristics	Units of measure	BODC Data Storage Units
Geographic Feature	Coordinate Reference System	International Association of Oil & Gas Producers EPSG Geodetic Parameter Dataset (http://www.opengis.net/def/crs/EPSG/0)
Geographic Feature	INSPIRE Feature Type	European Commission Joint Research Council (JRC) INSPIRE Feature Catalogue
Geographic Feature	Feature Type	ISO / Open Geospatial Consortium Simple Features Access instantiable classes
Organisation	Country	Two-letter country codes defined in ISO 3166-1
Organisation	SeaDataNet EDMO	SeaDataNet European Directory of Marine Organisations (EDMO) codes
Platform	International Council for the Exploration of the Seas (ICES) Platform Code	ICES Platform Code (http://vocab.nerc.ac.uk/collection/C17)
Platform	Measurement Regime	JRC INSPIRE Environmental Monitoring Facilities (EMF) Measurement Regime
Platform	Platform type	SeaVox Platform Category (http://vocab.nerc.ac.uk/collection/L06)
Platform	Result Acquisition Source	JRC INSPIRE Result Acquisition Source
Platform	Specialised EMF Type	JRC INSPIRE Specialised EMF Type

The first component (component 1 in Fig. 6) is an internal repository of metadata, developed using the Drupal content management system. Drupal is an open-source, community based framework enabling rapid development of web applications and is particularly suited to content management systems such as the Data Catalogue. The flexible native content management ability of a framework such as Drupal was key to the decision to input metadata in it rather than in a more familiar data cataloguing platform such as CKAN or GeoNetwork. In addition to core Drupal functionality provided 'out-of-the-box', the Data Catalogue also makes use of extended functionality through the inclusion of contributed software modules which can be managed within the Drupal framework. It is also possible to develop new modules to provide custom functionality that may not be available as core or contributed modules. It should be noted that:

- This repository is designed as an internal intranet portal only and not for general public access. A subset of relevant and appropriately classified data descriptions as defined by the actionable data policy are shared externally, only after criteria for external publication have been met
- The Data Catalogue implements role based access control allowing user access to be appropriately managed, e.g. limit create/update privileges to data owners and administrators.
- The Data Catalogue is available in read-only mode to any users already authenticated on the internal network. In this case a restricted view is provided, ensuring that any restricted access information is hidden.

The Data Catalogue has been developed to export metadata for datasets and services in ISO 19115/19139 based XML format in compliance with the INSPIRE implementing rules for metadata (component 2 of Fig. 6). This allows dataset descriptions and associated information (e.g. owning organisation, programme, sensor information etc.) to be published

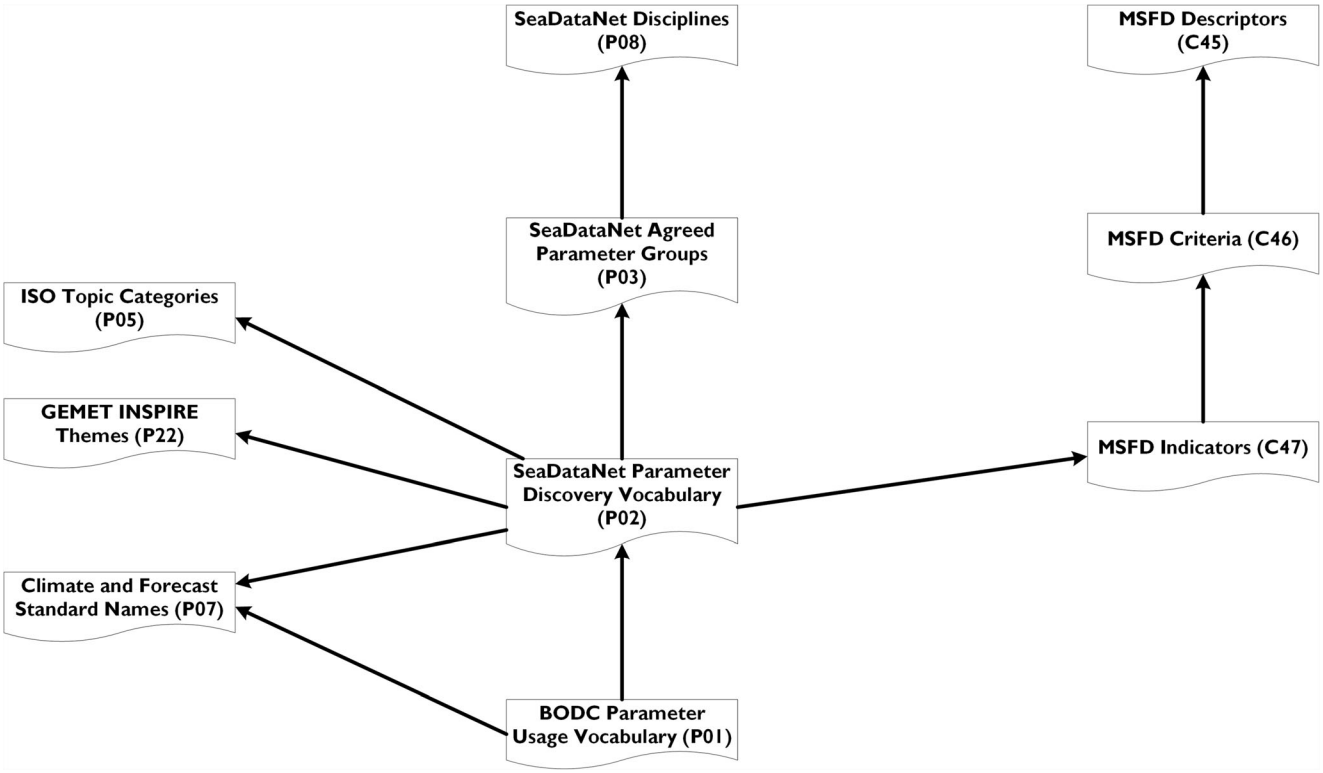

Fig. 5 Hierarchy of inferred vocabulary terms as a result of tagging a Dataset with a term from the British Oceanographic Data Centre (BODC) Parameter Usage Vocabulary. The codes in brackets - e.g. P01, P02 - indicate the collection identifier from the NERC Vocabulary Server, such that http://vocab.nerc.ac.uk/collection/P01/current/ returns the BODC Parameter Usage Vocabulary. MSFD indicates the European Commission's Marine Strategy Framework Directive; ISO indicates the International Organisation for Standardisation; GEMET indicates the European Environment Agency's General Multilingual Environmental Thesaurus; and INSPIRE is the European Commission's Spatial Data Infrastructure

and/or harvested from the Data Catalogue using industry standard formats and metadata rules. In addition, the internal Catalogue supports the DataCite metadata schema, allowing a completed data description entry to be exported in support of the minting of Digital Object Identifiers (DOI) for published data. The assignment of a DOI to a dataset is a well-documented paradigm to allow data to be cited within the scientific literature, and for a data centre or data publishing organisation to assert that an assessment of the technical quality (metadata, data format) of the dataset has been passed allowing the data to be maintained and served for the foreseeable future (Callaghan et al., 2012). This is the only place in the Data Catalogue system where individual's names are made publically available alongside the dataset, as dataset authors for the citation. As can be seen in Fig. 2, this is not the only place where individual's names are stored in the Data Catalogue system. In these other occurrences of an individual's name only organisational-level contact information (such as an info@example.org email address) and contact points based on an individual's organisational affiliation are made available to public view. The operational procedure which has been established around this is to obtain explicit consent from all dataset authors for this prior to the publication step in order to comply with the European General Data Protection Regulation. When a DOI is assigned to an entity in the Data Catalogue, it is recommended best practice to create and store the shortened form of the DOI from the ShortDOI.org service at the time the DOI is minted.

A subset of the content maintained within the internal Data Catalogue is shared externally. This publication process has been developed to make use of the standard metadata export functionality (XML formatted files) and the external facing GeoNetwork instance (component 3 of Fig. 6). GeoNetwork is an open source catalogue application to manage spatially referenced resources. It provides powerful metadata editing and search functions as well as an interactive web map viewer. It is currently used in numerous Spatial Data Infrastructure initiatives worldwide. A custom implementation of GeoNetwork has been developed to serve as the external/public facing web portal for the Data Catalogue. A number of steps are involved in the publication process, which are described below and illustrated in Fig. 7. Content is regularly exported from the internal data catalogue in ISO 19139 XML. This process can be configured to run as a background task or be manually initiated if updates are required immediately. Publication criteria and rules are applied through the

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?url ?label ?date WHERE
{
      {
            SELECT (?a AS ?url) (?c AS ?label)
            WHERE {

      BIND(<http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/> AS ?a)
                  ?a skos:prefLabel ?c
            }
      }
      UNION
      {
      SELECT(?d AS ?url) (?e AS ?label)
      WHERE {
            BIND(<http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/>
AS ?a)
                  ?a ?b ?d.
                  ?d skos:prefLabel ?e.
                  FILTER regex(str(?d),"P02|P07|S26|L05|L06")
            }
      }
      UNION {
            SELECT(?g AS ?url) (?h AS ?label)
            WHERE {

      BIND(<http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/> AS ?a)
                  ?a ?b ?d.
                  ?d ?f ?g.
                  ?g skos:prefLabel ?h.
                  FILTER regex(str(?d),"P02").
                  FILTER regex(str(?g),"P05|P22|P03|C47").
                  FILTER (lang(?h) = 'en')
            }
      }
      UNION {
            SELECT(?g AS ?url) (?h AS ?label)
            WHERE {

      BIND(<http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/> AS ?a)
                  ?a ?b ?d.
```

Listing 1 An example SPARQL query to be issued against the NERC Vocabulary Server to build the hierarchy shown in Fig. 5 for the code which represents "Practical salinity of the water body by CTD and computation using UNESCO 1983 algorithm and NO calibration against independent measurements" with the URL http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/

```
                    ?d ?f ?g.
                    ?g skos:prefLabel ?h.
                    FILTER regex(str(?d),"S26").
                    FILTER regex(str(?g),"S21").
                    FILTER (lang(?h) = 'en')
              }
  }
  UNION {
              SELECT(?j AS ?url) (?k AS ?label)
              WHERE {

        BIND(<http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/> AS ?a)
                    ?a ?b ?d.
                    ?d ?f ?g.
                    ?g ?i ?j.
                    ?j skos:prefLabel ?k.
                    FILTER regex(str(?d),"P02").
                    FILTER regex(str(?g),"C47").
                    FILTER regex(str(?j),"C46").
              }
        }
  UNION {
              SELECT(?m AS ?url) (?n AS ?label)
              WHERE {

        BIND(<http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/> AS ?a)
                    ?a ?b ?d.
                    ?d ?f ?g.
                    ?g ?i ?j.
                    ?j ?l ?m.
                    ?m skos:prefLabel ?n.
                    FILTER regex(str(?d),"P02").
                    FILTER regex(str(?g),"C47").
                    FILTER regex(str(?j),"C46").
                    FILTER regex(str(?m),"C45").
              }
        }

  UNION {
              SELECT(?j AS ?url) (?k AS ?label)
              WHERE {

        BIND(<http://vocab.nerc.ac.uk/collection/P01/current/PSALCU01/> AS ?a)
```

Listing 1 (continued)

```
                    ?a ?b ?d.
                    ?d ?f ?g.
                    ?g ?i ?j.
                    ?j skos:prefLabel ?k.
                    FILTER regex(str(?d),"P02").
                    FILTER regex(str(?g),"P03").
                    FILTER regex(str(?j),"P08").
              }
        }
  UNION {
        Select (?a AS ?url) (?c AS ?label) (?d AS ?date)
              WHERE{
                    ?a ?b skos:Collection.
                    ?a skos:prefLabel ?c.
                    ?a dct:date ?d.
                    FILTER
  regex(str(?a),"C45|C46|C47|L05|L06|L22|P01|P02|P03|P05|P08|P22|S21|S26")
              }
        }
  } ORDER BY ?url
```

Listing 1 (continued)

machine-actionable data policy to ensure that only content appropriate for publication is included in the export process. These rules are based on data classification, publication status, licensing etc. and can be updated as required. Once exported, metadata XML files are moved to a central staging area located on the external perimeter network or 'demilitarized zone' (DMZ). This serves as the collection point for the GeoNetwork instance. The GeoNetwork instance includes an automated and configurable harvest capability. This allows the previously exported data descriptions to be imported and published on the public facing portal.

While not a core component of the Data Catalogue, the external facing Catalogue, through GeoNetwork, supports integration with other data serving applications (components 4 and 5 of Fig. 6). This allows Data Catalogue users to download or link to the underlying data as described by in the Data Catalogue. For spatial data this is achieved via Open Geospatial Consortium compliant web services from a GeoServer instance. GeoServer implements a number of standards such as Web Feature Services, Web Map Services, and Web Coverage Services. Another important data serving application is ERDDAP; a data server that gives a simple,

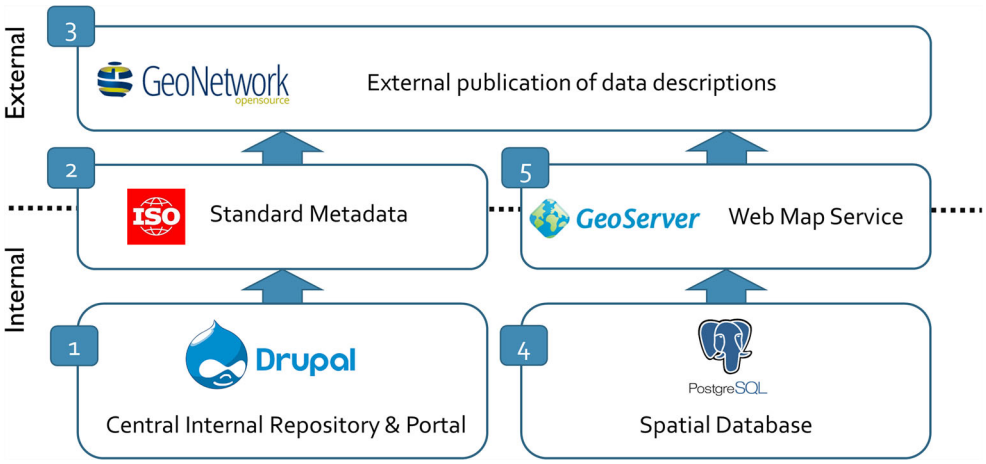

Fig. 6 A high-level view of the adopted system architecture. The component numbers are identified in the main body of the text

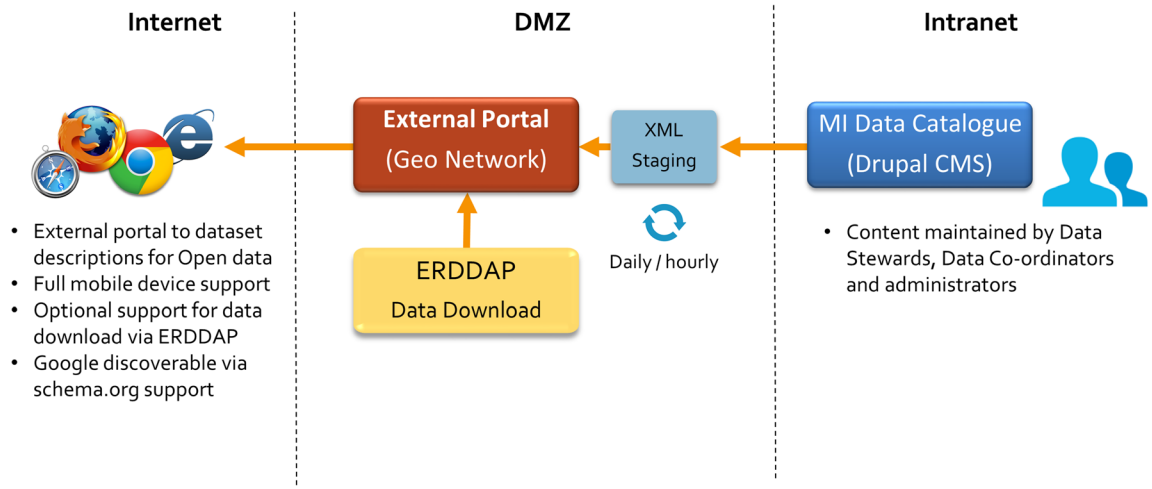

Fig. 7 The metadata publication process from internal data catalogue to external GeoNetwork instance. The Data Steward and Data Coordinator Roles are described in Leadbetter et al. (2019)

consistent way to download subsets of scientific datasets in common file formats and make graphs and maps (Simons, 2019). ERDDAP has been developed by the National Oceanographic and Atmospheric Administration in the United States to provide access to data stored in multiple different formats through a web interface, and using RESTful URLs through a web service, brokering the storage formats to a number of data delivery formats. ERDDAP is a useful tool in scientific data delivery, not just for marine science, as it can access and serve any tabular or gridded data. These data integration components are included here to provide a complete, unified solution view of metadata and data delivery.

This approach provides a clear decoupling and separation of potentially sensitive internal data descriptions and

Table 4 How the data cataloguing platform described in this paper addresses the requirements of the FAIR principles of Data Management

FAIR Principle	Approach adopted in this Data Catalogue
F1. (Meta)data are assigned a globally unique and persistent identifier	The catalogue defines a unique identifier to the metadata records, of the form ie.marine.data:dataset.2740. This may be supplemented with a Digital Object Identifier, such as doi.org/csgf for the example given above.
F2. Data are described with rich metadata	The Data Catalogue model in Figs. 1-5 presents a rich metadata model for data description.
F3. Metadata clearly and explicitly include the identifier of the data they describe	The metadata records clearly show the Data Catalogue identifier and DOI and
F4. (Meta)data are registered or indexed in a searchable resource	Both the internally facing Drupal instance and the externally facing GeoNetwork instance are fully searchable metadata repositories.
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol	The identifier from Principle *F1* can be simply appended to the GeoNetwork service URL to retrieve the metadata record over HTTP.
A2. Metadata are accessible, even when the data are no longer available	Data Catalogue records remain visible when deprecated or superceded, with a note and link to replacement records appended to the original record.
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	The ISO19115 metadata standard is used for the metadata, and it is also made available as Schema.org via JSON-LD and via RDF using the DCAT vocabulary.
I2. (Meta)data use vocabularies that follow FAIR principles	The use of controlled vocabularies in the Data Catalogue metadata records is highlighted in Table 3. Where vocabulary terms are addressable by HTTP URLs, these are included within the metadata records.
I3. (Meta)data include qualified references to other (meta)data	References to other metadata are available through the catalogue, but future work should focus on making the qualifications more verbose.
R1. Meta(data) are richly described with a plurality of accurate and relevant attributes	Data licenses are made clear to users as they are assigned to metadata records through the Classification. Through analysing the various domain requirements and working with international data networks, the metadata meet best practices in marine science data management. Where possible, detailed provenance is shown in a metadata record, but for older datasets this can be an issue as not all the details may have been recorded.

```
{
   "@context":"http://schema.org",
   "@type":"Dataset",
   "name":"Lough Furnace Automatic Water Quality Monitoring Station (AWQMS)
profiles 2009-2014",
   "description":"Full water column profiles of temperature, conductivity,
pressure and dissolved oxygen are routinely measured in the coastal basin
Lough Furnace as part of the LTER (long-term ecological research) monitoring
programme. Profiles are measured by a multi-parameter sonde attached to an
automated undulating winch that initiates downcasts at 4 daily intervals (00,
06, 12, 18 hours). This dataset includes profiles recorded during the period
2009-2014. Analysis of this dataset can be found here: Kelly, S., Eyto, E.
de, Dillane, M., Poole, R., Brett, G., and White, M. (2018). Hydrographic
maintenance of deep anoxia in a tidally influenced saline lagoon. Marine and
Freshwater Research 69(3) 432-445 https://doi.org/10.1071/MF17199",
   "url":"http://data.marine.ie/geonetwork/srv/eng/catalog.
      search#/metadata/ie.marine.data:dataset.2752",
   "temporalCoverage":"2009-02-01T00:00:00/2014-10-29T00:00:00",
   "keywords":[
      "Chemical oceanography",
      "Physical oceanography",
      "Dissolved gases",
      "Other physical oceanographic measurements",
      "Water column temperature and salinity"
   ],
   "variablesMeasured":[
      "Concentration of oxygen {O2 CAS 7782-44-7} per unit volume of the
water body [dissolved plus reactive particulate phase]",
      "Date and time",
      "Density (potential) of the water body by computation from salinity and
potential temperature using UNESCO algorithm with 0 decibar reference
pressure",
      "Depth below surface of the water body",
      "Electrical conductivity of the water body",
      "Mass Concentration Oxygen in Sea Water",
      "Practical salinity of the water body by computation using UNESCO 1983
algorithm",
      "Sea Water Density",
      "Temperature of the water body"
   ],
   "creator":{
      "@type":"Organization",
      "url":"http://www.marine.ie",
```


Listing 2 A Schema.org representation of a dataset from within this data catalogue model. The original record is available at http://data.marine.ie/geonetwork/srv/eng/catalog.search#/metadata/ie.marine.data:dataset.2752

```
      "name":"Marine Institute"
    },
    "license":[
      "Marine Institute Licence"
    ],
    "spatialCoverage":{
      "@type":"Place",
      "geo":{
        "@type":"GeoShape",
        "polygon":"53.921815982253,-9.5740141040776 53.92242251448,-
9.5723833209966 53.922877407867,-9.5709241992924 53.922725777289,-
9.5677484638188 53.922371970464,-9.5665468341801 53.92156325789,-
9.5668043262456 53.921057804576,-9.5658601886723 53.920046879589,-
9.5646585590337 53.919288669784,-9.5665468341801 53.917974406838,-
9.5668901569341 53.916862305878,-9.5678342945073 53.91630624429,-
9.5686926013921 53.915547966555,-9.5689500934575 53.91478967505,-
9.5682634479497 53.914840228245,-9.5668043262456 53.914334693537,-
9.5646585590337 53.913728043809,-9.5613969928716 53.9139808156,-
9.5584787494633 53.914536908155,-9.5570196277592 53.91377859829,-
9.5577921039555 53.912919163795,-9.5583929187749 53.912110268114,-
9.561654484937 53.91140247154,-9.5613969928716 53.910593546481,-
9.5597662097905 53.910947453123,-9.5584787494633 53.911250799285,-
9.5573629505132 53.910391312769,-9.5577062732671 53.910189078077,-
9.5590795642827 53.909734046439,-9.5593370563481 53.90862172601,-
9.5577062732671 53.908470043656,-9.5591653949712 53.908722847273,-
9.5604528552983 53.908571165286,-9.5622552997563 53.90932956971,-
9.5637144214604 53.910239636841,-9.5637144214604 53.910846337245,-
9.5636285907719 53.910896895215,-9.5650018817876 53.910542988145,-
9.5678342945073 53.910087960363,-9.5689500934575 53.909582368125,-
9.5684351093266 53.908925089065,-9.5689500934575 53.908470043656,-
9.570838368604 53.908520604502,-9.5722116596196 53.909885724202,-
9.5714391834233 53.911250799285,-9.570838368604 53.912009155052,-
9.5698084003423 53.912767497048,-9.5709241992924 53.912565273863,-
9.5722116596196 53.912767497048,-9.5734991199467 53.91342471564,-
9.5725549823735 53.914536908155,-9.5734132892583 53.914536908155,-
9.5754732257817 53.914031369775,-9.5765890247319 53.914536908155,-
9.5774473316167 53.91529520425,-9.576074040601 53.916407346948,-
9.5758165485356 53.917064508256,-9.576245701978 53.916407346948,-
9.5783056385014 53.915547966555,-9.5804514057133 53.915497414217,-
9.583627141187 53.91625569287,-9.5824255115483 53.91635679565,-
9.5805372364018 53.917013957753,-9.5783056385014 53.919743597319,-
9.578048146436 53.920602891361,-9.5764173633549 53.921815982253,-
9.5740141040776"
```


Listing 2 (continued)

```
        }
    },
    "distribution":[
        {
            "@type":"DataDownload",
            "encodingFormat":"CSV",
            "contentUrl":"http://data.marine.ie/data/657328d9-bbed-4427-ac73-
85a2dae184b2.zip"
        }
    ]
}
```


Listing 2 (continued)

externally published metadata. Only information explicitly categorised as suitable for publication is persisted on the external system. The external facing portal publishes a 'read-only' copy of the metadata contained within the Data Catalogue and mitigates any data loss if the external system has been compromised. Core user accounts and system provisioning details are maintained on the internal Drupal system, residing on a controlled and secure network. The publication criteria can be updated and modified at any time if requirements or user needs change. It makes use of industry standard metadata and provides an excellent reference system for other implementers that may be interested in using metadata in a similar way, for example Ireland's Open Data Portal (https://data.gov.ie/).

Conclusions

We have presented a reusable, modular approach to cataloguing marine science data which meets a number of functional requirements derived from both academic literature and legislative drivers. The Data Catalogue system presented above also meets, at a base level, the requirements of the FAIR principles of data management (see Table 4). One particular development of note in the "Findability" principle is that the data model is presented within the HTML representations of the metadata landing pages using JSON-LD encoded Schema.org (see Listing 2). This improves the discoverability of the content of the Data Catalogue through exposing it to tools such as Google's Dataset Search.

Although Table 4 shows a good alignment of the work presented above with the FAIR principles, there remains work to complete on the formalised representation of the data in structured formats beyond Schema.org and in the provenance of the datasets described in the Data Catalogue. Firstly, although GeoNetwork supports a generic Resource Description Framework (Miller, 1998) description of metadata records using the Data Catalog vocabulary (Maali et al., 2014) this requires extension to add in specific terms from domain specific ontologies such as GeoLink. This should also allow for more formalised descriptions of linkages between various datasets using richer semantics to describe the connections. Better connectivity between datasets and reports which use them is also required in the future. A further semantic application would be the use of spatial semantics to provide textual geographic search, which requires extensions to the existing structured thesauri describing geographic regions of the sea, such as the SeaVoX salt and freshwater body gazetteer (http://vocab.nerc.ac.uk/collection/C19).

Marine science programmes often collect biological samples in combination with environmental data. A collection of physical samples is analogous to a Dataset, with added complexity due to the samples tangibility. For example, a Dataset Collection Activity in the form of a marine research vessel survey may have a primary goal to measure stock abundance of a specific fishery (i.e. haddock or cod). In this example, a large part of the survey will involve taking biological samples (in the form of fish otoliths) for aging the population to report on stock recruitment. The resulting age dataset will be used to inform policy advice on regulatory measures regarding fishing effort in succeeding years (Marine Institute, 2018). The biological samples (in this case, the otoliths), and the associated fish metadata, are often stored for an extended period of time after the Dataset Collection Activity, for scientific reproducibility and transparency of the age dataset generated. In addition, otoliths can be used for microchemical analyses to investigate fish diet and habitat (Campana & Thorrold, 2001), which can be valuable for fisheries conservation efforts in subsequent years. Therefore, the necessity for appropriate physical and digital storage of biological samples and their associated metadata is evident. We anticipate the development of an optional accessory extension to the Data Catalogue to model biological samples and their associated metadata. The extension will utilize select concepts from the Data Catalogue, such as Geographic Feature and Programme, but also include additional metadata. For example, in the fisheries use-case,

phenotype data (i.e. fish length and weight) will be associated with each biological sample (i.e. otolith). We expect the physical sample's extension to the Data Catalogue to become a useful tool for long term archiving and reusability of physical samples resulting from various marine science programmes.

Finally, there is ongoing work in the data catalogue beyond the FAIR principles, as these offer a base level of good data stewardship (Boeckhout et al., 2018). One example is to automate assessments of the maturity of the stewardship of datasets within the Data Catalogue system. This takes the Data Stewardship Maturity Framework of Peng et al. (2015) as its starting point and will assess the values encoded for various elements in the Data Catalogue's data model to produce a rating for a given dataset. As discussed by Flynn et al. (2019), this approach can also be specialised in order to provide an assessment of the suitability of a dataset for a given application, in the case of their study for Marine Spatial Planning.

Acknowledgments The authors wish to thank Mr. Dirk Fleischer of Christian-Albrechts University Kiel for conversations which informed the concept of the machine actionable data policy and Mr. Trevor Alcorn at the Geological Survey Ireland, and previously at the Marine Institute, for his early contributions to the project.

A number of projects and programmes have contributed to the development of this work.

This work is part supported by the Irish Government and the European Maritime & Fisheries Fund as part of the EMFF Operational Programme for 2014-2020.

This work is part supported by the Marine Institute's Digital Ocean programme.

This work was part carried out under the COMPASS project. This project is supported by the European Union's INTERREG Iva Programme, managed by the Special EU Programmes Body (SEUPB). The views and opinions expressed in this document do not necessarily reflect those of the European Commission or the Special EU Programmes Body (SEUPB).

The 'Unlocking the Archive' project (Grant-Aid Agreement No. PBA/FS/16/03) is carried out with the support of the Marine Institute and is funded under the Marine Research Programme by the Irish Government.

References

Boeckhout M, Zielhuis ZA, Bredenoord AL (2018) The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet 26:931–936

Botts M, Robin A (2007) OpenGIS Sensor Model Language (SensorML) implementation specification. Open Geospatial Consortium

Callaghan S, Donegan S, Pepler S, Thorley M, Cunningham N, Kirsch P et al (2012) Making data a first class scientific output: data citation and publication by NERC's environmental data centres. Int J Digit Curation 7(1):107–113

Campana SE, Thorrold SR (2001) Otoliths, increments, and elements: keys to a comprehensive understanding of fish populations? Can J Fish Aquat Sci 58(1):30–38. https://doi.org/10.1139/f00-177

Craglia M, Annoni A (2007) INSPIRE: an innovative approach to the development of spatial data infrastructures in Europe. Research and Theory in Advancing Spatial Data Infrastructure Concepts, 93–105

Ehler C, Zaucha J, Gee K (2019) Maritime/Marine Spatial Planning at the interface of research and practice. In *Maritime Spatial Planning* (pp. 1–21). Palgrave Macmillan, Cham

European Parliament (2003). *Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC*. Retrieved 7th January 2020 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32003L0004

European Parliament (2008). *Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive) (Text with EEA relevance)*. Retrieved 7th August 2019 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008L0056

European Parliament (2014). *Directive 2014/89/EU of the European Parliament and of the Council of 23 July 2014 establishing a framework for maritime spatial planning*. Retrieved 7th August 2019 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.257.01.0135.01.ENG%20

Flynn S, Meaney W, Leadbetter A, Fisher J, Nic Aonghusa C (2019) *A Data Management and Storage Process for Marine Spatial Planning in Ireland*. Manuscript submitted for publication

Friddell JE, LeDrew EF, Vincent WF (2014) The Polar Data Catalogue: best practices for sharing and archiving Canada's polar data. Data Science Journal: IFPDA-01

Günther O, Lessing H, Swoboda W (1996) UDK: A European environmental data catalogue. In *Proceedings of the Third International Conference in Integrating GIS and Environmental Modeling*. National Center for Geographic Information and Analysis, Santa Barbara (USA)

Herring J (2011) OpenGIS implementation standard for geographic information-simple feature access-part 1: common architecture. Open Geospatial Consortium

Hilbring D, Usländer T (2006, September) Catalogue services enabling syntactical and semantic interoperability in environmental risk management architectures. In *EnviroInfo* (pp. 39-46)

INSPIRE Thematic Working Group Environmental Monitoring Facilities (2013) *D2.8.II/III.7 INSPIRE Data Specification on Environmental Monitoring Facilities – Technical Guidelines*. European Commission Joint Research Centre

Krisnadhi A, Hu Y, Janowicz K, Hitzler P, Arko R, Carbotte S, ... Ji P (2015, October) The GeoLink modular oceanography ontology. In *International Semantic Web Conference* (pp. 301–309). Springer, Cham

Leadbetter AM, Lowry RK, Clements DO (2014) Putting meaning into NETMAR–the open service network for marine environmental data. International Journal of Digital Earth 7(10):811–828

Leadbetter AM, Vodden PN (2016) Semantic linking of complex properties, monitoring processes and facilities in web-based representations of the environment. International Journal of Digital Earth 9(3):300–324

Leadbetter AM, Carr R, Flynn S, Meaney W, Moran S, Bogan Y, Brophy L, Lyons K, Stokes D, Thomas R (2019) Implementation of a data management quality management framework at the Marine

Institute. Ireland *Earth Science Informatics*:1–13. https://doi.org/10.1007/s12145-019-00432-w

Maali F, Erickson J, Archer P (2014) *Data catalog vocabulary (DCAT)*. W3C recommendation, 16

Marine Institute (2017) Marine Institute Data Policy. Marine Institute, Galway, Ireland. https://www.marine.ie/Home/sites/default/files/MIFiles/Docs/DataServices/Marine%20Institute%20Data%20Policy%202017.pdf

Marine Institute (2018) The Stock Book 2018: annual review of fish stocks in 2018 with management advice for 2019. Marine Institute, Galway

Miller E (1998) An introduction to the Resource Description Framework. Bull Am Soc Inf Sci Technol 25(1):15–19

Nylén T, Tolvanen H, Erkkilä-Välimäki A, Roose M (2019) Guide for cross-border spatial data analysis in maritime spatial planning. University of Turku, Turku

McGuinness DL (2002) Ontologies come of age. Spinning the semantic web: bringing the World Wide Web to its full potential, 171–194

Peng G, Privette JL, Kearns EJ, Ritchey NA, Ansari S (2015) A unified framework for measuring stewardship practices applied to digital environmental datasets. Data Science Journal 13:231–253

Proctor R, Roberts K, Ward BJ (2010) A data delivery system for IMOS, the Australian Integrated Marine Observing System. Adv Geosci 28:11–16

Schaap DMA, Lowry RK (2010) SeaDataNet–pan-European infrastructure for marine and ocean data management: unified access to distributed data sets. International Journal of Digital Earth 3(S1):50–69

Simons RA 2019 *ERDDAP*. https://coastwatch.pfeg.noaa.gov/erddap. Monterey, CA: NOAA/NMFS/SWFSC/ERD

Tanhua T, Pouliquen S, Hausman J, O'Brien K, Bricher P, de Bruin T, Buck JJH, Burger EF, Carval T, Casey KS, Diggs S, Giorgetti A, Glaves H, Harscoat V, Kinkade D, Muelbert JH, Novellino A, Pfeil B, Pulsifer PL, Van de Putte A, Robinson E, Schaap D, Smirnov A, Smith N, Snowden D, Spears T, Stall S, Tacoma M, Thijsse P, Tronstad S, Vandenberghe T, Wengren M, Wyborn L, Zhao Z (2019) Ocean FAIR data services. Front Mar Sci 6:440. https://doi.org/10.3389/fmars.2019.00440

Vitolo C, Elkhatib Y, Reusser D, Macleod CJ, Buytaert W (2015) Web technologies for environmental big data. Environ Model Softw 63:185–198

Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR). A Practical Guide, 1st Ed.*, Cham: Springer International Publishing

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N et al. (2016) "The FAIR guiding principles for scientific data management and stewardship." Scientific Data 3