2018

Interoperable Ocean Observing Using Archetypes: A Use-case Based Evaluation

Paul Stacey
Institute of Technology, Blanchardstown, paul.stacey@tudublin.ie

Damon Berry
damon.berry@tudublin.ie

Recommended Citation

P. Stacey and D. Berry, "Interoperable Ocean Observing using Archetypes: A use-case based evaluation, *MTS/IEEE Oceans conference, 2018, Charleston, South Carolina, USA, 22-25th. October 2018*. doi: 10.1109/OCEANS.2018.8604834

Interoperable Ocean Observing using Archetypes: A use-case based evaluation

Paul Stacey
School of Informatics & Engineering
Institute of Technology Blanchardstown
Dublin, Ireland
paul.stacey@itb.ie

Damon Berry
School of Electrical and Electronic Engineering
Dublin Institute of Technology
Dublin, Ireland
damon.berry@dit.ie

***Abstract*-This paper presents a use-case based evaluation of the impact of two-level modeling on the automatic federation of ocean observational data. The goal of the work is to increase the interoperability and data quality of aggregated ocean observations to support convenient discovery and consumption by applications. An assessment of the interoperability of served data flows from publicly available ocean observing spatial data infrastructures was performed. Barriers to consumption of existing standards-compliant ocean-observing data streams were examined, including the impact of adherence to agreed data standards. Historical data flows were mapped to a set of archetypes and a backward integration experiment was performed to assess the incremental benefit of using two level models to federate data streams. The outcome of the evaluation demonstrates the feasibility of building a two-level model based ocean observing system using a combination of existing open source components, the adaptation of existing standards and the development of new software tools. The automatic integration of data flows becomes possible. This technique also allows real-time applications to automatically discover and federate newly discovered data flows and observations.**

Keywords—two-level modeling, interoperability, O&M, archetypes

I. INTRODUCTION

In order for ocean observations to be re-useable beyond their original purpose, they must contain rich contextual meta-data. This is not always the case. The Berlin Declaration on Open Access to Knowledge in Sciences and Humanities [1] seeks to promote the Internet and Web as a functional instrument to promote and advance human knowledge. Open access to data and knowledge can also act as a key economic driver. Pooling existing resources can save significant amounts of public money. The European Union green paper on Marine knowledge 2020 strategy [2] estimates that a shared marine data infrastructure consisting of high quality marine data collected by EU public bodies could save €1Billion per year. There are many barriers to building such marine data infrastructures; discoverable and interoperable data are the focus of much research [3].

Today the European Commission is advancing the goal of access to open data in a transparent way. This goal has prompted several initiatives such as INSPIRE [4], EMODnet [5], SeaDataNet [6], JericoNEXT [7] and AtlantOS [8]. These initiatives subsequently advance a complimentary international goal of interoperable and open ocean data. For example, SeaDataNet contributes to the Ocean Data Interoperability Platform (ODIP) [9]. ODIP brings together all the key ocean data management organizations from the EU, US and Australia. ODIP in turn is promoted by IOC/IODE [10] and other international consortia to help achieve global ocean data interoperability. Through ODIP, EU projects such as INSPIRE are having a global impact. For example, the adoption of the Observations & Measurements (O&M) standard ISO/DIS 19156 [11] within INSPIRE has seen O&M become a key component of the GEO-DAB discovery and access broker [12]. GEO-DAB connects more than 150 international providers of high quality Earth Observations. The continued investment in open and interoperable ocean spatial data infrastructures (SDI) around the world is beginning to realize dividends. However, there are still many challenges to overcome.

The Columbus project [3] has performed a broad review of ocean data portals. Their work is not exhaustive but highlights the wealth of available SDIs and portals. The Columbus review is unique as its goal is to create measurable growth in the blue economy. It is also tasked with monitoring the implementation of the Marine Strategy Framework Directive (MSFD) [13]. Thus the focus is on the ability of marine spatial data infrastructures to encourage and enable end users develop value added services and products. In their analysis it was found many marine data portals are built from a developer's perspective on the intended purpose, and not the end user. Therefore ease of use and user friendliness of data sharing facilities can impede the wider sharing of collected data [3].

Recently the authors have shown how two-level informational modeling techniques may be translated to geo-observational scenarios [14] [15]. The attributes of a two-level based system of the type described in [14] are highly desirable within the ocean observing community. The promises of a two-level modeling systems design approach are in keeping with the aims of ODIP, GEOSS [16] and other ocean data interoperability initiatives. In [15] the authors propose a novel technique to enable *archetypes*, and consequently a two-level modeling approach within technologically constrained ocean observing platforms. In [14] a rudimentary technical validation of the approach is described. While the preliminary study has served to highlight the potential of two-level modeling for achieving

978-1-5386-4814-8/18/$31.00 ©2018 IEEE

interoperable ocean observations, additional robust use-case based evaluations are needed; considering the peculiarities of the current state-of-the-art ocean data sharing frameworks and spatial data infrastructures (SDI). This paper presents a use-case based evaluation of the applicability of two-level modeling within ocean based observing systems. The initial focus is on using two-level modeling to increase *interoperability* and *discoverability* of ocean observational data flows.

This paper is organized as follows. Section II gives a brief overview of two-level modeling. Section III describes the use-case used in this work. Section IV describes the tools & methods used. Section V presents the outcomes of an interoperability review of the current state-of-the-art European based spatial data infrastructures for the marine environment. Section V also presents the results of the use-case based assessment of two-level modeling for the estimation of chlorophyll-α in the southern North Sea. Section VI presents a short discussion and conclusion.

II. Two-Level Modeling

Traditionally information systems are designed using a single-level approach for modeling information. In the single-level approach, information and knowledge concepts are tightly coupled in a single data model and hard coded into the system software. This coupling happens early in the design process when data models are defined [17].

In domains such as oceanography, where data models are subject to constant evolution - as the domain knowledge itself is constantly evolving - hard coded systems soon become obsolete as they no longer represent the current domain knowledge [17]. Interoperability suffers over time as heterogeneous information systems begin to emerge, all representing different implementations of the domain data and with no clear mechanism for integration of information objects [17].

To avoid this scenario, many standards development bodies such as the Open Geospatial Consortium (OGC) [18] avoid overly constraining data models and standards such as ISO/DIS 19156. However, this results in abstract models that need to be specialized for use-cases. Developers are typically left to iron out the details themselves, resulting in many heterogeneous system implementations. Although these systems will adhere to the abstract standardized data model, the particulars of the implementations are not standardized and therefore inhibit interoperability.

A two-level modeling system design approach defines two levels, or models. The *reference model* and the *knowledge model* (Fig. 1).

- The reference model contains non-volatile concepts, or classes with an abstract meaning that are not subject to change over time. These classes are hard-coded into the system software.
- The knowledge model captures the concepts that will undergo evolution over time. This model is not hard-coded into the system software but rather processed at runtime. Concepts are captured in an Archetype Model using archetypes. Archetypes act as a problem specific constraint model on the underlying reference model.

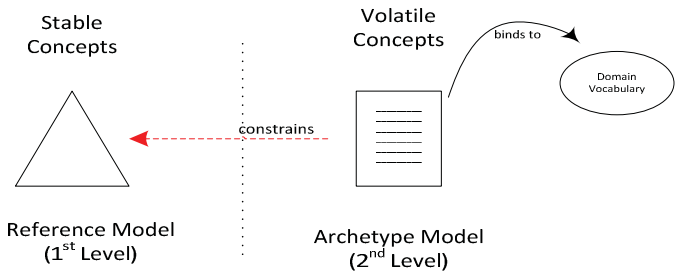

Fig. 1. The archetype model is a set of constraint statements against the underlying stable reference model. The archetype model evolves over time.

A. Archetypes

Archetypes are a set of constraint statements, normally captured using the Archetype Definition Language (ADL) [19]. Archetypes are developed by a community of supporting domain experts and may be further specialized by these communities for particular use-cases in different jurisdictions. Systems generate information instances at run-time using *operational templates* (OPT) that adhere to the archetype model and the underlying reference model. For a more thorough overview of two-level modeling techniques the reader is directed towards [15] and [17].

III. Background & Study Overview

In order to further evaluate the benefits of two-level modeling in ocean observing scenarios, a use-case study has been developed. The aim of the study is to demonstrate the automatic backward federation [20] of observational data flows, governed by the use of community agreed archetypes. Here the approach is developed to show its applicability in understanding and estimating the mechanisms governing chlorophyll-α concentrations within a defined sea region.

A. Chlorophyll-α

It is believed that anthropogenic warming of oceans is increasing the level of phytoplankton in the water column [21]. Phytoplankton are microscopic algae and are an important source of aquatic food. However, in large concentrations algae can have a detrimental effect on marine life and water quality [22]. Excessive growth can starve aqua-culture sites of dissolved oxygen and devastate fish stock. Chlorophyll-α (Chlfa) is a photosynthetic pigment and common to all phytoplankton [22]. Chlfa concentrations are used to quantify levels of phytoplankton and can be measured using in-situ sensors known as fluorometers or satellite based sensors. High levels of Chlfa can indicate an *algae bloom* and is an important indicator of eutrophication [22]. There are many drivers of excessive phytoplankton growth. Typically, there are two primary production drivers, light (irradiance) and nutrients within the body of water [22].

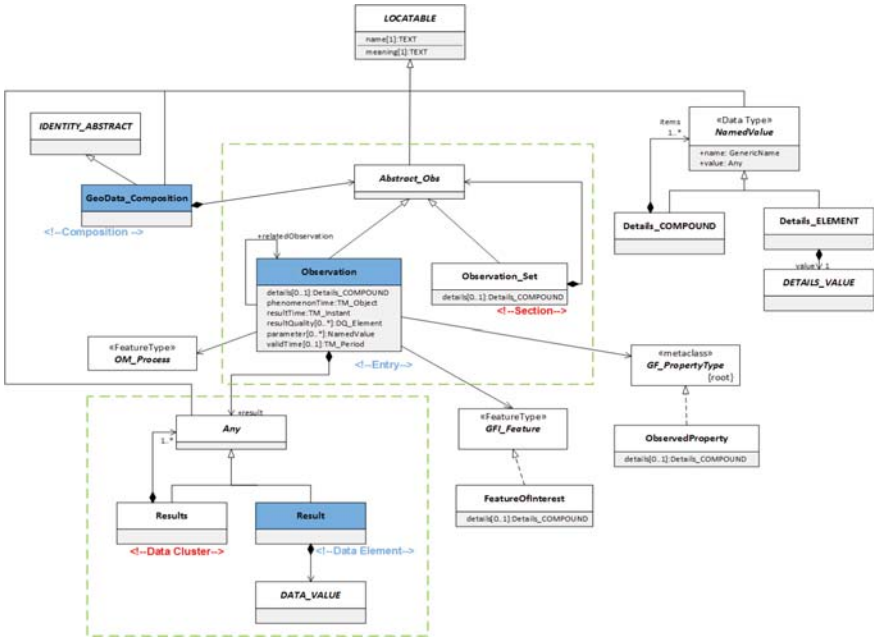

Fig. 2. An augmented Observations and Measurements model [11]. This model serves as the reference model for the two-level modeling approach. Compound/element patterns (highlighted in green) are necessary for two-level modeling.

The Development of accurate Chlfa estimation models and prediction systems for individual sea regions is an important area of research. The focus is often on developing computationally efficient estimation models, using other oceanic parameters to estimate Chlfa levels. For example, Irwin and Finkel have shown that sea-surface temperature combined with latitude/longitude, surface nitrate and irradiance can predict 83% of the log variance in chlorophyll-α in the north Atlantic sea region [23]. In [24] it was found that sea surface temperature is the best single predictor of log chlorophyll-α.

Observations are key inputs to Chlfa estimation models. Satellite based sensors are an important source of observational data but can only read at or close to the sea surface. Also, satellite data may be diminished with cloud cover, therefore in-situ monitoring stations are needed.

B. Sea Regions & Ocean Models

The North West Shelf (NWS) sea region covers a large area. Sub regions include the Irish Sea and Southern North Sea, among others. The NWS operational oceanography organization (NOOS) [25] includes nine countries that collaborate together to develop ocean observing and prediction systems for the NWS area. The NWS data portal [26] is one product arising from NOOS. NOOS is also part of the European Global Ocean Observing System (EuroGOOS) [27].

NOOS also operates in the context of the Global Ocean Observing System (GOOS) [28]. One of the core goals for GOOS and associated GOOS Regional Alliances (GRA) [29] (of which EuroGOOS is part of) is to develop advanced ocean model based products. Today there is now a wealth of ocean dynamics models available. The EuroGOOS ocean models Web tool [30] provides a convenient way to browse and filter the various ocean models that are available for the EuroGOOS area.

There are a wide range of ocean models available for the NWS area. The Dutch Continental Shelf Model (DCSM) model is a well-established hydrodynamic model developed by the Dutch government to improve accurate water-level forecasting [31]. The Nemo Nordic model [32] is a specialized model for the Baltic & North Sea, based on the well-known NEMO ocean engine. The GEM/BLOOM model developed by Deltares can be used to estimate chlorophyll-α concentrations and water quality in the North Sea [33]. Other generalized statistical models such as the Generalized Additive Model (GAM) [34] are also often used as a linear predictive model for ocean dynamics.

The Southern North Sea region was selected as the focus for the presented use-case. The use-case is motivated by the previous work performed as part of the INSPIRE Marine Pilot [35]. In this use-case salinity and temperature observations are the data flows of choice. It is reasonable to focus on salinity and temperature as they have been shown to have a strong correlation with chlorophyll-α concentrations in the NWS sea region [23] [36]. Also, typically salinity and temperature in-situ observations are more readily available within sea regions.

IV. Tools & Methods

To ensure that real observational data was used, a review of publicly available ocean observational portals was performed. Of the portals reviewed the EMODnet-Physics data portal was chosen [37]. Three ocean observing platforms were selected within the area of the southern North Sea. This area is chosen as it is composed of a number of bordering jurisdictions (UK, Netherlands, Belguim, France) whom are subject to EU INSPIRE compliance [4].

Data for a 60 day period is downloaded from each site through the EMODnet-Physics data portal. The data was retrieved in netCDF format. NetCDF data files were converted to JSON using the netCDF operator tool suite NCO toolkit [38] for ease of parsing and assessment. An assessment of data interoperability was performed using mapping tables.

An additional mapping of each dataset was then performed to produce INSPIRE (O&M) compliant data flows. A data assimilation exercise was performed using the OpenDA toolbox [39]. Further constraining of the now INSPIRE compliant datasets using notional community agreed archetypes and the O&M profile described in [15] and shown in Fig 2 (above) was performed. Data assimilation was again performed using the OpenDA toolbox. The interoperability of the data for the purpose of automatic discovery and assimilation was assessed.

Next, each dataset was loaded onto the external flash memory of three separate ARM 1GHz Cortex A8 processor-based boards with wired LAN connectivity (Fig 3). Each board represents each dataset's source observing platform. Experimental time spin-up was of the order of 60:1, meaning the 60 day period of data was re-run over a 24 hour period. The data was reported using the operational-templates-as-a-service (OPTaaS) and Linked Data knowledge graph method described in [14] (Fig 4). Data assimilation was again performed using the OpenDA toolbox, with experimental real-time assimilation of the reporting test rig system performed.

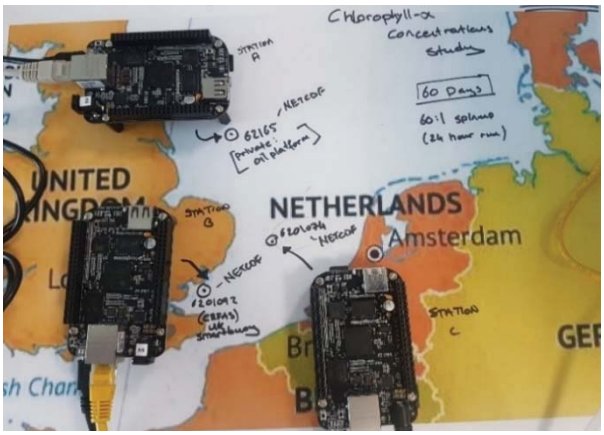

Fig. 3. Test rig. Each board represents a real deployed platform. Data for each platform was acquired from the EMODnet-physics portal.

A. Data Assimilation for Ocean Models

Data assimilation (DA) is commonly used with ocean models to improve model estimation. Data assimilation optimally blends all information available about a geophysical system to give a consistent picture of its state [43]. The most useful information to improve ocean models is obtained from in-situ sensor based observations. Data assimilation uses measured observations in combination with a dynamic system model to improve the estimates of an ocean system's states [44]. Lopez at al. [45] note the importance of assimilation of appropriate and relevant observations when estimating hydrological variables. However, the discovery, interoperability and thus assessment of the relevant observations can be challenging when meta-data describing the raw observational measurements is sparse.

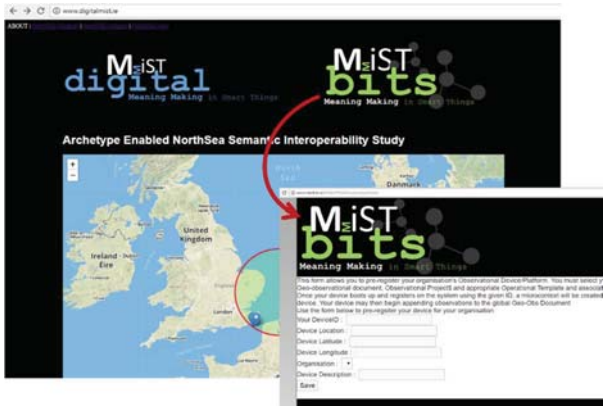

Fig. 4. The OPTaaS backend infrastructure is implemented as a set of RESTful [40] Web services using Groovy/Grails [41] and Java [42]. New platforms can register against community agreed archetypes/opts where the platform then receives a micro-context template to constrain their observational data

Increasing the number of observations and observation points that are assimilated into estimation models greatly improves model forecasting results. In-situ observational data are typically accurate and timely and thus once properly described, they can present an opportunity for more accurate estimations [46]. In [47] its was shown that a seven day forecast for sea levels and ocean currents was significantly improved when moving from one altimeter to two. There are numerous methods used for assimilating observations with ocean models. The two main categories are variational methods and sequential methods. Sequential methods are used when assimilation takes place when new observations become available. Kalman filters [48] are commonly used as a sequential method for assimilating ocean observations. A Kalman filter is used with linear systems and the extended Kalman filter can be used for non-linear systems.

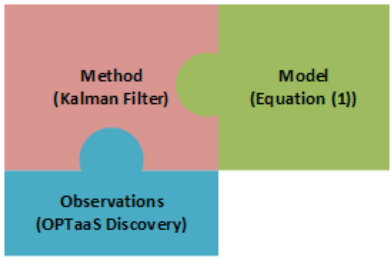

Fig. 5. OpenDA. Method represents data assimilation algorithm. Observations can be stored in netCDF, CSV, NOOS format for time-series or an SQL Database [39]

Improving the assimilation process is an active area of research. The ensemble Kalman filter is an updated version of the extended Kalman filter and is more computationally efficient. Today ensembles are used to improve forecasting. Ensembles are the combination of results from numerous models. The singular evolutive extended Kalman filter (SEEK) [43] further improves the assimilation process for oceanography. These developments are largely driven by the increasing availability of ocean observational data, such as

satellite oceanography [49] and the ability of the filter to evolve as new data becomes available.

There are many tools to aid assimilation such as OpenDA [39]. MOVE [50], ECMWF [51] and PEODAS [52]. OpenDA is a free open source data assimilation tool box primarily written in Java. OpenDA is actively used in several other assimilation projects and tools such as SANGOMA [53].

B. Predictability of chlorophyll-a fluctuations

In [24], Blauw shows how the predictability of chlorophyll concentrations from environmental variables increases greatly when environmental variables monitored from in-situ mooring stations are included within GAM models. Blauw highlights the need for fine grained monitoring of ocean regions through the deployment of in-situ observing platforms. Blauw's results show that the driving forces for chlorophyll fluctuation differ in different regions of the North Sea.

For this work a *simple model*, *simple method* and *lots of observations* approach is adopted. If the model is simple, it is less computationally intensive. Maximizing observations means less grid interpolation is necessary. Therefore, the approach seeks to harvest as many useable observations as possible. For the purposes of investigation, a deliberately oversimplified GAM model is used (1). Assumed, is an ideal and simplified linear relationship between temperature, salinity and chlorophyll-α concentrations within the southern North Sea region. In (1) μ represents mean chlorophyll-α concentrations from previous model runs. A 2-dimensional square grid with 6 grid points is used, constant depth is assumed.

$$\begin{aligned} Chla = \mu + & f_1(salinity) \\ & + f_2(sea\_surface\_temp) \\ & + f_3(lng, lat) \\ & + f_4(month) \end{aligned} \quad (1)$$

A Kalman filter is used for assimilation of observations into the model. As new observations are discovered using semantic search and a semantic reasoner, using the OPTaaS system, they are automatically assimilated in real-time into the model (Fig 5).

V. Ocean Spatial Data Infrastructure (SDI) Review & Results

Downstream services such as EMODnet-physics greatly enhance the ability of end users to consume high quality marine data products. New applications arising from the availability of high quality data need to be cognizant of the EU Inspire Directive. With a combination of Copernicus Marine Environment Monitoring Service's (CMEMS) [54] In Situ Thematic Centre IN STAC [55] and EMODnet users have access to harmonized open access data that has under gone automatic and manual data quality checks, and have been augmented with additional metadata. EMODnet's gateway contains seven thematic data portals.

The EMODnet-physics data ingestion process allows data providers to contribute their dataset directly to the EMODnet operational oceanography data exchange. Data providers will typically collect, control and distribute their data based on their own rules [56]. EMODnet provides regional coordinators to work with data providers to enable the setup a new data flows. Where data providers are not in the position to harmonize their datasets with the EMODnet system, regional coordinators perform the task of data harvesting and harmonization.

EMODnet-physics acts as a downstream service for CMEMS-INSTAC and SeaDataNet. The CMEMS-INSTAC service performs the harmonization and automatic quality control on datasets at one of five regional centers. Quality checks are defined by the EuroGOOS Data Management Exchange and Quality Working Group (DATAMEQ) [57]. A conversion to a unique NetCDF format is performed at Regional Data Acquisition Centers (RDAC) by trained staff. INS-TAC uses the OceanSITES netCDF format [58]. OceanSITES netCDF is Climate and Forecast (CF) standard [59] compliant and is recommended by CMEMS and EuroGOOS. INS-TAC produces quality controlled aggregations of in-situ observational data using OceanSITES netCDF. In order to aid this process, CMEMS provides the *oceanotron* server to manage the dissemination of data collections [60]. The data model employed by oceanotron is based on the Climate Science Modeling Language (CSML) [61] and aims to be compliant with O&M and CF discrete sampling feature. CSML is in fact a specialist profile of O&M. CSML 3.0 is based on O&M and is aligned with binary CF netCDF.

A. NetCDF-CF

The netCDF standardized data model is domain independent [62]. NetCDF specifies that datasets should be self-describing. However netCDF files are not mandated to be self-describing. NetCDF files contain both array-oriented data and meta-data. Due to its generic nature, netCDF is not specific to any domain, and so has wide applicability. Also, due its generic data model, further metadata standards are usually employed within a domain to ensure data served in netCDF are interoperable. As is the case with OceanSITES netCDF mentioned above, the CF metadata standard is often combined with netCDF to describe in further detail how to encode oceanographic and other geographical feature based datasets. CF enables additional constraints to be applied to netCDF data sets in terms of space, time, units and standard naming conventions etc. CF conventions require implementing datasets to contain sufficient self-describing meta-data so that each variable has an appropriate level of descriptive meta-data.

One of the core advantages of using the CF conventions to describe data is the CF standard-names controlled vocabulary [63]. The standard names are used when describing geophysical quantities. For example, sea water temperature is standardized to the entry id *sea_water_temperature*. CF standard names include associated units and a description of the represented quantity. For example, to further describe sea water temperature at a particular depth, a vertical coordinate variable should also be included in the dataset. There has been some criticism of CF conventions, as many attributes are optional. This means that data providers have typically omitted the attributes that are

needed to fully understand the meaning of the structure of the data [64].

CF conventions are based on an open governance model with a bottom up standards process. This means that any community member can propose changes to the conventions. The community consensus approach employed by CF conventions have been key to its success. This approach has allowed the bridging of a diverse group of earth system modeling communities. CF conventions are documented in online resources. However, these resources do not currently allow for immediate discovery and integration of datasets. The netCDF-LD extension [65] seeks to allow the creation of netCDF compliant files that can also support linked open data principles. Implementing CF conventions with Attribute Conventions for Data Discovery (ACDD) [66] can also enhance data linking and data discovery when processing data sets.

B. INSPIRE & Oceansites netCDF Format

Within INSPIRE IR Requirement Annex IV [67] it states that any data related to the theme *oceanographic geographical features* (OF) shall be made available using a number of types, such as:

- PointObservation
- PointTimeSeriesObservation

All of the types listed in [68] and above are constraints to the O&M model. INSPIRE maintains a managed code list of recommended terms including the CF standard names. The INSPIRE ocean geographical features theme uses the O&M standard to ensure consistent encoding of observations. Observations can be measured, modeled and simulated. As O&M is a generic model, INSPIRE provides numerous extensions. One important extension to O&M is the complex properties model [68]. The complex properties model allows system developers to produce interoperable observational data with the necessary fine grained detail to describe the properties of the observation. However, Leadbetter et al. [69] argue that the existing INSPIRE complex properties extension is too abstract in terms of real-world implementation. Highlighted is the fact that ocean observations typically require a quantity and a mathematical approach to describe the observed property. The initial captured quantity may undergo statistical transformation and adjustment before being encoded in the data stream. However the details of the statistical process used is not captured in the data set. This is typically important information, needed for re-use of the processed data.

Oceansites includes a quality check (QC) meta-data for each data item. The reported QC indicator is typically on a simple scale (0-6 for example). However, the more detailed process of how the QC indicator was arrived at is not automatically linked with the actual dataset. It has been proposed that netCDF-LD can provide a solution to this, allowing provenance to be captured in the meta-data, separate to the actual data and thus reducing the overhead of quality information tied to datasets.

By the end of 2020 all INSPIRE obligations must be implemented by EU member states. EMODnet aims to use INSPIRE standards. However as noted in [70] EMODnet may require solutions that diverge from INSPIRE. [71] Gives a good overview of EMODnet compliance with INSPIRE. Also EMODnet has conducted a number of pilots such as the real-time oceanography data exchange pilot using SWE [72].

C. Interoperability Assessment

Blauw et al. [73] illustrate the complexity of using in-situ observed ocean data sets. In their work observations from the Cefas operated WARP (TH1) NMMP SmartBuoy (WARP CEFAS- 62010720) were used to examine the interplay between coastal phytoplankton and the tidal cycle. They obtained observations directly from the Cefas website [74]. Based on the instruments used and the calibration information available, a number of data cleansing steps were required to ensure the data were suitable for analysis. Datasets for WARP CEFAS-62010720 obtained from the EMODnet-physics portal were examined by the authors. The datasets include the quality check data from the CMEMS INS-TAC processing centers. These quality checks perform a number of functions such as spike detection and statistical controls; more details can be found in [75]. However, the additional information required for the data cleansing steps conducted in [73] is not encoded directly or indirectly in the dataset. Currently O&M extensions do not mandate this level of interoperability. This example illustrates the requirement for a mechanism that allows organizations to further constrain and describe their information based on individual platform deployments.

As described previously, INS-TAC processes data in a number of regional centers. The regional centers provide the quality and validation steps for the final data product. The regional centers use the oceanotron server, which disseminates data flows using the OceanSITES for Copernicus standard, consisting of netCDF CF and to an extent O&M compliant data representations. The OceanSITES for Copernicus standard is hard-coded into the oceanotron software. There for oceanotron will be subject to creeping obsolescence; as the standards evolve based on the rich and growing community of supporters. This is already evident as oceanotron uses CF conventions version 1.6. At the time of writing CF conventions are at version 1.8-draft. This requires the oceanotron software to be updated and re-distributed to centers. Presently this is not a difficult task as the number of centers using the software is small. However, the scalability of this approach must be questioned. Ideally integration services such as CMEMS INS-TAC should happen in a more distributed manner, using a total data quality approach from the point of capture.

The EMODnet-physics hosted platform WARP CEFAS-62010720 has undergone the CMEMS INS-TAC integration process. At the platform's dashboard, SOAP API, GEOSERVER OGC, THREDDS and ERDAP services are

```
archetype (adl version 1.4)
  TPOT-OM-Geo_Data_Document.north_sea.v1
concept
  [at0000]
Language original_language = <[ISO_639-1::en]>
Description original_author = < lifecycle_state = <"Draft">
    details = <["en"] = <language = <[ISO_639-1::en]>>
    >
definition
  Geo_Data_Document[at0000] occurrences matches {1..1} matches {  -- north_sea
  archetype_id existence matches {0..1} matches {*}
  details existence matches {1..1} matches { ......}
  geoDataComposition existence matches {0..1} cardinality matches {0..*; unordered; unique}
  matches {
    GeoData_COMPOSITION[at0001] occurrences matches {0..*} matches {  -- Slot
      observation_Set_ existence matches {1..1} cardinality matches {1..*; unordered; unique}
      matches {
        OBSERVATION[at0002] occurrences matches {0..*} matches {  -- Slot
          featureofinterest existence matches {1..1} matches {..}
          obsproperty existence matches {1..1} matches {
            ObservedProperty[at0006] occurrences matches {1..1} matches {*} --Slot
         details existence matches {1..1} matches {
             DETAILS_COMPOUND [at0008] occurrences matches {*} -- Slot
          }
         }
          resultTime existence matches {1..1} cardinality matches {...}
          results_cluster existence matches {1..1} cardinality matches {1..*; unordered;
          unique} matches {
            Results[at0009] occurrences matches {1..*} matches {*}  -- Slot
          }
              procedure existence matches {1..1} matches {*}
        } } } } }
ontology
  term_definitions = <
    ["en"] = <
      items = < ....
      ["at0001"] = < . . . < . . solved to {TPOT-OM-GeoData_COMPOSITION.platform-oceanSITES-moorings.v1}">>
      ["at0002"] = < . . . < . . solved to {TPOT-OM-OBSERVATION.PSAL_Obs.v1}">>
      ["at0006"] = < . . . < . . solved to {TPOT-OM-ObservedProperty.PSAL.v1}">>
      ["at0008"] = < . . . < . . solved to {TPOT-OM-DETAILS_COMPOUND.ComplexProperties.v1}">>
      ["at0009"] = < . . . < . . solved to {TPOT-OM-Results.PointTimeSeries.v1}">>
>  >  >
```

Listing 1. ADL Snippet of an archetype for the north_sea. The north_sea archetype is constructed using many other archetypes, a number are shown here in the summarized ADL file. Where concepts are described as external archetypes these are labelled as – ***Slot***. Slots are bound to external archetypes using at-codes. For example, above it can be seen that the details attribute at0008 is in fact governed by the complex properties archetype.

provided. Also a sensorML descriptor is provided. The OGC and SensorML descriptors are provided at a minimum requirement level for compliance. SensorML provides a mechanism to further describe the sensing process used to obtain observations, such as sensor calibration data. However, this is level of detail is not currently available for this platform. WMS and WFS minimum compliance is provided. Within the Copernicus hosted platform page, Sensor Observation Services are not yet available and full O&M compliance is not observed. For example the feature-of-interest is not encoded in an O&M compliant manner. Two other platforms listed below were also examined using the data flows obtained from the EMODnet-physics downstream service.

- EMODnet-physics hosted platform TWEms BSH – 10004 platform.
- EMODnet-physics hosted platform FoxtrottLightship Met Office – 62170 platform.

D. Archetype Modeling and Mapping

Archetypes and two level modeling provide a way to model and organize documentation about topics of interest in a standardized way. For the platform based observations under investigation, the netCDF data model acts as an organizer, it does not represent a documentation model or a conceptual data model. Archetypes and two level modeling provide a way to model and organize documentation about topics of interest in a standardized way. In two level modeling *compositions* represent storage concepts; *sections* represent organization concepts; and an *entry* represents content concepts. Composition, section and entry can be seen highlighted in the augmented O&M model in Fig 2. Identity and topic-of-information must also be modeled.

For this work, *region* serves as the identity-model. *Sea region* is a sub theme of *region* and *OceanRegion* within CMEMS and INS TAC. The CF standard-name for the region under investigation is used - *north_sea* -, meaning the north_sea OceanRegion is the topic of information for this study (see Listing 1).

A COMPOSITION concept can be considered to be a transaction and a unit of committal. Within the reference model (Fig. 2) GeoData_Composition represents a stable composition concept from which further concepts can be defined using archetypes. As observing platforms may have short deployment times and therefore may only exist temporally, here an observing platform deployment is considered a unit of committal. Its purpose in this study is to capture a passing ocean observing event or a longer term observing deployment. Thus the following archetype is defined: *TPOT-OM-GeoData_COMPOSITION.platform-oceanSITES-moorings.v1* (Shown as Archetype B in Fig 6).

A SECTION represents an organization concept. Within the reference model Observation_Set represents a stable section concept. The purpose of a netCDF file is somewhat analogous to a section. Here a section is an ordered list of content items, this is also true of netCDF files However netCDF files contain

much more information besides. In fact, much of the additional meta-data within a netCDF file is repeated per netCDF file.

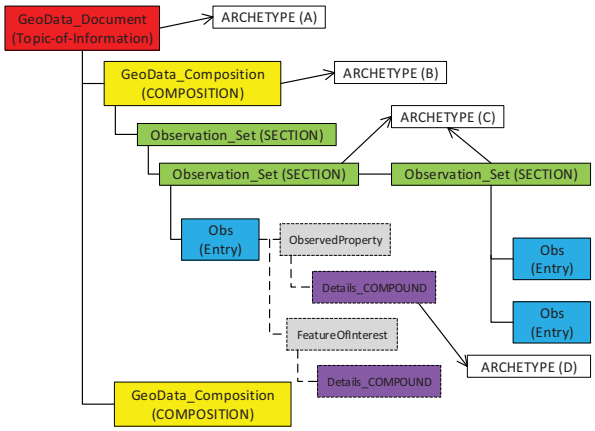

Fig. 6. Shown is the extent to which each archetype defines the overall model. GeoData_Document represents the top level document, which contains an aggregation of compositions. Compositions are storage level concepts, in this case the document about the north_sea has numerous observing platforms which are COMPOSITIONS and governed by Archetype B. Archetype C is defined based on part of the OceanSITES netCDF model where observations are organized daily. Archetype D represents the INSPIRE defined complex properties profile of O&M, which has been further specialized.

Sections may contain more sections or entries. For this study the netCDF *variables.attributes* concept is chosen as a constraint on the Observation_Set reference model concept. For convenience the archetype name netCDF-attr is used. Therefore the following archetype is defined: *TPOT-OM-Observation_Set.netCDF.netCDFattrdaily.v1* (Shown as Archetype C in Fig 6).

An ENTRY represents details of data elements. Within the reference model (Fig 2) OM_Observation represents a stable ENTRY concept. Here the practical salinity concept is mapped to an OceanSITES/INSPIRE/O&M compliant data model using the following archetypes:

- *TPOT-OM-OBSERVATION.PSAL_Obs.v1*
- *TPOT-OM-ObservedProperty.PSAL.v1*
- *TPOT-OM-OM_Observation.oceansitesObs.pointtimeseries.v1*

Shown in Fig 2. ObservedProperty contains a COMPOUND type attribute called details. Details_COMPOUND allows for the further constraining and specialization of observed properties. As mentioned previously, INSPIRE already defines an O&M extension called the complex properties model. Here the complex properties model is redefined as an archetype *TPOT-OM-Details_COMPOUND.complex_properties.v1* (Shown as Archetype D in Fig 6). Redefining the complex properties model as an archetype allows for further managed specialization and helps address the issue (described in [69]) of the complex properties model being overly abstract.

The archetypes listed above are combined to create a set of operational templates. OPTs are then used by the prototype embedded observing platforms to create information instances (Fig 6). In the prototype system, when a platform is ready to come online, the provider pre-registers the platform on the OPTaaS backend system, selecting which set of templates the platform should use Fig 4. A pre-registration ID is returned. This pre-registration ID is then used by the platform to register fully on the backend system when the platform is live. Platforms register by calling the following URL and passing their unique pre-registrationID: *http://mistbits.ie:8080/OPTaaSDev/register/{pre-red-ID}* The OPTaaS backend system then builds a constrained micro context which acts as a micro template for the platform to create information instances (see [14] for more details). When observational platforms need to report new observations they use the OPTaaS observations append Web service. Platforms call the URL below, using a POST method and passing the observations in the format defined in the platforms micro context template. *http://mistbits.ie:8080/OPTaaSDev/obsappend/{patformID}*. The observation append Web service appends the new observations as a new section with associated entries for the particular composition relating to the reporting platform. The act of appending observations involves a validation step to ensure the information instance adheres to the platforms set of operational templates. It is important to note that appending observations is adding information to the overall document about the topic-of-interest. In this case the north_sea.

E. Automatic Discovery and Assimilation

Systems that use archetypes may also use the Archetype Query Language (AQL) [76]. AQL queries are expressed based on semantics defined within the archetype level. An AQL query statement may be scoped within a particular record/geo-data-document or all documents based on a particular archetype. Using AQL a fined grained automatic assessment of newly discovered data-flows relevant to an application can be made. This is enabled by the rich meta-data associated with each information object, standardized to meet the community agreed constraints. Currently the testing framework does not support AQL. However, the OPTaaS infrastructure further described in [10] uses a linked data approach to build information instances. In the OPTaaS backend archetypes are represented using OWL (converted from ADL). Archetype/OWL governed documents are captured as knowledge graphs and SPARQL endpoints are available. In [77] it is shown how archetyped SPARQL queries may be constructed using quality indicators. Here a similar approach is adopted to enable the automatic discovery of relevant observing platforms. In this use-case as each platform becomes live it is discovered using an archetype SPARQL query. The data-flow is assessed for relevance to the study with fine grained search terms against the platforms governed archetypes. The quality of the data flow is also assessed. In this instance the system is configured to accept the data-flow and

assimilate it into the ocean model. Using the OPTaaS infrastructure as new platforms register they are automatically discovered. Once discovered their data flows are accepted for automatic assimilation into the model.

VI. Conclusion

Data pre-processing is an important step when assimilating data from heterogeneous sources. To ensure data sources are truly interoperable the meta-data must be detailed enough for systems to manually assess the data-sets suitability for automatic assimilation into the system. Also, adhering to principle of collect once use multiple times, and *find-bind-publish*, data providers may wish to publish the cleansed data-set including data provenance in an interoperable way, appending to an overall document relating to a topic-of-interest, such as the North Sea.

Retrieving data from current spatial data infrastructures can be a cumbersome process. Current SDI implementations do not allow for easy automatic discovery and federation of ocean observational data flows. The results of the use-case presented here show that discovery and assimilation of data can be automated with a high degree of confidence when systems adhere to community generated archetype models. This evaluation has shown the approach to be flexible and robust in real-world scenarios. It has also shown that the approach is in keeping with current interoperability efforts and is compatible with existing standards. The real advantage of the approach is that it improves the ability of systems to automatically discover relevant data flows and data sets and due to the verbosity of the quality data enables the federation and automatic assimilation of the data into applications.

Acknowledgement

Data used in this work was accessed from the EMODnet-physics portal. These data were collected and made freely available by the Copernicus project and the programs that contribute to it.

References

[1] B. D. Borges, "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities," Revista Negotium, (10), pp. 89-91, 2008.

[2] European Commission, "Green Paper-Marine Knowledge 2020—from seabed mapping to ocean forecasting," 2012.

[3] Columbus Consortium, "Marine Portals and Repositories and their role in Knowledge Transfer to support Blue Growth," 2016.

[4] I. Directive, "Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)," Published in the Official Journal on the 25th April, 2007.

[5] European Commission, "European Marine Observation and Data Network", 2010. [online]. Available http://www.emodnet-physics.eu/. [Accessed: April 2018]

[6] D. M. Schaap and R. K. Lowry, "SeaDataNet–Pan-European infrastructure for marine and ocean data management: unified access to distributed data sets," International Journal of Digital Earth, vol. 3, (S1), pp. 50-69, 2010.

[7] C. Antoine, V. Sandrine and F. Jean-Valery, "JERICO-NEXT. Report on Developments Dedicated to Monitor and Study Benthic Comportment and Processes.D3.10", 2017.

[8] A. Fischer et al, "Initial AtlantOS Requirements Report," 2016.

[9] H. Glaves et al, "Ocean data interoperability platform (ODIP): Supporting the development of a common global framework for marine data management through international collaboration," in EGU General Assembly Conference Abstracts, 2014.

[10] "IOC Oceans | United Nations Educational, Scientific and Cultural Organization, UNESCO". [Online]. Available: http://www.unesco.org/new/en/natural-sciences/ioc-oceans. [Accessed: 02-Jul-2018].

[11] ISO 19156:2011 Geographic information – Observations and measurements. 2011. doi:10.13140/2.1.1142.3042.

[12] S. Nativi, and & J. Bemmelen, "GEO DAB and GEOSS Portal", in AGU Abstract Fall Meeting, 2016.

[13] S. Olenin et al, "Marine strategy framework directive," Task Group, vol. 2, 2010.

[14] P. Stacey and D. Berry, "Design and implementation of an archetype based interoperable knowledge eco-system for data buoys", OCEANS 2017 - Aberdeen, Aberdeen, 2017, pp. 1-9. doi: 10.1109/OCEANSE.2017.8084936

[15] P. Stacey and D. Berry, "Towards a Digital Earth: using archetypes to enable knowledge interoperability within geo-observational sensor systems design", Journal of Earth Science Informatics, Springer, 2018. Doi:10.1007/s12145-018-0340-z

[16] Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan, ESA Publications Division, Feb. 2005.

[17] T. Beale, "Archetypes: Constraint-based domain models for future-proof information systems," in OOPSLA 2002 Workshop on Behavioral Semantics, 2002.

[18] "Open Geospatial Consortium". [Online]. Available: http://www.opengeospatial.org/. [Accessed: 02-Jul-2018].

[19] T. Beale, and S. Heard, "Archetype Definition Language". The openEHR Foundation, London, .2007.

[20] A. P. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," ACM Computing Surveys (CSUR), vol. 22, (3), pp. 183-236, 1990.

[21] T. P. Barnett et al, "Penetration of human-induced warming into the world's oceans," Science, vol. 309, (5732), pp. 284-287, Jul 8, 2005.

[22] Deltares, "D-Water Quality Manual". July, 2018.

[23] A. J. Irwin and Z. V. Finkel, "Mining a sea of data: Deducing the environmental controls of ocean chlorophyll," PloS One, vol. 3, (11), pp. e3836, 2008.

[24] A. N. Blauw, Monitoring and Prediction of Phytoplankton Dynamics in the North Sea. PhD thesis, 2015.

[25] M. Holt, "Towards NOOS—The EuroGOOS NW shelf task," in Building the European Capacity in Operational Oceanography: Proceedings 3rd EuroGOOS Conference, 2003, pp. 461.

[26] "North West Shelf Data Portal". [Online]. Available: http://nwsportal.bsh.de/. [Accessed: 02-Jul-2018].

[27] J. Woods et al, "The strategy for EuroGOOS," EuroGOOS Publication, vol. 1, 1996.

[28] P. Dexter and C. Summerhayes, "Ocean observations–the Global Ocean Observing System (GOOS)," D.Pugh, G.Holland G (Eds.), Troubled Waters: Ocean Science and Governance, Cambridge University Press, Cambridge, pp. 161-178, 2010.

[29] T. C. Malone and W. D. Nowlin, "GOOS regional alliances and the development of the coastal module of GOOS," in US/EU Baltic International Symposium, 2006 IEEE, 2006, pp. 1-16.

[30] "EuroGOOS Ocean Models", [Online]. Available: http://eurogoos.net/models. [Accessed: 02-Jul-2018].

[31] H. Gerritsen, H. De Vries and M. Philippart, "The Dutch continental shelf model," Coastal and Estuarine Studies, pp. 425-425, 1995.

[32] R. Hordoir et al, "Nemo-Nordic 1.0: A NEMO based ocean model for Baltic & North Seas, research and operational applications," Geoscientific Model Development Discussions, 2018.

[33] F. Los, M. Villars and M. Van der Tol, "A 3-dimensional primary production model (BLOOM/GEM) and its applications to the (southern)

North Sea (coupled physical–chemical–ecological model)," J. Mar. Syst., vol. 74, (1-2), pp. 259-294, 2008.

[34] T. J. Hastie, "Generalized additive models," in Statistical Models in S, Anonymous Routledge, 2017, pp. 249-307.

[35] European Commission, "Marine INSPIRE Pilot", 2016. [online]. Available: http://inspire-marine.jrc.ec.europa.eu/. [Accessed: Accessed: 02-Jul-2018]

[36] X. Desmit, K. Ruddick and G. Lacroix, "Salinity predicts the distribution of chlorophyll a spring peak in the southern North Sea continental waters," J. Sea Res., vol. 103, pp. 59-74, 2015.

[37] A. Novellino et al, "European marine observation data network—EMODnet physics," in OCEANS 2015-Genova, 2015, pp. 1-6.

[38] C. Zender, P. Vicente and W. Wang, "NCO: Simpler and faster model evaluation by NASA satellite data via unified file-level netCDF and HDF-EOS data post-processing tools," in AGU Fall Meeting Abstracts, 2012.

[39] M. Verlaan et al, "OpenDA, a generic toolbox for data-assimilation in numerical modelling," in 15th Biennial Conference of the Joint Numerical Sea Modelling Group, Delft, the Netherlands, 2010.

[40] R. T. Fielding and R. N. Taylor, "Principled design of the modern Web architecture," ACM Transactions on Internet Technology (TOIT), vol. 2, (2), pp. 115-150, 2002.

[41] G. Rocher et al, "The Grails Framework-Reference Documentation," Environments, vol. 4, pp. 3, 2009.

[42] J. Gosling and H. McGilton, "The Java language environment," Sun Microsystems Computer Company, vol. 2550, 1995.

[43] D. T. Pham, J. Verron and M. C. Roubaud, "A singular evolutive extended Kalman filter for data assimilation in oceanography," J. Mar. Syst., vol. 16, (3-4), pp. 323-340, 1998.

[44] G. Markensteijn, "Performing data assimilation experiments with hydrodynamic models: A Java Sea case," 2017.

[45] P. L. Lopez et al, "Improved large-scale hydrological modelling through the assimilation of streamflow and downscaled satellite soil moisture observations," Hydrology and Earth System Sciences, vol. 20, (7), pp. 3059-3076, 2016.

[46] M. E. Ridler et al, "Data assimilation framework: Linking an open data assimilation library (OpenDA) to a widely adopted model interface (OpenMI)," Environmental Modelling & Software, vol. 57, pp. 76-89, 2014.

[47] S. Verrier, P. Le Traon and E. Remy, "Assessing the impact of multiple altimeter missions and Argo in a global eddy-permitting data assimilation system." Ocean Science, vol. 13, (6), 2017.

[48] R. E. Kalman, "A new approach to linear filtering and prediction problems," Journal of Basic Engineering, vol. 82, (1), pp. 35-45, 1960.

[49] C. L. Parkinson, A. Ward and M. D. King, "Earth science reference handbook: a guide to NASA's earth science program and earth observing satellite missions," National Aeronautics and Space Administration, pp. 277, 2006.

[50] N. Usui et al, "Meteorological Research Institute multivariate ocean variational estimation (MOVE) system: Some early results," Advances in Space Research, vol. 37, (4), pp. 806-822, 2006.

[51] M. A. Balmaseda, K. Mogensen and A. T. Weaver, "Evaluation of the ECMWF ocean reanalysis system ORAS4," Q. J. R. Meteorol. Soc., vol. 139, (674), pp. 1132-1161, 2013.

[52] Y. Yin, O. Alves and P. R. Oke, "An ensemble ocean data assimilation system for seasonal prediction," Mon. Weather Rev., vol. 139, (3), pp. 786-808, 2011.

[53] L. Nerger et al, "SANGOMA: Stochastic Assimilation for the Next Generation Ocean Model Applications EU FP7 SPACE-2011-1 project 283580," 2011.

[54] K. von Schuckmann, "the CMEMS OSR task team: The Copernicus Marine Environment Monitoring Service Ocean State Report," J.Oper.Oceanogr, vol. 9, pp. s235-s320,.

[55] Copernicus Marine, [Online]. Available: http://marine.copernicus.eu/situ-thematic-centre-ins-tac/. [Accessed: 02-Jul-2018].

[56] EMODnet Ingestion Portal, [Online]. Available: https://www.emodnet-ingestion.eu/. [Accessed: 02-Jul-2018].

[57] S. Pouliquen, "Recommendations for in-situ data Real Time Quality Control," 2011.

[58] OceanSites, "OceanSITES Data Format Reference Manual", 2015. [Online]. Available: http://www.oceansites.org/docs/oceansites_data_format_reference_manual.pdf. [Accessed: 02-Jul-2018].

[59] J. Gregory, "The CF metadata standard," CLIVAR Exchanges, vol. 8, (4), pp. 4, 2003.

[60] Copernicus Marine In Situ Tac Data Management Team, "Copernicus in situ TAC - CMEMS System Requirements Document." http://doi.org/10.13155/40846, 2017.

[61] A. Woolf et al, "Climate science modelling language: Standards-based markup for metocean data," in Proceedings of 85th Meeting of American Meteorological Society, 2005.

[62] R. Rew and G. Davis, "NetCDF: an interface for scientific data access," IEEE Comput. Graphics Appl., vol. 10, (4), pp. 76-82, 1990.

[63] CF standard names, [Online]. Available: http://cfconventions.org/standard-names.html. [Accessed: 02-Jul-2018]

[64] NASA CF Conventions, [Online]. Available: https://earthdata.nasa.gov/user-resources/standards-and-references/climate-and-forecast-cf-metadata-conventions. [Accessed: 02-Jul-2018].

[65] N. J. Car, A. Ip and K. Druken, "netCDF-LD SKOS: Demonstrating linked data vocabulary use within netCDF-compliant files," in Environmental Software Systems. Computer Science for Environmental Protection: 12th IFIP WG 5.11 International Symposium, ISESS 2017, Croatia, 2017, Proceedings 12, 2017, pp. 329-337.

[66] "Attribute Conventions for Data Discovery", [Online]. Available: https://www.unidata.ucar.edu/software/thredds/v4.3/netcdf-java/formats/DataDiscoveryAttConvention.html. [Accessed: 02-Jul-2018].

[67] INSPIRE, "Technical Guidance for the implementation of INSPIRE Download Services version 3.0", 2013.

[68] INSPIRE, "D2.9 Draft Guidelines for the use of Observations & Measurements and Sensor Web Enablement-related standards in INSPIRE Annex II and III data specification development", 2013.

[69] A. M. Leadbetter and P. N. Vodden, "Semantic linking of complex properties, monitoring processes and facilities in web-based representations of the environment," International Journal of Digital Earth, vol. 9, (3), pp. 300-324, 2016.

[70] K. Millard, P. Smits, A. Abramic, J. B. Calewaert, I. Shepherd, "Marine Pilot - EMODnet and INSPIRE: Benefits of Closer Collaboration and a Framework for Action". 2015.

[71] "EMODnet compliance with the INSPIRE Directive: a matter of fact", [Online]. Available: http://Emodnet.eu/emodnet-compliance-inspire-directive-matter-fact. [Accessed: 02-Jul-2018].

[72] "Marine Profiles of the OGC Sensor Web Enablement Standards", [Online]. Available: https://odip.github.io/MarineProfilesForSWE/. [Accessed: 02-Jul-2018].

[73] A. N. Blauw et al, "Dancing with the tides: fluctuations of coastal phytoplankton orchestrated by different oscillatory modes of the tidal cycle," PLoS One, vol. 7, (11), pp. e49319, 2012.

[74] Centre for Environment, Fisheries and Aquaculture Science, "Cefas Data Hub", [Online]. Available: https://www.cefas.co.uk/cefas-data-hub/cefas-data-hub-apis/. [Accessed: 02-Jul-2018].

[75] H. Wehde et al., "CMEMS Quality Information Document". Available: http://marine.copernicus.eu/documents/QUID/CMEMS-INS-QUID-013-030-036.pdf, 2016.

[76] AQL, "Archetype Query Language 1.0.3". Available: http://www.openehr.org/releases/QUERY/latest/docs/AQL/AQL.htmlAQL. 2015.

[77] K. Dentler et al, "Semantic integration of patient data and quality indicators based on openEHR archetypes," in Process Support and Knowledge Representation in Health Care, Springer, 2012, pp. 85-97.