A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences

Valentina Beretta, Jean-Christophe Desconnets, I. Mougenot, Muhammad Arslan, Julien Barde, Véronique Chaffard

Computers & Geosciences 154 (2021) 104807

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences

Valentina Beretta [a,*], Jean-Christophe Desconnets [a], Isabelle Mougenot [b], Muhammad Arslan [a], Julien Barde [c], Véronique Chaffard [d]

[a] Mission Science Ouverte, IRD, Marseille, France
[b] ESPACE-DEV, IRD, Université de Montpellier, Université Antilles, Université de Guyane, Université de la Réunion, Montpellier, France
[c] UMR MARBEC, IRD, IFREMER, CNRS, Université de Montpellier, Montpellier, France
[d] IGE, CNRS, IRDn Université Grenoble Alpes, Grenoble, France

ARTICLE INFO

Keywords:
Interdisciplinary datasets
Semantic metadata model
Semantics
FAIR principles

ABSTRACT

The recent technological advancements and emergence of the open data in environmental and life sciences are opening new research opportunities while creating new challenges around data management. They make available an unprecedented amount of data that can be exploited for studying complex phenomena. However, new challenges related to data management need to be addressed to ensure effective data sharing, discovery and reuse, especially when dealing with interdisciplinary research contexts. These issues are magnified in interdisciplinary context, by the fact that each discipline has its practices, e.g., specific formats and metadata standards. Moreover, the majority of current data management practices do not consider semantic heterogeneity existing among disciplines. For this reason, we introduce a flexible metadata model that describes the datasets of various disciplines using a common paradigm based on the observation concept. It provides a key vision for articulating the user point of view and underlying scientific domains. In this study, we therefore decide to mainly reuse the SOSA lightweight ontology (Sensor, Observation, Sample, and Actuator) to efficiently leverage others existing ontologies to improve datasets discovery and reuse coming from Earth and life observation. The main benefit of the proposed metadata model is that it extends the technical description, usually provided by existing metadata models, with the observation context description enabling the need of a user viewpoint. Moreover, following the FAIR principles, the metadata model specifies the semantics of its elements using ontologies and vocabularies, and reuses as much as possible ontological and terminological existing resources. We show the benefit and applicability of the model through a case study we identified as representative after interviewing researchers in environmental and life sciences.

1. Introduction

For tackling broader and complex questions about the natural world, nowadays scientists in environmental and life sciences can exploit the vast amount of data that is available through different platforms and services (Kelling et al., 2009), thanks to both the increasing advancement in technologies and the advent of open science. New data acquisition systems and abilities to process voluminous data made it possible to easily record and store measurements generated by scientific communities through laboratory analysis, scientific experiments (settings and results), microscopes, ground and satellite monitoring systems, and so on. At the same time, open science promotes the sharing and reuse of collected datasets across communities. Among this abundance of datasets, the discovery of relevant ones is not straightforward, especially when considering the intrinsic multi-disciplinarity of research questions in environmental and life sciences. The Earth is a complex system composed of highly interconnected sub-systems interacting with each other where a small change in a sphere may lead to consequences in the others (Donner et al., 2009). The projects related to this domain, therefore, require to simultaneously consider information related to aspects handled by different disciplines (Rahimi et al., 2014; Argent, 2004). For instance, a scientist interested in studying the population of a

* Corresponding author.
E-mail addresses: vberetta10@gmail.com (V. Beretta), jean-christophe.desconnets@ird.fr (J.-C. Desconnets), isabelle.mougenot@umontpellier.fr (I. Mougenot), muhammad.arlsan@ird.fr (M. Arslan), julien.barde@ird.fr (J. Barde), veronique.chaffard@ird.fr (V. Chaffard).

https://doi.org/10.1016/j.cageo.2021.104807
Received 12 July 2019; Received in revised form 18 January 2021; Accepted 28 April 2021
Available online 5 May 2021

particular fish species in a specific area should consider data from both biosphere and hydrosphere because the fish population dynamics depend on parameters like chlorophyll concentration and sea surface temperature. In this case, when the scientist will publish its dataset on the web, he will describe "chlorophyll concentration" and "sea surface temperature" as features that have an impact of its object of the study (the considered fish species). Considering its research, he will not distinguish the fact that, in turn, the feature "chlorophyll concentration" contains a potential object of interest. Therefore, other scientists that have the chlorophyll as the object of their study could not directly take advantage of this dataset. In this work, we study a flexible model that will permit scientists that study chlorophyll to easily discover and reuse the dataset initially used for investigating a different object of study.

Currently, given a research question, it is hard to identify to existing relevant in the realms of available resources. Each discipline has its formats, vocabularies, standards, methodologies, and best practices (Gray et al., 2005, Hey and Trefethen, 2003). These differences lead to structural and semantic heterogeneities that make data highly compartmentalized, i.e., data produced by a scientific community is often shared and reused only in that community. As a result, scientists that want to reuse datasets originating from other disciplines need to put considerable effort to learn and understand the different discipline's points of view and insights.

Over the last decades, several initiatives have been promoted and implemented the effective sharing of very large amount of Earth and life observations both at international and national levels, such as GEOSS (Battrick, 2005), LTER (Franklin et al., 1990) and GBIF (Edwards, 2004). These projects aim to let researchers and organizations continue collecting data using their methods while enabling them to share and reuse the data with researchers of other disciplines. Most often these initiatives are based on the description of homogeneous datasets (dataset-level) based on semantically restricted metadata models. It primarily provides the functions of data discovery and location. The producer's point of view is also privileged. To strengthen the discovery and reuse of very large amount of data for interdisciplinary purposes, we offer a user-centric metadata model which allows the integration of data across the silos of various Earth and life sciences domains. The contributions of this study are manifold:

- we suggest to uniformly represent datasets originating from different disciplines using a common description, which is based on the observation paradigm; more precisely, we suggest the exploitation of the SOSA observation model as a metadata model. Our motivation is based on the fact that it is more relevant and realistic, from an implementation point of view, to exploit it at metadata-level rather than observational data-level. Environmental data are very heterogeneous, extremely numerous and voluminous on which efficient discovery and integration functions can be built.
- we provide a metadata model that, embodying a user-centric viewpoint, satisfies the description needs of different communities; it characterizes a dataset based on multiple aspects that are associated with an observation (i.e. object of interest, observed property, collection protocol, spatial and temporal extents); these elements of high level of abstraction and shared and understood by the main part of the environmental community are then used, simultaneously or not, for discovering and evaluating the relevance of a dataset depending on the focus of the disciplines involved in a study (usually different disciplines privilege different aspects for discovering and evaluating datasets);
- we present an ontology[1]-based approach which reuses and take advantage of the SOSA ontology and its extension, SSN-EXT as core model. We also reuse and articulate SOSA with well-defined ontologies which offer data provenance information using PROV-O or introduce data representation which contextualize observation and scientific point of view on temporal, spatial or thematic dimensions reusing SWEET, TIME or SKOS ontologies. Finally, the Complex Properties Model is mapped with SOSA to support observable complex properties meet frequently in environmental observation.

The metadata model resulting from this study, adding semantics to metadata, provides a more efficient discoverability of data, a high level of semantic interoperability and enables highly added-value services for portals dedicated to interdisciplinary research projects such as visualization, data analysis or on-demand processing service. For instance, the researcher that is interested to study the tuna habitat in the oceans in 2012 and specifying only this information, will retrieve datasets related to occurrences of all tuna species (the model will also look for tuna synonyms), and also datasets related to the environment where tuna lives (i.e., environmental properties of that oceanic area). Indeed, the model can potentially take advantage of reasoning capabilities enabled by ontologies for inferring relevant environmental parameters associated with the ecosystem where tuna lives. The rest of this paper is structured as follows. Section 2 presents the notion of datasets in environmental and life sciences and metadata standards. Section 3 introduces the user-centric metadata model, which is discussed in Section 4. Finally, Section 5 summarizes the main findings of this study.

2. Related works

The aim of this study, as previously anticipated, is to introduce a user-centric metadata model for facilitating the discovery and reuse of datasets in interdisciplinary settings. Before presenting it, it is therefore important to specify the meaning of dataset and the meaning of metadata model.

2.1. Dataset, data model and data format

The World Wide Web Consortium (W3C), in the Data Catalog Vocabulary, defines a dataset as a collection of data published by a single agent (Albertoni et al., 2020). The fact that a single agent creates a dataset is essential because agent goals are usually determining factors for establishing which data are gathered together into a single dataset. For instance, a data producer such as a water cycle long term observatory tends to create datasets per station, each of them containing measurements on multiple properties (e.g. air temperature, humidity) of the atmosphere around the station at a given period. Instead, the researcher studying the tuna habitat in the oceans may publish an aggregate version of the datasets he used for increasing the reproducibility of his study. The dataset should contain data on tuna occurrences, chlorophyll concentration and sea surface temperature gathered in the Indian Ocean in a given time interval. Since agent goals differ, heterogeneous datasets are created. Datasets in environmental and life sciences may contain data about a large variety of real-world entities and phenomena, resulting from very different sampling protocols. Datasets may also contain measurements from remote or in-situ sensors, results from campaigns or surveys, but also research products such as experimental results and model outputs. In this study, we therefore consider a dataset as a collection of data, published by as single agent, about the measurements of one or more properties related to one or more real world entities or phenomena collected using one or more procedures in a given area and time. As shown in Fig. 1, a dataset, independently from its content, can be stored using from flexible data formats (txt, eXtensible Markup Language (XML), JSON (Crockford, 2006), csv, spreadsheet) to fixed data formats (shape-file (ESRI, 1998), GeoTIFF (Ritter and Ruth, 1997) and NetCDF (Rew and Netcdf, 1990)). Moreover, a dataset can be organized in extremely different ways according to a data model that indicates its logical organization at the data level.

[1] An ontology is a formal specification of the shared conceptualization of a domain (Knublauch & Knublauch, 2017).

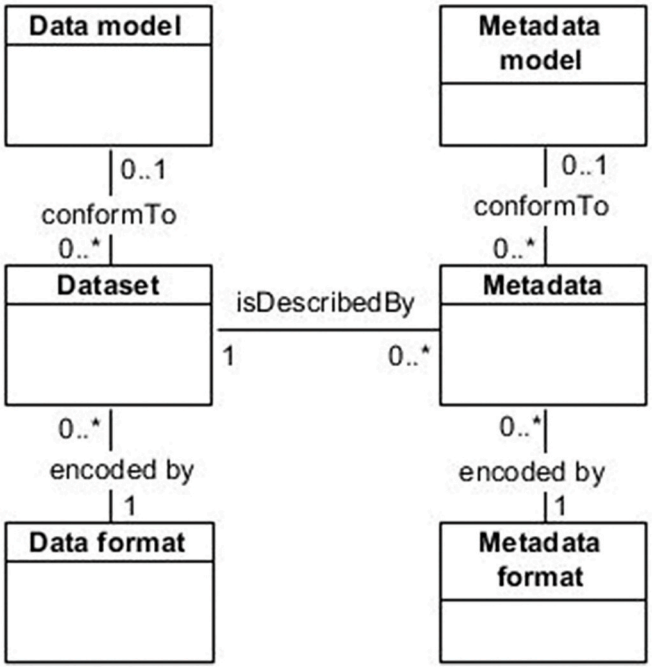

Fig. 1. Relationships among dataset, data model, data format, metadata model and metadata format.

2.1.1. Data models for representing scientific data

Data models based on the observation paradigm have been developed over the years. This paradigm considers an observation as an act that results in the estimation of the value of a feature property and involves the application of a specified protocol, such as using a sensor, instrument, algorithm or process chain (Cox, 2010). The most usual data models are 0bservations and Measurements (Cox, 2010) for environmental sciences and to a lesser extent the Extensible Observation Ontology (OBOE) (Madin et al., 2007) for ecology. SOSA (Sensor, Observation, Sample, and Actuator) (Janowicz et al., 2019) and its extension, SSN-EXT, proposed by (Cox and Little, 2020) is a standalone core ontology which aligns primarily with O&M data model (and secondarily with OBOE) and that allows the use of observation paradigm with W3C semantic Web ontologies.

Aligned with O&M data model specified by the Open Geospatial Consortium[2] (OGC), SOSA describes itself as a light-weight ontology for modelling acts of observation, sampling and actuation, using sensors, samplers and actuator respectively. With respect to acts of observation, SOSA share a similar if not identical conceptual basis as OGC/ISO O&M. SSN/SOSA is formulated as an RDF vocabulary expressed in RDF/OWL. Focused mainly on observation perspective, we will use only the classes and relationships involved in Observation part of SOSA. Fig. 2 provides an overview of the core classes and properties that are specifically related to modeling Observations.

SSN-EXT (Cox and Little, 2020) extends SOSA ontology with two new concepts very relevant for our purposes 1) the notion of ultimate feature-of-interest for an act of observation, sampling, or actuation, alongside the link to the (proximate) feature-of-interest, which might be a sample and allow to well differentiate feature closed to the observation protocol (proximate feature) and ultimate feature of interest which real study feature 2) the notion of homogeneous collections of observations, in which one or more observations properties (feature-of-interest, observed-property, procedure, sensor, …) may be shared by all members of the collection (Fig. 3). This extension is also aligned with OBOE as shown in (Cox and Little, 2020; Madin et al., 2007).

Data models are mainly adopted in environmental and life sciences for providing a description at data level. This description has to be complemented with metadata, presented hereinafter, to describe datasets in its entirely, see Fig. 1 and promoting their sharing and reuse. We consider SOSA as intermediate metamodel, less general than DCMI or DCAT metadata model but not domain specific. It is neutral-domain model bringing by the observation paradigm which is shared over disciplines.

2.2. Metadata, metadata model and metadata format

Metadata aims to specify the dataset contextual information for describing, explaining, locating or otherwise making it easier to discover, determine whether a resource is relevant and reuse it in the proper manner (Qin and D'ignazio, 2010; Desconnets, 2017). Metadata can be categorized into three classes: descriptive, structural and administrative (National Information Standards Organization, 2016). The descriptive metadata details the data content. It is usually used to search and identify a dataset. The structural metadata describes the structure of the data, and the relationship existing among different elements. The administrative metadata instead describes access and rights information. As shown in Fig. 1, metadata are independent of the data model and data format. Moreover, they can be stored using a given format such as XML or JSON-LD and can be conformed to a given metadata model. Metadata models, therefore, describe information at the dataset level, while data models describe information at the data level. The metamodel presented is close to an application profile (Tennis, 2015). The general principle is indeed the same, i.e., to reinforce the structuring of metadata elements describing a dataset. However, the objective is a little different, since we do not seek to constrain the description of the datasets, but rather to enrich it as much as possible to facilitate the discovery and reuse of these datasets in a second step. In this sense, we do not use the SHapes Constraint Language (SHACL) (Knublauch et al., 2017), which is too prescriptive.

2.2.1. Metadata models for representing scientific data

A metadata model is composed of metadata elements that are defined on the basis of both a name (property) and a literal or a non-literal value. Interestingly, literal values may be labels of ontological concepts selected from an onto-terminology, e.g., thesauri, taxonomies (see Fig. 4). The use of ontological properties and onto terminology is promoted by the FAIR principles (Wilkinson et al., 2016) to make data interoperable avoiding ambiguous semantics across disciplines.

FAIR principles are also dedicated to improving the find ability, accessibility, interoperability and reuse of digital resources. For instance, they suggest the use of rich metadata description that explicitly contain globally unique and persistent identifier (i.e., IRI or DOI) to identify and easily find a resource on the Web as well as the dataset, metadata schema vocabulary or controlled terms used to valuate metadata. This metadata has to be accessible using open standardized protocol and remains accessible even when data are no longer available. Moreover, metadata has to be integrated with license information as well as provenance to improve the resource reusability.

Scientific communities have suggested a large variety of metadata models due to the heterogeneity of their datasets. The best information for describing datasets varies from discipline to another according to the key aspect that is used for evaluating the relevance of a dataset.

For instance, geospatial scientists evaluate datasets based on their geographic location, while biologists evaluate and choose datasets considering the experimental protocol that was adopted. For instance, some standards are generic metadata that describe any resource, such as Dublin Core (DC) (Weibel et al., 1998) and.

DCTERMS (Klyne and Carroll, 2006), or datasets, such as the Data Catalog Vocabulary (DCAT) (Albertoni et al., 2020), while others are

[2] The OGC is an international consortium of academic, industry and government organizations that collaboratively develop open standards for geospatial and location services.

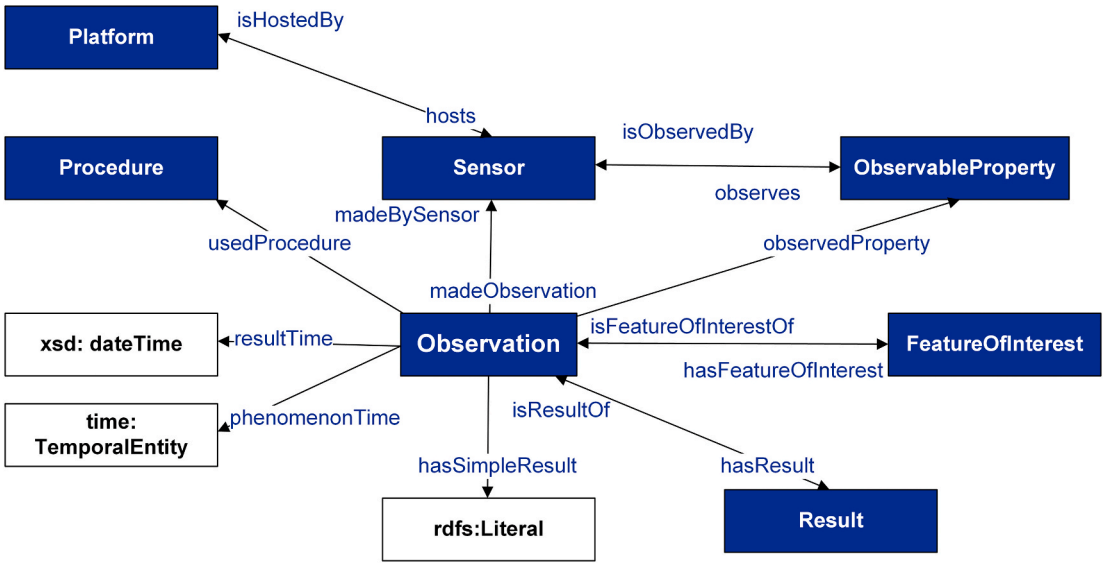

Fig. 2. Classes and relationships involved in Observation (SOSA) (Janowicz et al., 2019)

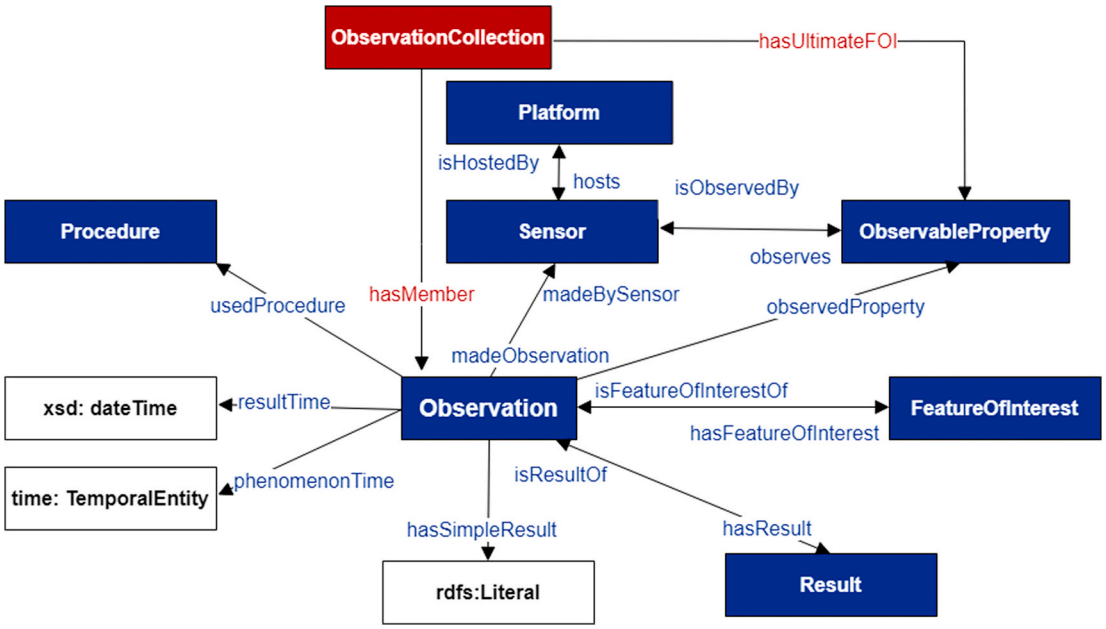

Fig. 3. Extensions to the semantic sensor network ontology (Cox and Little, 2020).

domain-dependant metadata with dedicated modules to describe datasets, such as the Ecological Metadata Language (EML) (Fegraus et al., 2005) and ISO-19115 (Danko, 2005). The advantage of domain-dependent metadata is that the publishing agent can specify fine-grained technical aspects. However, this edge becomes a disadvantage when considering multi-disciplinary settings. Highly specific metadata are very difficult to understand by non-expert users.

Moreover, considering existing metadata, only a few elements are dedicated to the context of the observation (what is being observed? how the observation was made? what features were measured?). They usually are the abstract and the keywords elements, the abstract usually contains natural language text and the keywords refer to multiple dataset aspects whose semantic relationships are not explicitly specified. An example is given by the EML metadata of the dataset from the LTER catalogue available at https://portal.lternet.edu/nis/metadataviewer?packageid=knb-lter-luq.76.223400. Scientists are therefore restrained in the discovery and reuse of datasets originating from other disciplines.

As a solution for facilitating the comprehension of a dataset and promoting a user-centric viewpoint, we decide to base the metadata model on the observation paradigm that is understood by all scientific communities.

This paradigm also enables to characterize a dataset making explicit the relationships among the main semantic aspects of its content, i.e., the observed properties, the entities that were measured. According to the FAIR principles, we reuse as much as possible existing resources. Intuitively, considering that a dataset can be seen as a result of observation, we exploit the SOSA ontology adapting the targeting granularity of the model description from data to the dataset. As O & M model, SOSA ontology because it explicitly distinguishes between sampled and sampling feature as ultimate feature-of-interest and sample. A sample is extracted from the ultimate feature-of-interest when observations cannot be made directly on a feature of interest. The SOSA extension (Cox, 2020) add the object property has Ultimate Feature Of Interest which allows to better distinguish intermediate feature, close to the

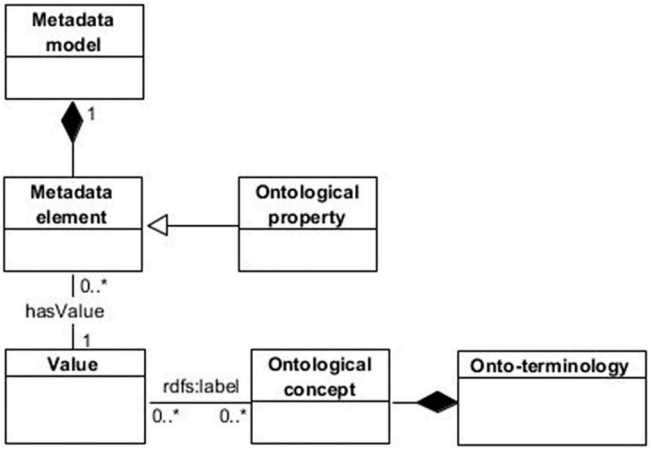

Fig. 4. Relationships among dataset, data model, data format, metadata model and metadata format.

observation protocol, and describe a specific relationship with ultimate feature of interest, most relevant for data end-users. This permits us to better encode a user-centric viewpoint.

3. A user-centric metadata model for facilitating interdisciplinary research

A user-centric metadata model is presented in this section. The aim of this model is to facilitate the multi-disciplinary research allowing scientific communities to reuse data produced by scientist with different expertise. The user-centric metadata model is sufficiently general to be widely applicable as well as compatible with a broad range of disciplines and related standards. Its primary purpose is to encode the user's point of view, making data semantically interoperable across disciplines (i.e., agreement on the considered meaning using the state-of-the-art controlled vocabularies). This means that scientists having different backgrounds can exchange and use each other datasets without difficulties. This is done by specifying a set of dataset aspects whose semantics does not vary based on the considered discipline. Moreover, to increase the level of interoperability, we adopt an ontology-based formalism that allows shared and unambiguous dataset descriptions. In this section, first, the ontologies and the vocabularies have been described on which the proposed metadata model is based on. Second, the presentation of the proposed value-added features is provided to facilitate the searching of multi-disciplinary datasets. Lastly, the proposed metadata model is instantiated, and its application is discussed using a real-life scenario.

3.1. Proposed metadata model

The proposed metadata model (see Fig. 5) uses the concept of observation as a common conceptualization paradigm across disciplines. An observation is an activity that results in the approximation of the values of a property of a feature of interest. It consists of the application of a particular procedure using a digital or human sensor, an algorithm, or a process chain (Cox, 2013). The procedure for acquiring the observation from the physical environment can be in-situ, remotely, or ex-situ based on the sampling location (Cox, 2013). Using a procedure, a dataset is generated which is the composition of results originating from different observations. Each result represents a dataset characterized by a single feature of interest (Cox, 2013). To construct a metadata model, an ontology-based formalism is used for representing its schema and the relations between its elements. The use of ontologies allows us to minimize heterogeneity problems (Kashyap and Sheth, 1998). For instance, IRI[3] associations enable the normalization of metadata records by avoiding duplicates. Moreover, we reused, as much as possible existing resources making extensive use of vocabulary included in the state-of-the-art ontologies or onto-terminologies mainly: DCAT (Albertoni et al., 2020), SOSA (Janowicz et al., 2019), SSN-EXT (Cox, 2020), TIME (Cox and Little, 2020), DCTERMS (DCMI Usage Board, 2020), PROV-O (Lebo et al., 2013), and CITO (Shotton and Peroni, 2018). The W3C web semantic vocabularies foundations are also considered, i.e., RDFS, FOAF, GEOSPARQL and SKOS ontologies. Although, the proposed model (see Fig. 5) is based on several ontologies as listed above, a few of them which are the most important towards our contribution are described below;

(i) SOSA: It is the major fundamental ontology on which our proposed model is based on (see Fig. 1). It is used for specifying metadata elements bringing observation context which is understandable for the main part of end-users targeted. It defines set of observations as a collection giving the main properties of the observation context (see Fig. 2) (Janowicz et al., 2019).
(ii) Data Catalog Vocabulary (DCAT): It is used in our model for describing the dataset characteristics. DCAT takes charge of structural and administrative metadata commonly implemented with data providers metadata schemas such as ISO 19115. DCAT is an RDF vocabulary that is designed to enable interoperability between existing published data catalogs on the web (Albertoni et al., 2020). It allows the data publishers to describe their datasets using a standard vocabulary to facilitate the usage and aggregation of metadata from different multiple data catalogs on the web.
(iii) Semantic Web for Earth and Environmental Terminology (SWEET) ontology: It is used for discovery and use of Earth science and environmental data (DiGiuseppe et al., 2014). For the development of the model, the class representation is used from this ontology to provide spatial and temporal dataset granularities in the model required to evaluate the fitness for use of the dataset and for some cases quantitative description (temporal or spatial resolution) of dataset to run a processing chain.
(vi) The PROV Ontology (PROV-O): It is used to provide a set of classes, proper ties, and restrictions for representing and interchanging provenance data generated in different systems (Lebo et al., 2013). The alignment of SOSA classes with PROV-O Activity, Plan classes offers to potentially describe precisely the dataset provenance as it is necessary for dataset coming from complex and nested processing chains (e.g., global climatic model simulation). It also allows to implement the lineage of various dataset following the different steps of processing.

With the help of the above-mentioned ontologies and vocabularies, a metadata model is built, and it is centered towards below-mentioned contributions.

3.1.1. Using the FOI and UFOI concepts to enable domain-neutral dataset searches

One of the main advantages of considering the SOSA model is that it makes it possible to represent the collected observations using the FeatureOfInterest (FOI) and UltimateFeatureOfInterest (UFOI) concepts. They represent respectively the actual entity (i.e., sample) that has been measured and, the broader real-world phenomenon that has been sampled. Using this distinction, the expert perspective of an observation can be related to its thematic perspective that better represents the general user's viewpoint. An FOI is usually centered towards a certain domain (i.e., its understanding may require a high level of expertise),

[3] IRI: Internationalized Resource Identifier.

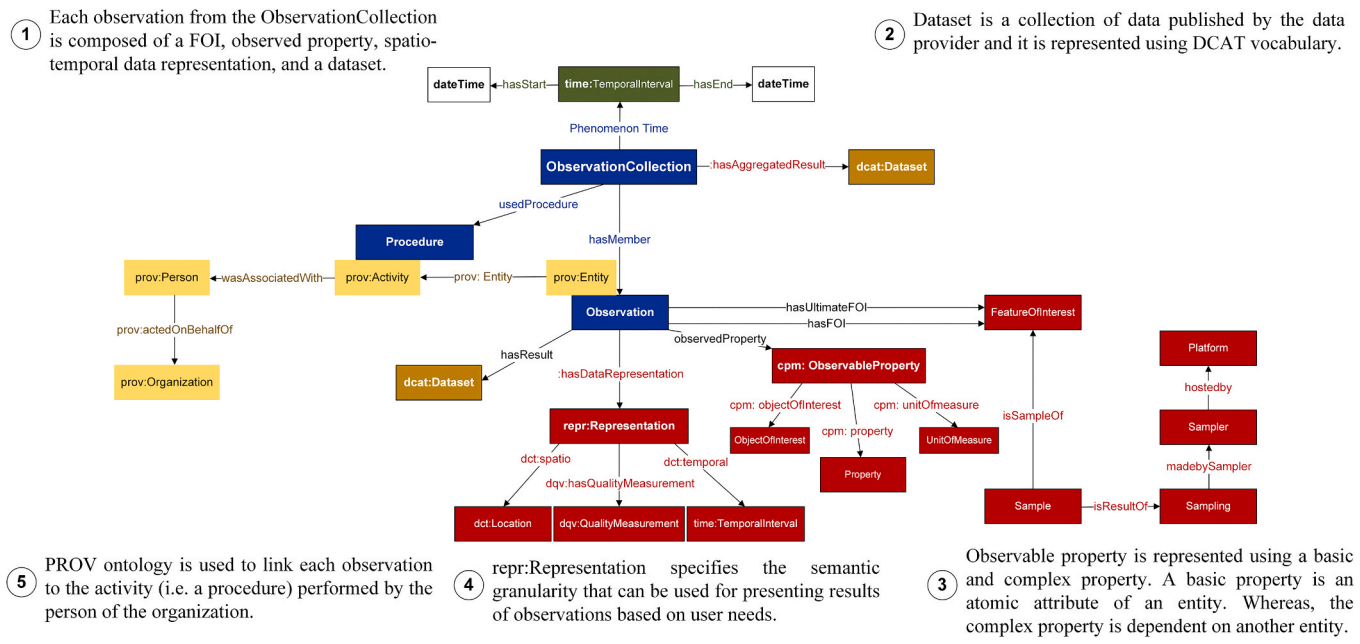

Fig. 5. Proposed metadata model. Prefix name spaces,: http://example.org/; cpm:http://purl.org/voc/cpm; dcat:http://www.w3.org/ns/dcat#; dct:http://purl.org/dc/terms/; dqv:http://www.w3.org/ns/dqv#; ; repr:http://sweetontology.net/repr/; sosa:http://www.w3.org/ns/sosa/; time:http://www.w3.org/2006/time#.

whereas the UFOI represents a general concept of environmental and life sciences indicating the topic or theme of an observation. Due to the generality of the UFOI concepts, high-level expertise is not required to understand it. This makes it easier to identify relevant datasets for discipline since it is not related to the expertise area of a researcher.

For instance, when considering a dataset related to the height of the water in a river measured by a hydrometric station,[4] the FOI for a hydrologist is the water height in the particular area around the station. However, the expert implicitly knows that the measures produced by this station are related to a sample of the river and, therefore, he knows that these measures can be used for studying all phenomena that are influenced by the river. For instance, this data can be used by a biologist interested in exploring the evolution of a given plant species in the proximity of the river since the river floods may influence it. If the fact that the sample is related to the river is explicitly stated, then a biologist can easily retrieve and reuse the dataset. This knowledge can be specified in our model using an UFOI that makes explicit the thematic context of the observation. In this way, scientists of other disciplines can easily understand the semantic content of a dataset through concepts representing general themes of environmental and life sciences. Moreover, the concept of a sample can also be useful for extrapolating general information on the value contained in a dataset, i.e., the maximum and minimum value when considering a dataset containing the temperature of the atmosphere.

3.1.2. *Use of complex properties for an improved dataset discovery*

To further endorse the data reuse across different disciplines, we have extended the Observable Property of the SOSA model using the concept of complex properties using the CPM ontology (Leadbetter and Vodden, 2016). A basic observable property is an atomic attribute of an entity, while a complex property is an attribute that defines the narrower objects of interests each having their own set of properties using constraint, matrix, property, statistical measure values, etc. For instance, considering the ocean as a FOI, its temperature is a basic observable property, while its chlorophyll concentration is a complex property (the concentration is related to the entity chlorophyll). Complex properties are used to breakdown complex concepts into "atomic" concepts (object, property, unit of measure …). This separation provides a clarification which enables more accurate discovery using FOI, observable property or matrix independently of each other. For instance, without considering complex properties, a researcher that studies the impact of water quality on the evolution of yellowfin tuna in the Indian Ocean would annotate the part of dataset related to water quality. It indicates the Indian Ocean as a feature of interest and chlorophyll concentration as property. However, the dataset could not be discovered by a researcher interested in analyzing the evolution of the chlorophyll (i.e., his FOI) in a given area. One of the main purposes is to increase the accuracy in the dataset discovery process.

3.1.3. *Adding spatial and temporal dataset granularity for discovery process*

Another objective of the user-centric metadata model is to encode the user-centric spatio-temporal viewpoint in the metadata model. To achieve this, a new class for data representation is required to be defined for specifying the most suitable dataset representation for the end-users as each stakeholder may have different data representation needs. Hence, we have added a Representation class (i.e., taken from the SWEET ontology). The purpose of this class is to specify the semantic granularity to be used for presenting the results of observations as per user needs. For instance, a user may prefer to combine the results concerning administrative units such as regions. Since spatial and temporal dimensions result to be essential dimensions also for presenting data to end-users, the Representation class is characterized by a Spatial Unit and Temporal Unit classes. The former indicates which spatial representation to adopt, i.e., administrative regions, land cover, grids, etc. The latter indicates which temporal representation to use, i.e., daily, monthly, annual, etc. The spatio-temporal dimensions are added using the GeoSPARQL and OWL-Time ontologies.

With the help of the above-mentioned extensions to the observation model (i.e., SOSA), the proposed metadata model provides a foundation to enhance the discovery of different datasets for conducting multidisciplinary research studies. In the next section, the metadata model is instantiated, and its application is discussed using a real-life scenario.

[4] It is a station on a river, lake, estuary, or reservoir that collects and records water quantity and quality data.

3.2. A practical example

This section presents a real-word use-case (see Figs. 6–9) that is used for instantiating the user-centric metadata model and shows its application for improving multi-disciplinary dataset searches. As an example, we consider a scientist that is an information architect working in the marine domain is interested to understand the phenomena of the Yellowfin tuna population evolution in the Indian Ocean in 2004. For addressing this research question, he manipulated both biological and environmental data. He must enrich in-situ biological observations (dated and localized) related to Yellowfin tuna catches, with remote sensing products of environmental parameters, such as the Sea Surface Temperature (SST) and the Chlorophyll concentration (CHLA). First, he decided which datasets to exploit based on his experience. He extracted Yellowfin catches data from a Structured Query Language (SQL) multi-dimensional data warehouse and converts it in a NetCDF format. Then, he downloaded the NetCDF files from NOAA[5] and NASA[6] Threads Data Servers (TDS) for SST and CHLA respectively. At this point, he converted all data to the same temporal and spatial resolution (i.e., 5°/monthly) to manually aggregate the data using the temporal and spatial dimensions. Thanks to his experience, the researcher was able to conduct his study. However, a user without his expertise may have difficulties in conducting this procedure consulting only current metadata. The metadata of Tuna catches[7] contains detailed information that is difficult to transfer when aggregating data in a new dataset. Moreover, the keyword metadata element does not distinguish the semantics of its values. Also, the metadata of SST and CHLA datasets, contained in the NetCDF files, has the same issue. Their keywords are based on the NASA Global Change Master Directory[8] (GCMD) Science Keywords and the CF-conventions and, for the SST and CHLA datasets, they are Oceans > Ocean Temperature > Sea Surface Temperature and Earth Science > Oceans > Ocean Chemistry > Pigments > Chlorophyll; Earth Science > Oceans > Ocean Chemistry > Chlorophyll, respectively. They, therefore, describe the semantic content of dataset. However, they do not explicitly distinguish between the observed entity or a phenomenon and the observed property. They result to be human-understandable but not machine-understandable, making difficult the automatic retrieval of this dataset. The researcher published the aggregated dataset providing ISO-19115 metadata. However, they do not explicitly refer to the Indian Ocean. Only the spatial extent in the form of coordinates is reported. This means that a non-expert user that does not know the coordinates cannot identify this dataset only typing the name of the area of interest. In such use-cases, the limitation of current metadata can be addressed using the user-centric metadata model that could facilitate the understanding of dataset contents by non-experts characterizing datasets using the observation framework (i.e., covering different key aspects of multiple disciplines). Moreover, it could also extend the keyword-based search of datasets to a semantic search leveraging the knowledge contained in the ontologies. Additional properties can be used to detail the sosa:Procedure, the sosa:Sensor and the sosa:Platform description. We omit them to facilitate the reading. The complete version of the metadata presented in the example is available at https://doi.org/10.23708/KMJ4CC. We privilege AGROVOC terms for representing hydrology concepts, and TAXREF-LD terms for biological concepts. In the next section, we discuss the main advantages of the proposed model and its limitations.

For the working example as described above, to specify the values associated with metadata elements, we have reused domain ontoterminology. For instance, GEMET[9] and AGROVOC[10] (Caracciolo et al., 2013) from Agroportal, EnvThes[11] (Schentz et al., 2013) from LTER. The complete metadata model is available at https://doi.org/10.23708/KMJ4CC.

Another example of the proposed metadata model is constructed using the data of Observation and Monitoring Network for Phytoplankton and Hydrology in coastal waters. The dataset (Cellule d'administration Quadrige,[2] 2017) contains several files of distinct parameters measured between 2006 and 2016 at the French surveillance sites (called REPHY network). However, to give the proof-of-concept illustration of the proposed model, only observations related to the concentration measurements of Pheopigment (i.e. a pigment which is the degradation product of algal chlorophyll pigments and commonly formed during and after marine phytoplankton blooms (Roy et al., 2011) is discussed in the provided Turtle script. Based on our model parameters, Ocean is defined as the Ultimate Feature of Interest. The concept "Ocean" is enriched with spatial information i.e. Mediterranean Sea that is further semantically linked with the Atlantic Ocean. Also, the observable property is defined using the CPM ontology, which links the concept "Pigments" to Object of Interest, whereas, Biomass to the Property of Object of Interest. The complete example is available at https://doi.org/10.23708/G4OM43.

4. Discussion

The first advantage of the user-centric metadata model is that it explicitly describes the semantic content of a dataset using a set of aspects that cover the different focus of several disciplines. This permits us to have a simplified overview of the dataset that enables scientists of different disciplines to evaluate the relevance of a dataset without being lost by technical details. Moreover, this description can be used for any observational datasets, independently of the types of their observations (in-situ, remote sensing, model outputs, or survey datasets) and their context (what, where, when and, how). Another advantage is the distinction between the FOI and UFOI and the use of the metadata values contained in ontologies. Since the sampled phenomena or an entity is explicitly stated, related phenomena or entities can be automatically identified using the knowledge expressed in our ontology. For instance, the dataset reporting observations of Yellowfin tuna catches can be discovered by a researcher interested in any tuna species, even if he does not explicitly state a specific tuna species. The ontology contains the information that Yellowfin tuna is a species of tuna and, therefore, the system based on the semantic metadata model can retrieve the Yellowfin dataset. Similarly, the user-centric metadata model allows the user to not specify the components of the Upper Oueme River Basin. This knowledge is automatically retrieved in the ontology. This means that even non-experts will be able to retrieve all relevant datasets for a certain phenomenon or an entity without having to specify all the exact terms to search. Instead, this is required when using current metadata models. Moreover, the user-centric metadata model introduces the possibility to indicate the most suitable data representation for the end-user. While traditional metadata specifies the spatial and temporal resolution and extent, the proposed metadata also indicates the semantics of the spatial and temporal representation expected by users for combining and representing data. For instance, a user may prefer to represent collected data based on administrative units rather than grids with a given resolution. In this case, the semantics of the spatial dimension is the administrative unit. Although several advantages of the model facilitate interdisciplinary research, limitations arise as well. First, an effort is required to manually annotate the datasets to add

[5] ftp://ftp.nodc.noaa.gov/pub/data.nodc/pathfinder/Version5.2/.
[6] https://oceancolor.gsfc.nasa.gov/l3/.
[7] https://tunaatlas.d4science.org/geonetwork/srv/eng/catalog.search#/metadata/global_catch_1deg_1m_ps_bb_tunaatlasIRD_level2.
[8] Global Change Master Directory Keywords, https://wiki.earthdata.nasa.gov/display/gcmdkey.
[9] https://www.eionet.europa.eu/gemet/en/themes/.
[10] http://aims.fao.org/vest-registry/vocabularies/agrovoc.
[11] http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn#http%3A%2F%2Fvocabs.lter-europe.net%2FEnvThes%2F21447.

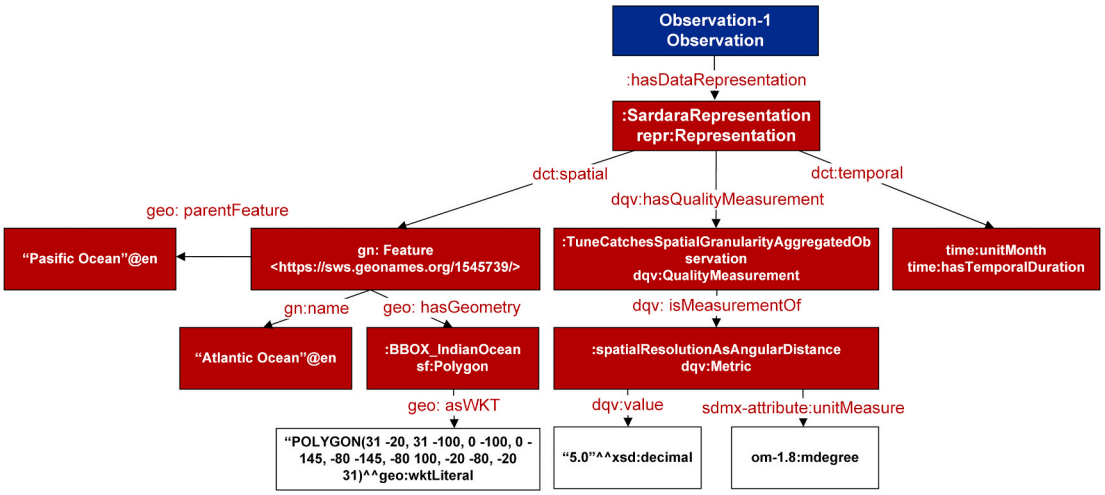

Fig. 6. Spatio-temporal representation along with the quality measurement for an Observation-1.

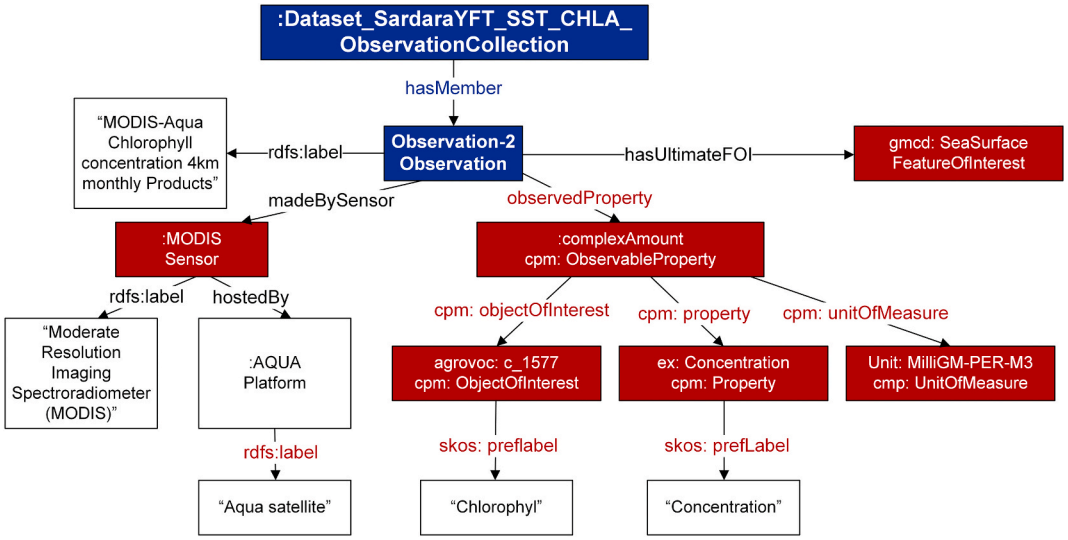

Fig. 7. Complex property representation for an Observation-2.

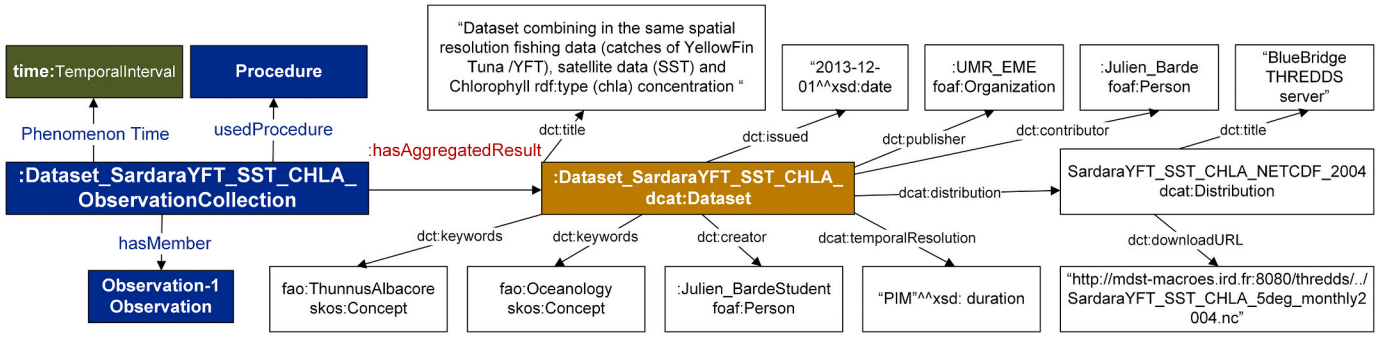

Fig. 8. Dataset description using DCAT vocabulary.

descriptions of it. For our case, the dataset was linked to a limited set of observations. Whereas, for annotating the bigger datasets the required effort will be higher. The semantic content is not described through the single element keywords but through a set of aspects. Second, different communities may prefer different ontologies and vocabularies. Thus, mappings between these resources are required. Even considering these limitations, we still consider the metadata model useful. It will increase the visibility of a dataset since a larger set of users (also experts of other disciplines) can discover it. Moreover, automatic models can be adopted for addressing the problem of ontology mappings (Euzenat and Shvaiko, 2007).

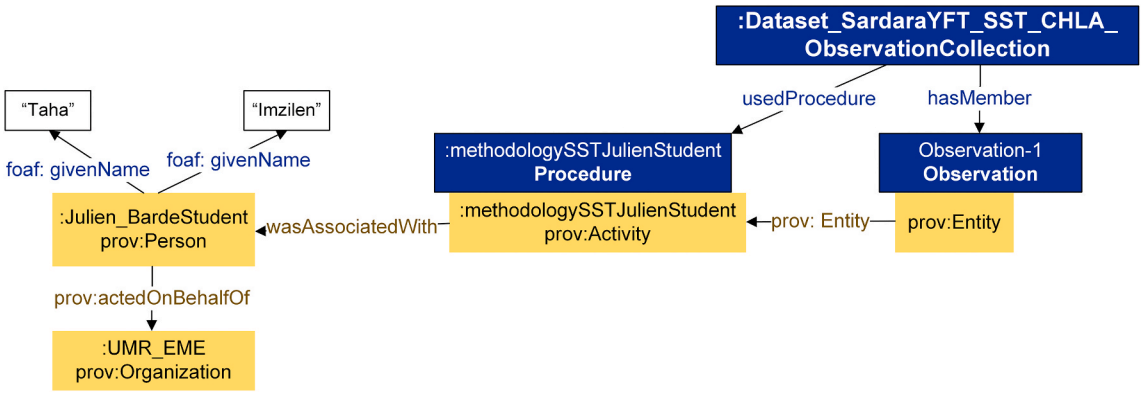

Fig. 9. Dataset provenance details.

5. Conclusion

Researchers conducting interdisciplinary studies in environmental and life sciences nowadays encounter difficulties for discovering existing datasets and evaluating their relevance. For this, we described a user-centric metadata model to foster multidisciplinary data discovery, access, and reuse. This study comes along with international initiatives such as Google Dataset Search (Halevy et al., 2016), schema.org (Patel-Schneider, 2014), its extension Bioschemas (Gray et al., 2017), and (DataCite Metadata Working Group, 2016). They promote the discovery and access of datasets in the Web of data through new search methods, the use of vocabularies and interlinked data. Indeed, traditional search techniques, applied for discovering relevant web pages in the Web of documents, are no longer effective. Moreover, current metadata does not provide clear and complete contextual information for fostering dataset reuse. This problem, the user-centric metadata model complements the technical description, already detailed by existing metadata models, with the thematic description that specifies the semantics of the involved entities with a higher level of abstraction. We formalize the model reusing as much as possible existing models, ontologies and vocabularies following the FAIR principles. In the near future, we will develop micro-services, which will make it possible to verify the validity and completeness of the description of the datasets, and consequently improve the dataset discovery process. To this end, we will define conformance rules using the SHACL (Shapes Constraint Language) language (Knublauch & Knublauch, 2017). The real-world use-case we present show the applicability of the metadata model that is available at https://doi.org/10.23708/FXIYQL.

Author contributions

V. Beretta, J.C. Desconnets, and I. Mougenot contributed to the conception, design, analysis and implementation of the metadata model introduced in this study. V. Beretta, J.C. Desconnets, and I. Mougenot took the lead in writing the manuscript. M. Arslan was responsible for improving the quality of figures and refining the text in the document during the revision. J. Barde and V. Chaffard provided real-world use-cases (and related datasets) helping shape the analysis and the validity of this study. Moreover, they also gave input and feedback for the manuscript. All authors gave their final approval of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the IRD that founded this project. We would like also to acknowledge researchers that enrich this work describing us their experience in multi-disciplinary studies. Thanks to Benjamin Sultan, Dimitri Defrance, and Thibault Catry (UMR 228 Espace-Dev, IRD), Anne-Sophie Archambeau (GBIF France), Benjamin Roche (UMMISCO, IRD), Jessica Abbate (TransVIHMI–MIVEGEC–CEPED, IRD), Charlotte Tollenaere (IPPS, IRD), Isabelle Braud (UR HHLY, IRSTEA), Sylvie Galle and Charly Coussot (PHyREV, IGE).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2021.104807.

References

Albertoni, R., Browning, D., Cox, S., Gonzalez Beltran, A., Perego, A., Winstanley, P., 04 February 2020. Data Catalog Vocabulary (DCAT) - Version 2. W3C Recommendation. https://www.w3.org/TR/vocab-dcat-2/.

Argent, R., 2004. Concepts, methods and applications in environmental model integration. Environ. Model. Software 19 (3), 217.

Battrick, B., 2005. Global Earth Observation System of Systems GEOSS: 10-year Implementation Plan Reference Document; Ad-Hoc Group on Earth Observations; Final Draft. ESA Publ. Division.

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J., 2013. The Agrovoc linked dataset. Semantic Web 4 (3), 341–348.

Cellule d'Administration Quadrige, 2017. Données du réseau de surveillance du REPHY. IFREMER. https://doi.org/10.12770/c5dd9e6f-b45f-4cd6-984d-95d13c8d1f1f, 2017.

Cox, S., 2010. Observations and Measurements. OGC Abstract Specification Topic 20. Open Geospatial Consortium document 10-004r3, 2010. http://portal.opengeospatial.org/files/41579.

Cox, S., 2013. An Explicit OWL Representation of ISO/OGC Observations and Measurements. SSN@ ISWC 1063, pp. 1–18.

Cox, S., 2020. Extensions to the Semantic Sensor Network Ontology. W3C Recommendation. https://www.w3.org/TR/2020/WD-vocab-ssn-ext-20200116/.

Cox, S., Little, C., 2020. Time Ontology in OWL. W3C recommendation. https://www.w3.org/TR/2020/CR-owl-time-20200326/.

Crockford, D., 2006. The Application/jsonmedia Type for Javascript Object Notation (Json). Technical report.

Danko, D.M., 2005. Iso/tc211: geographic information–metadata iso 19115. In: World Spatial Metadata Standards. Elsevier, pp. 535–555.

DataCite Metadata Working Group, 2016. DataCite Metadata Schema for the Publication and Citation of Research Data. DataCite e.V. https://doi.org/10.5438/0013. Version 4.0.

DCMI Usage Board, 2020. DCMI Metadata Terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

Desconnets, J.-C., 2017. Recherche D'information Spatio-Temporelle : Application Aux Images Satellitaires. Habilitation à diriger des recherches, Université de Montpellier. https://hal.archives-ouvertes.fr/tel-01649173.

DiGiuseppe, N., Pouchard, L.C., Noy, N.F., 2014. SWEET ontology coverage for earth system sciences. Earth Sci. India 7 (4), 249–264.

Donner, R., Barbosa, S., Kurths, J., Marwan, N., Jul 2009. Understanding the earth as a complex system - recent advances in data analysis and modelling in earth sciences. Eur. Phys. J. Spec. Top. 174 (1), 1–9. https://doi.org/10.1140/epjst/. ISSN 1951-6401. URL.

Edwards, J.L., 2004. Research and societal benefits of the global biodiversity information facility. Bioscience 54 (6), 485–486.
ESRI, 1998. Shapefile Technical Description. White Paper.
Euzenat, J., Shvaiko, P., 2007. Ontology Matching, vol. 18. Springer, Heidelberg.
Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M., 2005. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bull. Ecol. Soc. Am. 86 (3), 158–168.
Franklin, J.F., Bledsoe, C.S., Callahan, J.T., 1990. Contributions of the long term ecological research program. Bioscience 40 (7), 509–523.
Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D.J., Heber, G., 2005. Scientific data management in the coming decade. Acm Sigmod Record 34 (4), 34–41.
Gray, A.J., Goble, C.A., Jimenez, R., 2017. Bioschemas: from potato salad to protein annotation. In: International Semantic Web Conference (Posters, Demos & Industry Tracks).
Knublauch, H., Knublauch, D., July 2017. Shapes constraint language (SHACL). In: W3C Recommendation, w3c. https://www.w3.org/TR/shacl.
Halevy, A., Korn, F., Noy, N.F., Olston, C., Polyzotis, N., Roy, S., Whang, S.E., 2016. Goods: organizing google's datasets. In: Proceedings of the 2016 International Conference on Management of Data. ACM, pp. 795–806.
Hey, T., Trefethen, A., 2003. The Data Deluge: an E-Science Perspective. Grid Computing: Making the Global Infrastructure a Reality, pp. 809–824.
Janowicz, K., Haller, A., Cox, S.J., Le Phuoc, D., Lefrançois, M., 2019. SOSA: a lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics 56, 1–10.
Kashyap, V., Sheth, A., 1998. Semantic heterogeneity in global information systems: the role of metadata, context and ontologies. Cooperative information systems: Current trends and directions 139–178.
Kelling, S., Hochachka, W.M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., Hooker, G., 2009. Data-intensive science: a new paradigm for biodiversity studies. Bioscience 59 (7), 613–620.
Klyne, G., Carroll, J.J., 2006. Resource Description Framework (RDF): Concepts and Abstract Synta. https://www.w3.org/TR/rdf11-concepts/.
Knublauch, H., Allemang, H., Steyskal, S., 2017. SHACL Advanced Features. https://www.w3.org/TR/shacl-af/.
Leadbetter, A.M., Vodden, P.N., 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.
Lebo, T., Sahoo, S., McGuinness, D., 2013. PROV-O: the PROV Ontology. https://www.w3.org/TR/prov-o/.
Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. Ecol. Inf. 2 (3), 279–296.
National Information Standards Organization, 2016. Understanding Metadata. Technical report. University City Science Center, Philadelphia, PA. https://doi.org/10.5438/0012. Version, 4.
Patel-Schneider, P.F., 2014. Analyzing schema.org. In: International Semantic Web Conference. Springer, pp. 261–276.
Qin, J., D'ignazio, J., 2010. The central role of metadata in a science data literacy course. J. Libr. Metadata 10 (2–3), 188–204.
Rahimi, S., Roodposhti, M.S., Abbaspour, R.A., 2014. Using combined AHP–genetic algorithm in artificial groundwater recharge site selection of Gareh Bygone Plain, Iran. Environmental earth sciences 72 (6), 1979–1992.
Rew, R., Netcdf, G. Davis, 1990. An interface for scientific data access. IEEE computer graphics and applications 10 (4), 76–82.
Ritter, N., Ruth, M., 1997. The GeoTiff data interchange standard for raster geographic images. Int. J. Rem. Sens. 18 (7), 1637–1647.
Roy, S., Llewellyn, C.A., Egeland, E.S., Johnsen, G., 2011. Phytoplankton Pigments: Characterization, Chemotaxonomy and Applications in Oceanography. Cambridge University Press.
Schentz, H., Peterseil, J., Bertrand, N., 2013. Envthes-interlinked thesaurus for long term ecological research, monitoring, and experiments. In: EnviroInfo, pp. 824–832.
Shotton, D., Peroni, S., 2018. CiTO, the Citation Typing Ontology. https://sparontologies.github.io/cito/current/cito.html.
Tennis, J., 2015. Metadata application profiles. Encyclopedia of Archival Concepts, Principles, and Practices. (Rowman & Littlefield)., 2015.
Weibel, S., Kunze, J., Lagoze, C., Wolf, M., 1998. Dublin Core Metadata for Resource Discovery. RFC 2413. The Internet Society.
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3 (1), 1–9.