Implementing a new Research Data Alliance recommendation, the I-ADOPT framework, for the naming of environmental variables of continental surfaces

Coussot Charly, Chaffard Véronique, Boudevillain Brice, Sylvie Galle, Braud Isabelle

Earth Science Informatics
https://doi.org/10.1007/s12145-024-01373-9

RESEARCH

Implementing a new Research Data Alliance recommendation, the I-ADOPT framework, for the naming of environmental variables of continental surfaces

Coussot Charly[1] · **Braud Isabelle**[2] · **Chaffard Véronique**[3] · **Boudevillain Brice**[4] · **Sylvie Galle**[3]

Received: 31 October 2023 / Accepted: 12 June 2024

Abstract
To improve data usage in an interdisciplinary context, a clear understanding of the variables being measured is required for both humans and machines. In this paper, the I-ADOPT framework, which decomposes variable names into atomic elements, was tested within the context of continental surfaces and critical zone science, characterized by a large number and variety of observed environmental variables. We showed that the I-ADOPT framework can be used effectively to describe environmental variables with precision and that it was flexible enough to be used in the critical zone science context. Variable names can be documented in detail while allowing alignment with other ontologies or thesauri. We have identified difficulties in modeling complex variables, such as those monitoring fluxes between different environmental compartments and for variables monitoring ratios of physical quantities. We also showed that, for some variables, different decompositions were possible, which could make alignments with other ontologies and thesauri more difficult. The precision of variable names proved inadequate for data discovery services and a non-standard label (*SimplifiedLabel*) had to be defined for this purpose. In the context of open science and interdisciplinary research, the I-ADOPT framework has the potential to improve the interoperability of information systems and the use of data from various sources and disciplines.

Keywords Thesaurus · Ontologies · I-ADOPT · Complex variables · Metadata · Critical zone

Introduction

Context

Environmental knowledge is essential for understanding and modeling the functioning of the complex Earth system and for predicting its evolution in response to global environmental change. This knowledge is also essential for designing mitigation measures that will preserve the habitability of the Earth, which has now entered the Anthropocene (Crutzen 2002), and for addressing urgent societal needs related to the United Nations Sustainable Development Goals[1], as well as for monitoring and predicting risks. Environmental issues such as climate change, biodiversity loss, natural resource depletion, air, or water pollution, are complex and interrelated. These so-called "wicked" environmental problems (Rittel and Webber 1973) require interdisciplinary and transdisciplinary research that relies on data from different disciplines and sources (Parson

Communicated by Hassan Babaie.

✉ Coussot Charly
charly.coussot@ird.fr

Braud Isabelle
isabelle.braud@inrae.fr

Chaffard Véronique
veronique.chaffard@ird.fr

Boudevillain Brice
brice.boudevillain@univ-grenoble-alpes.fr

[1] Université Grenoble Alpes, IRD, CNRS, Météo-France, INRAE, OSUG, Grenoble 38000, France

[2] INRAE, RiverLy, Villeurbanne, France

[3] Université Grenoble Alpes, IRD, CNRS, Grenoble-INP, IGE, Grenoble 38000, France

[4] Université Grenoble Alpes, CNRS, IRD, Grenoble-INP, IGE, Grenoble 38000, France

[1] https://sdgs.un.org/goals.

Published online: 02 July 2024

et al. 2011) to understand the system and build integrated modeling tools using data-driven approaches (e.g. McDowell 2015; Bui 2016). Models will only provide reliable responses if they integrate all existing multidisciplinary data sources, which raises the question of the discoverability and accessibility of these data for new uses, not necessarily foreseen at the outset.

Open science has been proposed to promote access to data beyond the communities that produced them (Finkel et al. 2020) and is a key element in encouraging collaboration and accelerating the pace of scientific discovery and innovation (e.g. Mosconi et al. 2019). In this context, to mitigate the effects due to the heterogeneity of data sources and improve the transdisciplinary use of data, the FAIR principles have been defined (Wilkinson et al. 2016). FAIR stands for "Findable, Accessible, Interoperable, and Reusable" and designates a set of principles and best practices aimed at making data more useful and valuable to a broader community. The FAIR principles involve standardizing data description elements to make data findable, interoperable, and reusable by providing standardized metadata, referencing persistent identifiers, and offering clear descriptions of data content. The use of community standards such as Observations and Measurements (Cox 2011) is recommended by the FAIR principles, but standardized classes that accurately describe the acquired data can be freely instantiated, resulting in semantic discrepancies between descriptions (INSPIRE Maintenance and Implementation Group (MIG) 2016; Leadbetter and Vodden 2016). Variables[2] play a critical role in achieving interoperability between the various digital resources used in scientific workflows and are essential to meeting the interoperability challenge (Peckham et al. 2013; Stoica and Peckham 2019). To this end, Cox et al. (2021) suggest that describing metadata elements using terms from web-published vocabularies and unique, resolvable persistent identifiers allows not only wider communities of users, but also machines, to interpret data unambiguously.

However, environmental monitoring networks and related information systems have often been developed in disciplinary and community silos, with their own vocabularies, information systems, and practices, without considering the potential reuse of their data by people external to their communities (see the example of critical zone[3] observatories in France (Gaillardet et al. 2018; Braud et al. 2020). In the context of the French critical zone observatory network, the absence of a cross-community naming convention leads each observatory to use different naming conventions and vocabularies to describe a similar variable. In addition, naming conventions for observatory variables are defined in the context of a given observatory and are generally not specific enough to be used in a broader context. Users from different scientific domains sometimes need to identify details that may be implicit at the observatory context level. Having these details explicitly mentioned in the variable name allows users from other scientific communities to avoid having to explore data or analyze other metadata to find out whether the variable corresponds to their needs. For example, only precipitation volumes acquired at a given time step may be of interest to a scientist wishing to feed a particular model. Or again, only soil moisture measurements acquired at a specific depth close to the soil surface will be useful for calibrating remote sensing measurements. As a result, it is becoming increasingly difficult to combine observations[4] collected and described by different organizations despite the proximity of scientific domains with, ultimately, vocabularies operating in "silos" by discipline (Lausch et al. 2015). The proliferation of community-specific vocabularies, often difficult to align semantically (Campos et al. 2020), leads to degraded interoperability between vocabularies reducing system interoperability. To avoid this, several ontologies have been proposed for naming variables, in order to harmonize variable names using shared concepts. To improve interoperability, Leadbetter and Vodden (2016) suggest "breaking down the complex concepts into "atomic concepts" and identifying where the same atomic concepts are present in different domains, following the approach in Weinberger's (2002)". This approach, which implements Linked Open Data (Lausch et al. 2015), is tested in this paper.

Enhancing semantic interoperability in the Theia/OZCAR Information System (Theia/OZCAR IS)

The work presented in this paper was conducted as part of the construction of the Theia/OZCAR Information System (IS) (Braud et al. 2020). The Theia/OZCAR IS aims at facilitating the discovery and reuse of in-situ observational data of continental surfaces and the critical zone collected by French research organizations and their partners in France and other parts of the world, in particular in the OZCAR Research Infrastructure (RI) (Observatories of the Critical

[2] In the remainder of the paper, the term "variable" will be used. This term is similar to the term "observed properties" in other papers (e.g. Magagna et al. 2021), metadata standards such as Observations and Measurements (Cox 2011), and ontologies (e.g. http://www.w3.org/ns/sosa/observedProperty, http://purl.org/voc/cpm#ObservableProperty). This term is also equivalent to schema.org https://schema.org/variableMeasured.

[3] Earth's critical zone is the "heterogeneous, near-surface environment, from the bedrock to the top of the atmospheric boundary layer, in which complex interactions involving rock, soil, water, air, and living organisms regulate the natural habitat and determine the availability of life-sustaining resources (National Research Council 2001).

[4] In the remainder of the paper, the term "observation" can be defined after Observations and Measurements (Cox 2011) such as: "observation is an act (event), whose result is an estimate of the value of a property of the feature of interest".

Zone: Applications and Research) (Gaillardet et al. 2018). OZCAR-RI coordinates 22 observatories in about eighty sites around the world, operated by French research institutes. These observatories collect in-situ data documenting the various environmental compartments of the critical zone over the long term, with over 50 years of legacy data for some of them. These observatories monitor a wide variety of variables (including data from meteorology, hydrology, geomorphology, geology, hydrogeology, biogeochemistry, geophysics, pedology, microbiology, and ecology) and have historically developed their own data management systems. As a result, Theia/OZCAR IS has to handle very heterogeneous data descriptions and formats. The need to share data in a broader national and international context further complicates the situation. The Theia/OZCAR IS is part of the French Earth System DATA TERRA Research Infrastructure (Huynh et al. 2019), which aims to facilitate access to Earth system data and strengthen interdisciplinary research. At the international level, OZCAR RI also contributes to European infrastructures such as eLTER-RI[5] (European Long Term Ecosystem, critical zone and socio-ecological Research). The development and conception of the Theia/OZCAR IS therefore had to integrate this interdisciplinary context in the design of its tools. It was also necessary to consider data reuse by other communities. Consequently, the Theia/OZCAR IS had to provide semantic interoperability to enable data discovery and reuse for those communities.

To understand our user community's needs, a survey was conducted among the OZCAR RI scientists and revealed that the most important data search feature they required was the ability to search for data using variable names (Braud et al. 2020). However, as mentioned above, the different data producers used their own vocabularies resulting in incomplete search results due to the lack of semantic harmonization. The variable names of the data producers therefore needed to be harmonized and the first version of the Theia/OZCAR thesaurus was released in 2018 (Theia/OZCAR 2018). This version was built by establishing correspondence maps between variables from the different data producers and then deriving unique variable names from these correspondences. Those variable names were defined and organized in a thesaurus using the Simple Knowledge Organization System (SKOS) standard (Miles and Bechholder 2009). The thesaurus focused on data discovery and offered generic simplified labels for variable names to develop user-friendly data discovery services. Variable names were simplified by removing details that are often too specific or only relevant to a given producer. For example, the variable "soil moisture at 5 meters depth" was referred to as "soil moisture" in the thesaurus. These simplifications facilitated the alignment of variable concepts with other thesauri in the field (GCMD[6], EnvThes[7], AGROVOC[8], GEMET[9], AnaEE Thesaurus[10]).

Objectives

INSPIRE Maintenance and Implementation Group (MIG) (2016) argued that, when downloading the data, the IS user should be able to obtain enough information about the context in which the variable was acquired to have a general understanding of what was measured and the nature of the data, directly in the variable name. For example, a search for "soil moisture" will return soil moisture data series acquired at different depths, but users should be informed from the variable name that they are viewing "soil moisture at 5 meters depth". Furthermore, the use of simplified labels as variable names does not address the need for semantic interoperability between the Theia/OZCAR IS and other IS, such as DATA TERRA, or the eLTER IS mentioned above. With these simplified labels, different variables acquired in different contexts can be represented under the same label, and this ambiguity would propagate to other IS interfaced with Theia/OZCAR IS. To solve this problem, the variable terms used in the thesaurus need to be more specific in order to match the original variable name of the data producer and preserve the information they provided. However, this variable name will be too specific to find overlapping parts to map to other thesauri. Therefore, variable entries must be broken down into atomic elements for which semantic alignments will be possible (Leadbetter and Voden 2016).

The use of an ontology describing variable names was considered the best way to achieve this goal of both accuracy and interoperability. Ontologies describing variable names would serve as a useful basis for adding precise semantic annotations to scientific data, clarifying the inherent meaning of observational data. The objective of this work was the implementation of an existing ontology used as a framework for decomposing the variable names used in the Theia/OZCAR IS in order to enhance semantic interoperability. This objective was achieved through a review of existing ontologies used in environmental research with respect to the needs of the continental surfaces and critical zone community, as described in the next section. Following this analysis, the I-ADOPT framework was selected, but implementing its theoretical concepts in the "real world" was a challenge. This implementation of the framework for the set of variables of the Theia/OZCAR IS is described in the following

[5] https://elter-projects.org/.

[6] https://gcmd.earthdata.nasa.gov/.

[7] https://vocabs.lter-europe.net/envthes/en/.

[8] https://agrovoc.fao.org/browse/agrovoc/en/.

[9] https://www.eionet.europa.eu/gemet/en/themes/.

[10] https://agroportal.lirmm.fr/ontologies/ANAEETHES.

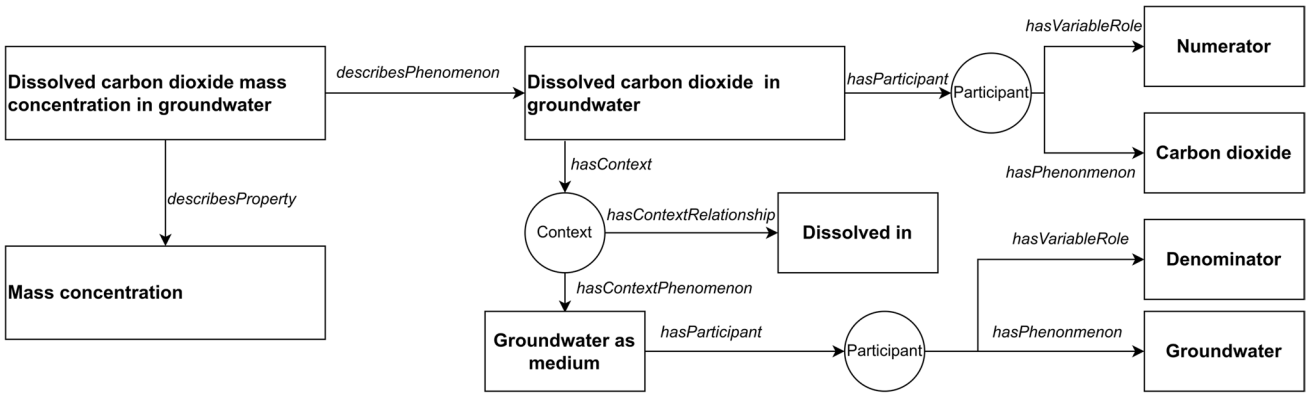

Fig. 1 Example of variable decomposition using SVO for the « Dissolved carbon dioxide mass concentration in groundwater » variable. It is described by the *Property* "Mass concentration" observing the *Phenomenon*"Dissolved carbon dioxide in groundwater". The *Phenomenon* is further described using *Inter-Phenomenon* classes which are classes used to link distinct phenomena together and create more complex *Phenomenon* systems. In this example, the complex *Phenomenon* "Dissolved carbon dioxide in groundwater" is contextualized using the *Phenomenon* "Groundwater as a medium" and is linked to the *Phenomenon* "Carbon dioxide"

section. The final part of the paper discusses the difficulties of using the I-ADOPT framework in the Theia/OZCAR IS, identifies solutions likely to alleviate them, and considers the benefits of this work in terms of interoperability.

Selection of the ontology used in the work

This section reviews four ontologies available to describe scientific variables in environmental research. There is a wide variety of environmental variables acquired in-situ to describe the critical zone (Gaillardet et al. 2018; Braud et al. 2020). While many of these are acquired by a large community and are well understood by the scientific community, others are more novel and generally involve more complex descriptions. These complex variables are not examined in this review since they are not representative of the variable set of Theia/OZCAR IS. Semantic decomposition involving such complex variables will be discussed later in this article. Instead, the four ontologies are evaluated using a single variable, chosen as a representative example of the level of complexity of the variables acquired in OZCAR-RI. This variable is "dissolved carbon dioxide mass concentration in groundwater".

Scientific variable ontology (SVO)

The Scientific variable ontology (SVO) was created to enable the unambiguous identification of scientific variables across different resources such as communications, structured digital data, or model inputs/outputs (Stoica et al. 2019). According to SVO, each scientific variable is composed of instances of each of the following core ontology classes:

- *Phenomenon*: anything (concrete) that can exist or occur in the physical universe and that has an independent existence. A distinct *Phenomenon* can be further described by a set of *Inter-Phenomenon* classes. *Interphenomenon* classes can be used to create more complex systems of *Phenomena*. Figure 1 shows the description of the "dissolved carbon dioxide" *Phenomenon*.
- *Property:* the observed property. It can be decomposed using a set of Property association classes. "Mass concentration" is the Property of the variable in Figure 1.

SVO consists of an upper-ontology that defines the elementary concept categories used to construct scientific variables and an extensible lower-ontology that contains the domain-specific instances of the classes defined in the upper-ontology. This ontology is therefore highly customizable and adapted for the fine description of complex variables. However, to take advantage of the precision of the ontology, we need to increase our knowledge of domain-specific instances of the ontology for the different families of variables in the Theia/OZCAR thesaurus. Customization would also be required for variables that cannot be described using the existing instances of the lower ontology. This would imply considerable effort and implementation time.

Complex property model (CPM)

The Complex Property Model (CPM) ontology is based on the INSPIRE extensions (INSPIRE Maintenance and Implementation Group (MIG) 2016) of the O&M model (Cox 2011). It provides a set of concepts for describing and linking complex environmental observations and observing systems. It was developed to facilitate the

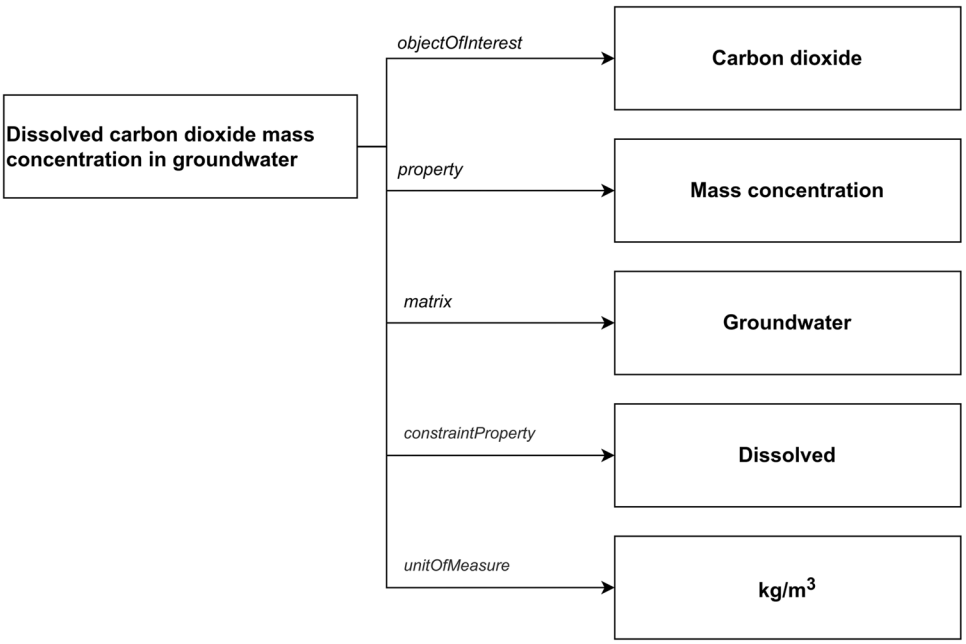

Fig. 2 Example of variable decomposition using the CPM ontology for the "dissolved carbon dioxide mass concentration in groundwater" variable

exchange and integration of data from various Earth Science domains (Leadbetter and Vodden 2016).

The *Observed Property* class described in the O&M model is an abstraction of the detailed phenomenon being measured. Detailed information on the phenomenon is considered part of the process and described in the *Procedure* class instance (INSPIRE Maintenance and Implementation Group (MIG) 2016). For example, high-frequency measurements of "dissolved carbon dioxide mass concentration in groundwater" could be described with "Mass concentration" as the *Observed property* and information describing the procedure of acquisition and the potential temporal aggregation would be described in the *Procedure* classes. This can be confusing to the scientific user, particularly during the data discovery process, as the procedure information is not usually presented in the foreground. The CPM ontology was created to expand the *Observed Property* and *Feature of Interest* components of O&M and to bring O&M *Observed Property* closer to what the user expects. A variable is described as an *Observable Property* in the CPM ontology and an *Observable property* is composed of at least the following elements:

- one *Object of Interest*, that is the feature of interest being observed (e.g. chemical species, environmental entities).
- one *Property*, that is the property being observed (e.g. temperature, concentration, height).

Additionally, an *Observable Property* can be decomposed using:

- a *Matrix*, that is a special feature of interest that provides context information for the *Observable Property* by documenting from where the O*bject of Interest* has been sampled. This can be the medium (e.g. air, water, soil) in which a *Property* is measured (e.g. concentration).
- a *Constraint* that provides constraint information for the *Observable Property*.
- a *Statistical Measure* that is used to describe statistical measures that are applied to the *Observable property* (e.g. daily maximum).
- a *Unit of Measure* that is the unit in which the *Property* is expressed (e.g. degree Celsius for the temperature property).

Figure 2 uses only five out of the six components provided by the CPM ontology since no *Statistical Measure* is relevant for decomposing the variable. *Statistical Measure* would have been used if the variable was representative of a value obtained after the application of a statistical method (e.g. average over a given period). As with SVO core classes *Phenomenon* and *Property*, only one *Object of Interest* and one *Property* are required to describe an *Observable Property*. Therefore, this ontology is lightweight and easy to implement. However, there is no mechanism to further contextualize complex variables. Information about contextual elements of the *Observable property* must be filled in using *Constraint* classes making them difficult to distinguish from other *Constraint* classes limiting the scope of the *Observable property*. Hence, for complex variables decomposed using CPM ontology, it is difficult to distinguish elements

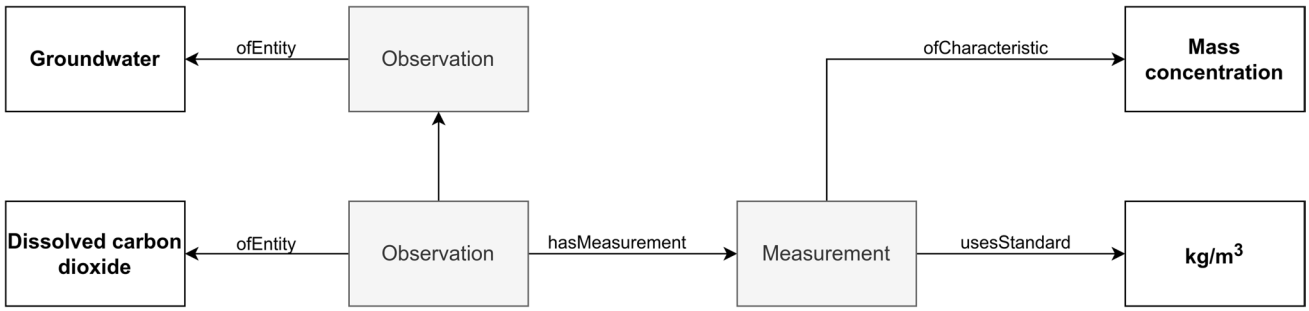

Fig. 3 Example of variable decomposition using the OBOE ontology for the "dissolved carbon dioxide mass concentration in groundwater" variable. The generic *hasContext* relation indicates that the measured *Entity* is part of the "Groundwater" *Entity*

providing additional context information from elements limiting the scope of the variable since those elements will be instantiated using the same *Constraint* class.

Extensible Observation Ontology (OBOE)

OBOE (Schildhauer et al. 2016) was developed in the context of ecological observation to describe the semantics of complex ecological data, although it can be used to generically describe scientific observations and measurements. It can be used to characterize contextual information surrounding an observation, such as location and time, and to document relationships between observations. In addition, the ontology allows for an accurate description of measurement units, including automatic conversions between units (Madin et al. 2007). The ontology is composed of four main classes. The *Observation* is the act of observing a particular *Entity*. An *Entity* is a generic class that represents all concrete and conceptual objects that are "observable". An *Observation* can be composed of several *Measurements*, which represent measurable *Characteristics* of the observed *Entity*. Figure 3 proposes a way to describe the variable "dissolved carbon dioxide mass concentration in groundwater" using OBOE. It involves the *hasContext* property which is used to reference a contextual relationship between *Entities* at the time of the *Observation*. This relation allows for exploiting the full potential of the ontology by describing the entire act of observation, such as the description of the spatio-temporal context of the observation.

Another Research Infrastructure (RI), AnaEE-France[11] (Mougin et al. 2015) develops services dedicated to the study of continental terrestrial and aquatic ecosystems. Its thematic areas concern biological diversity and the functioning of grassland, crop, forest, and lake ecosystems, relying on experimental platforms and the types of variables they deal with are close to the ones handled in Theia/OZCAR IS. AnaEE-France RI chose to use OBOE to perform a semantic decomposition of each of its variables. The variable decomposition consists of identifying an observed *Entity* on which a particular *Characteristic* is measured with a unit of measurement. Each variable has a name and categories derived from two naming standards (the AnaEE-France standard for naming variables and the AnaEE-France standard for naming variable categories). AnaEE-France information system (IS) extended the OBOE ontology by adding the *hasVariableContext* property, specializing the *hasContext* property, to link the variable standard name and categories to the observation while differentiating the result of the semantic decomposition of the observation in OBOE. The choice of using OBOE ontology to describe variables is appropriate for the AnaEE-France RI since the IS already uses this ontology to describe the observation context (Pichot et al. 2021). AnaEE-France IS can thus benefit from using a single model to document all the information related to the observations.

However, although the ontology is very generic, it is not the most effective when used in the context of variable description. There is no way to differentiate between the sampled matrix, the contextual information, or the constraints applied to the observation. Figure 3 shows that the sampled matrix of the variable is documented as another OBOE *Observation* using the *hasContext* relationship. Moreover, the "Dissolved" constraint, documented using the SVO and CPM ontologies in Figs. 1 and 2, is not instantiated using OBOE and relies on the *Entity* class instantiation.

Interoperable descriptions of observable property terminology framework (I-ADOPT framework)

The InteroperAble Descriptions of Observable Property Terminology working group (I-ADOPT working group 2021) of the Research Data Alliance (RDA)[12] has proposed the

[11] AnaEE (Analysis and Experimentations on Ecosystems): https://www.anaee-france.fr/en/.

[12] https://www.rd-alliance.org/.

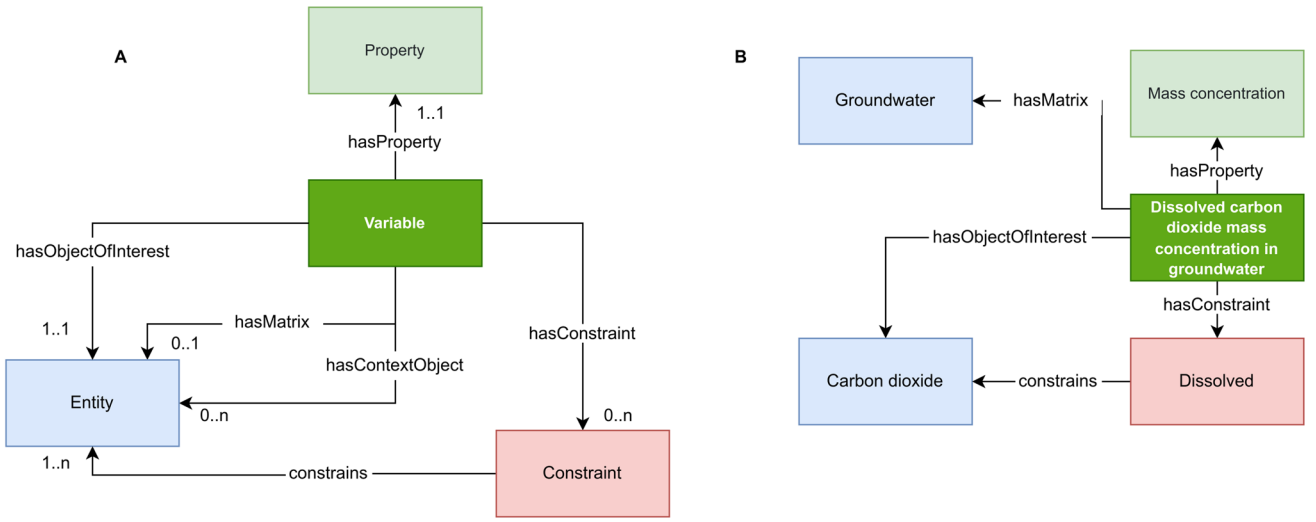

Fig. 4 **A**- The four classes and six relations of the I-ADOPT framework. A *Variable* is composed of exactly one *Entity* with the role *ObjectOfInterest* and one *Property* being measured. *Variable* often needs to be formalized using other *Entity* classes contextualizing the observation and *Constraint* classes confining the scope of the observation. Table 1 documents the definitions of the components of the framework. **B**- Example of variable decomposition using the I-ADOPT framework for the «dissolved carbon dioxide mass concentration per unit volume in groundwater » variable

I-ADOPT framework ontology. It is designed to improve interoperability between the various existing semantic models used to describe variable names, and to expand the usage of machine-readable variable descriptions (Magagna et al. 2021). The I-ADOPT framework is inspired by the atomization approach of the Complex Property Model (Leadbetter and Vodden 2016) and the SVO ontology and aims to describe the *WHAT* is observed and to some extent the *HOW* a variable is observed (Magagna et al. 2021). The semantic models presented above, which can be used to decompose variables into atomic elements, share similar components: an object being observed, and an observable characteristic of the object. The I-ADOPT framework decomposes each *Variable* using these two elements. The framework is also designed to allow the addition of restrictions to the observation and the addition of contextualization elements.

The I-ADOPT framework consists of four classes and six relations (Magagna et al. 2022). A *Variable* is composed of at least one *Entity* being observed and one *Property* being measured. Restrictions (i.e. precisions/details) on the observation can be described using one or several *Constraint* classes. Elements of the context of the observations are also documented using one or several *Entity* classes. The documentation defines an *Entity* as "an object or occurrence that has a role in an observation". An *Entity* may play one of the following roles: *ObjectOfInterest*, *ContextObject*, *Matrix*. Whether the involvement of a particular entity is meaningful enough to be included in the variable description depends on the particular observation. These roles can be defined in the context of an observation by associating *Entity* using three different relations; *hasObjectOfInterest* for the *Entity* whose *Property* is observed; *hasMatrix* for an *Entity* from which an *Entity* with the role *ObjectOfInterest* is sampled; *hasContextObject* for an *Entity* that provides additional background information regarding the *Entity* with the role *ObjectOfInterest*. Figure 4A shows the framework conceptual model and Fig. 4B presents the decomposition of the *Variable* "dissolved carbon dioxide mass concentration in groundwater". Table 1 provides the definitions of the framework conceptual model classes and relations.

Requirements of the Theia/OZCAR IS for adopting an ontology for decomposing variables

Several requirements have been identified for selecting an ontology to decompose Theia/OZCAR IS variables. Generally speaking, the ontology must fit in with the existing version of the vocabulary that is using SKOS. It must be usable for all Theia/OZCAR IS variables. The time required to implement the ontology for all variables must be taken into account. Regarding the data discovery services of the Theia/OZCAR IS, three requirements can be formulated.

- The ontology must enable the use of a simple label to support search services based on variable names.
- In the future, we would also like to use this work to develop search services on features of interest (i.e. entity that is of interest in the act of collecting data related to

Table 1 Definition of the 10 components of the I-ADOPT ontology according to the I-ADOPT Framework ontology (Magagna et al. 2023)

Class	**Definition**
Variable	A description of something observed or derived, minimally consisting of an *ObjectOfInterest* and its *Property*.
Property	A type of characteristic of the *ObjectOfInterest*.
Entity	An object or occurrence that has a role in an observation. An *Entity* may play one of the following roles: *ObjectOfInterest*, *ContextObject*, *Matrix*. Whether the involvement of a particular *Entity* is meaningful enough to include in the *Variable* description depends on the specific context.
Constraint	A *Constraint* limits the scope of the observation and restricts the context to a particular state. It describes the conditions of the involved *Entities* that are relevant to the particular observation.
Property	**Definition**
hasProperty	A *Variable* has a *Property* that characterizes an *Entity*.
hasObjectOfInterest	A *Variable* has an *Entity* whose *Property* is observed.
hasMatrix	A *Variable* might have an *Entity* in which the *ObjectOfInterest* is contained.
hasContextObject	A *Variable* has an *Entity* that provides additional background information regarding the *ObjectOfInterest*.
hasConstraint	A *Variable* has a *Constraint*, that confines an *Entity* involved in the observation.
constrains	A *Constraint* constrains an *Entity* having a role in the *Variable* description.

Table 2 Matrix summarizing whether the ontologies evaluated meet the Theia/OZCAR IS requirements

	SVO	OBOE	CPM	I-ADOPT
Combines with the existing version of the Theia/OZCAR thesaurus (combine with SKOS)	*Yes*	*Yes*	*Yes*	*Yes*
Can be used on the whole variable set	*Yes*	*Yes*	*Yes*	*Yes*
Has conveniences of implementation (considering skill development and implementation time)	*No*	*Yes*	*Yes*	*Yes*
Provides simple variable labels that can be used to support data discovery services	*No*	*No*	*No*	*No*
Can be used to support data discovery services using features of interest	*Yes, using Phenomenon instances*	*No, not enough semantics to differentiate Entity instances with the role of feature of interest from the others*	*Partially, using Matrix instances*	*Yes, using Entity instances*
Can be used to implement filtering services using variable constraints	*Yes*	*No, no classes are provided to describe constraints on variable*	*Yes*	*Yes*
Can be used to describe time steps of time series data and statistical measures on this time step	*Not evaluated*	*Not evaluated*	*Yes*	*No*
Is semantically rich enough to describe complex variables	*Yes*	*No, no classes are provided to describe constraints on variable*	*No*	*Yes*

a given variable), and filtering using time steps of time series data and other constraints applied to the variable.
- Finally, to promote system interoperability, the ontology must provides sufficiently rich semantics to decompose variables.

Table 2 compiles these requirements for the different ontologies evaluated.

The I-ADOPT framework was chosen to describe variables in Theia/OZCAR IS. Table 2 shows that the I-ADOPT framework best meets the requirements of Theia/OZCAR IS, although the ability of SVO to describe time steps of time series data has not been evaluated. The I-ADOPT framework was preferred to SVO because of its simplicity and its ease of implementation. Despite its simplicity and its lightness, the various roles that can be assigned to an *Entity* and the relation between *Constraint* and *Entity* make it suitable for describing complex variables. Also, the authors of the I-ADOPT framework provide alignments with other ontological frameworks (I-ADOPT working group 2023a) making the I-ADOPT framework directly compatible with SVO and CPM ontologies. The authors also provide alignments to O&M (O&M, OGC, and ISO 19156:2011) which is the core standard used by Theia/OZCAR IS to document observations made by data producers (data acquisition context, spatio-temporal context…) (Braud et al. 2020). Grellet et al.

(2021), support the choice of I-ADOPT framework, arguing that the framework can be used to complement O&M and to meet emerging needs to detail measured variables. Furthermore, the I-ADOPT framework is currently being implemented in the EnvThes environmental thesaurus (EnvThes 2023). Thus, the implementation of the I-ADOPT framework for OZCAR-RI variables will facilitate semantic alignments between concepts of the two thesauri and, to some extent, improve the interoperability of OZCAR-RI data on a European scale within the eLTER-RI data infrastructure which uses EnvThes (Wohner et al. 2022). The choice of the I-ADOPT farmework was also motivated by its endorsement as a RDA recommendation for describing scientific variables (Magagna et al. 2022) which ensures that its use is likely to increase in the future.

Two requirements are not met by using the I-ADOPT framework. The possibility to use simplified labels as variable names to provide data discovery services and the ability to describe temporal steps in time series data. None of the ontologies evaluated meet the first requirement. To this end, a *simplifiedLabel* property has been designed and is discussed further in this article. To meet the second requirement and describe time steps of time series data, we take advantage of the compatibility between the CPM ontology and the I-ADOPT framework. CPM ontology class *cpm: StatisticalMeasure* and properties *cpm: statisticalMeasure* and *cpm: aggregationTimePeriod* are used to describe time steps of time series data and statistical measures applied to time series data.

Method to implement the I-ADOPT framework in the Theia/OZCAR thesaurus

General methodology for modeling variables in Theia/OZCAR thesaurus

Before conducting this work of implementing the I-ADOPT framework in the Theia/OZCAR thesaurus, a previous version of the thesaurus already existed. As mentioned in the second section of this article, the initial thesaurus offered generic simplified labels for variable names and was used by Theia/OZCAR IS to provide data discovery services. This thesaurus was developed in accordance with the SKOS standard (Miles and Bechholder 2009). At the beginning of this work, more specific variables carrying specific information provided by the data producers had not yet been created in the thesaurus.

The definition of the atomic elements that compose the variables of the Theia/OZCAR IS was conducted by two scientists and two technical engineers from the OZCAR-RI network. Validation of the variable modeling was submitted to observatory scientists once all the variables from an observatory were modeled. Variable names need to be encompassing enough to allow observations to be grouped, and sufficiently discriminating to give the scientific user a clear idea of what is being measured. For this reason, variables in the Theia/OZCAR IS must identify the *Property* being measured, the *Entity* being observed, the environmental compartment from the critical zone involved if it cannot be known implicitly, and, if necessary, any constraints limiting the scope of the variable and contextual elements essential to understanding the observation (e.g. inside a borehole). Optionally, the statistical operations applied to the data and the time steps of the data series can be specified. Other information is not included in the variable. This information includes sensors (unless the *Property* being measured relates directly to the sensor, for instance, "tiltmeter sensor level"), units, or the ultimate feature of interest (e.g. the Amazon River, the Bernadouze peatland) of the observation. They are considered too discriminating and do not provide essential information for understanding the nature of what has been measured for data exploration purposes.

The creation and the modeling of a specific variable using the I-ADOPT framework in the Theia/OZCAR vocabulary follows these different steps:

1 – Create the *Variable* as a *skos: Concept* type with its full name described in the *skos: prefLabel* property and organize it in the thesaurus hierarchy using SKOS hierarchical relationships. Isaac and Summers (2009) provided an informative guide for users seeking to implement the SKOS standard. Qualify the *Variable* term as a *iadopt: Variable* using the I-ADOPT framework.

2 – Create or identify in another thesaurus the atomic terms to be used for modeling the *Variable*. In Theia/OZCAR IS atomic terms must be used to create discovery services (see the requirements in Table 1). For this reason, atomic terms composing a *Variable* are created in Theia/OZCAR thesaurus to get full control of them. However, reusing terms from existing vocabularies is another option for decomposing a *Variable*. As with the first step, each creation involves each new term being qualified as *skos: Concept* and organized in the vocabulary. New concepts must also be qualified with the I-ADOPT framework according to the *iadopt: Property*, *iadopt: Entity*, or *iadopt: Constraint*, depending on the role they play in modeling the variable. The following procedure can be used to identify the elements composing a *Variable*. The *Variable* "1 day mean dissolved carbon dioxide mass concentration in groundwater" is taken as an example:

a) The *Property* element is the easiest to identify because it is directly expressed in the observed values. Thus, observed values will be expressed in the dimensions of the property element. The I-ADOPT working group provides a "unit to property"[13] tool to facilitate this identifi-

[13] https://i-adopt.github.io/terminologies/unit2property/.

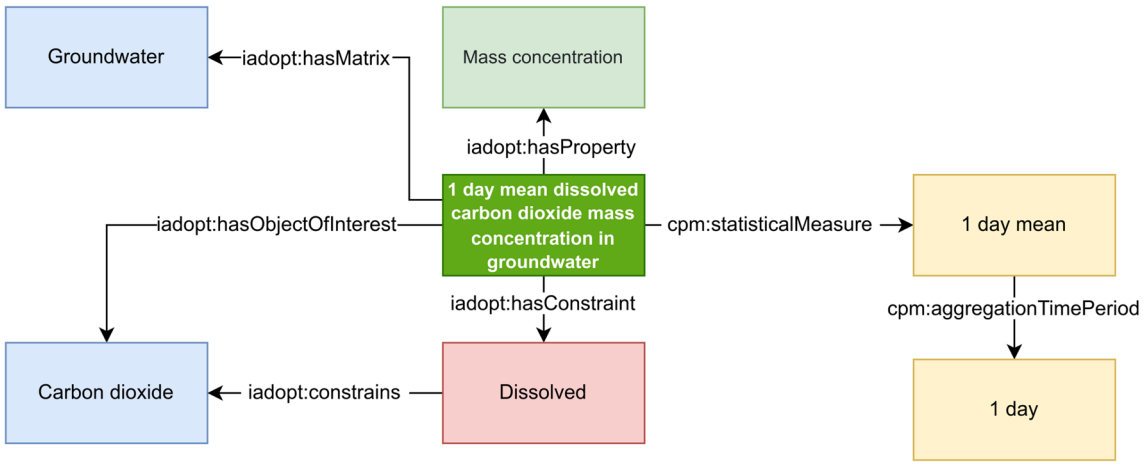

Fig. 5 Modeling of the time series variable for which data are aggregated over a time period. The description of the statistical aggregation is outside the scope of the I-ADOPT framework but the modeling can be extended using the CPM ontology for this purpose. The relationship *cpm: statisticalMeasure* is used to describe an aggregation over time and *cpm: aggregationTimePeriod* is used to describe the aggregation period

cation. Observed values related to the example variable can be expressed in kg/m^3 which can indicate that the *Property* is "Mass concentration".

b) The *ObjectOfInterest* entity can be identified once the *Property* element has been identified, as it corresponds to the *Entity* whose *Property* is observed. In our example, the concentration of "Carbon dioxide" is observed. So "Carbon dioxide" is the *Entity* that plays the role of *ObjectOfInterest*.

c) If the identified *ObjectOfInterest* is sampled from an *Entity* that contains it, the *Matrix* element can be identified. Here, the "Groundwater" *Entity* plays the role of *Matrix* since it is the *Entity* that contains the "Carbon dioxide" *Entity*.

d) If further *Entity* elements, that are different from the *Matrix*, are necessary to describe the *Variable*, *ContextObject* elements can be identified.

e) If the scope of any *Entity* identified in the previous step needs to be restricted, a *Constraint* element can be applied to an *Entity*. Here, only the "Carbon dioxide" that is dissolved in "Groundwater" is observed by the *Property*. So "Dissolved" is the *Constraint* applied to the "Carbon dioxide" *Entity*.

The I-ADOPT framework focuses on the observed variable. Information about the acquisition procedure is explicitly excluded from the I-ADOPT framework since it is not an intrinsic feature of the variable itself. However, Theia/OZCAR IS data are often time series data and we want to be able to provide the information about the temporal aggregation procedure, used to calculate a mean for example, in the variable names decomposition. The relations *cpm: statisticalMeasure* and *cpm: aggregationTimePeriod* of the CPM ontology (INSPIRE Maintenance and Implementation Group (MIG) 2016) are used for this purpose. Figure 5 presents the modeling of a time series *Variable* using the I-ADOPT framework in combination with the *cpm: StatisticalMeasure* class.

3 – Finalize the variable modeling by associating the elementary concepts with the *Variable* using the I-ADOPT framework relations *iadopt: hasProperty, iadopt: hasObjectOfInterest, iadopt: hasContextObject, iadopt: hasMatrix, iadopt: hasConstraint.* Appendix A provides a complete modeling of the variable represented in RDF/turtle format.

To make a thesaurus interoperable, similarity relations must be defined between terms in the thesaurus and terms in other thesauri and semantic alignments can be performed using SKOS associative relationships. This work does not have to be carried out each time a concept is created because batch operations can be performed for all concepts in the thesaurus. Also, special care must be taken to ensure that a definition will be attributed to each elementary concept in order to disambiguate the term. Definitions can be associated using the *skos: definition* property. Definitions from sources available online are recommended. As this work is intended for an interdisciplinary audience, more general definitions of concepts are preferred to increase the chance of finding similarities with other vocabularies. For chemical entities, ChEBI[14] definitions are preferred where available. When no satisfying definition is found online for a given concept, one has to be created with the help of a domain expert.

This work of creating and modeling a variable is time-consuming and tedious but some operations can be

[14] https://www.ebi.ac.uk/chebi/.

automated using SPARQL[15] queries when identifying patterns in variables. For example, « soil moisture » is measured at several depths and the modeling of the variables will follow the same pattern for each depth. The same approach can be used for the variable measuring the concentration of an element in a matrix. Appendix B provides an example of a SPARQL operation used to automatically create terms for variables describing the same measurements at different depths.

Software tools used

The thesaurus is managed using Vocbench 3 (Stellato et al. 2020). It is used to facilitate the creation of the terms of the thesaurus and to perform *Variable* modeling using the I-ADOPT framework. Each vocabulary term, including variables and the atomic concepts used to describe them, are *skos: Concept* that were created and organized one by one using the Vocbench application. The ontologies used for the modeling were previously imported into the Vocbench application. The w3id.org redirection service[16]guarantees the persistence of Uniform Resource Identifiers (URI)[17] over time. The creation of new terms in the thesaurus involves numerous manual operations and the integrity of the vocabulary must be tested to reduce the risk of errors. To this end, the Vocbench application provides methods for validating integrity constraints. To check the integrity of the I-ADOPT framework modeling, Appendix C provides a SPARQL query that highlights all terms that are qualified with the *iadopt: Variable type* and that do not conform to the I-ADOPT framework modeling.

The OnAGUI desktop application (Mazuel and Charlet 2023) is used to perform batch operations for thesauri alignments. It identifies similarities between vocabularies, thesauri, or ontologies based on the proximity of the labels of each of their elements. The result of the alignment of two thesauri using the OnAGUI application can be obtained in Expressive and Declarative Ontology Alignment Language (EDOAL) (David et al. 2011). The thesauri alignment results expressed using EDOAL are then translated into the SKOS associative relationships using a set of features provided by the Vocbench application (Stellato et al. 2021).

The vocabulary is stored on a RDF4J triple store (Eclipse RDF4J 2021). A user-friendly visualization interface is proposed using Skosmos (Suominen et al. 2015). Although the Skosmos documentation suggests using the Jena triple store to improve query performance, the Skosmos application is compatible with generic SPARQL endpoints (Skosmos 2023). On the other hand, Vocbench requires the triplestore to expose the RDF4J Sail API for remote connection (Vocbench 2023). Therefore, the RDF4J triple store was chosen so the thesaurus management operation is performed on a remote triple store with changes visible directly on the Skosmos viewer interface.

Results of the implementation in Theia/OZCAR thesaurus

The results of variable modelings using the I-ADOPT framework are all available in Theia/OZCAR thesaurus[18]. At the time of writing, 2773 variables from 22 observatories of the OZCAR-RI were analyzed. The initial strategy adopted by the project team for this analysis relied on the variable names provided by the producer, together with its description and unit. This allowed us to model a large proportion of the variables. The modeling of specific variables, especially those requiring the *ContextObject* entity to be specified, required exchanges with the scientists in charge of the observatories.

At the time of writing, the 2773 variables have been harmonized into 969 I-ADOPT *Variable* and decomposed into 282 I-ADOPT *Entity* elements, 135 I-ADOPT *Property* elements, 116 I-ADOPT *Constraint* elements, 30 CPM *StatisticalMeasure* elements. These numbers will evolve as the thesaurus construction is an ongoing process.

The diversity of the I-ADOPT framework atomic terms arising from the variables has prompted us to categorize them into the following groups that represent the top-concepts of the vocabulary:

- “Physical entity”, that gathers chemical entities (e.g. carbon dioxide, oxygen), environmental entities (e.g. atmosphere, water table), structure (e.g. borehole), sample (e.g. ice core).
- “Phenomenon” (e.g. evapotranspiration, erosion, snowmelt).
- “Process” (e.g. ecosystem respiration).
- “Instrument” (e.g. gravimeter).
- “Method”, that groups experiments (e.g. groundwater pumping test) and statistical methods.
- “Time”, that groups different periods that are used for statistical aggregation.
- “Property” (e.g. temperature, concentration).
- “Constraint”, that group the constraints limiting the scope of the variables of this thesaurus.

[15] https://www.w3.org/TR/rdf-sparql-query/.

[16] https://w3id.org/.

[17] A Uniform Resource Identifier (URI) is a unique sequence of characters that identifies a logical or physical resource used by web technologies. https://en.wikipedia.org/wiki/Uniform_Resource_Identifier (accessed 9th January 2024).

[18] https://w3id.org/ozcar-theia.

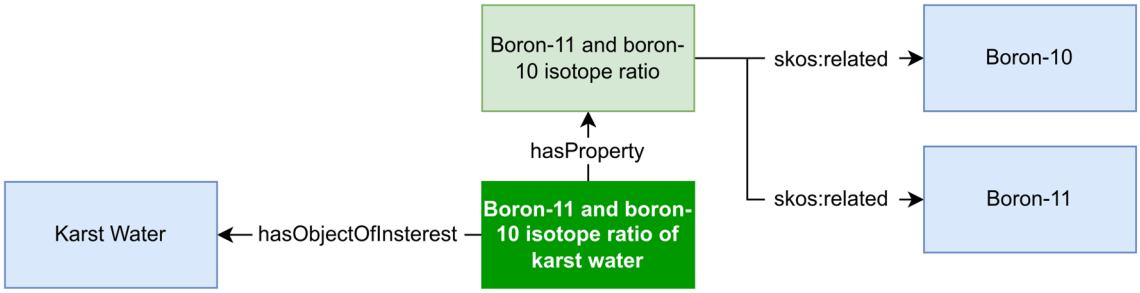

Fig. 6 Example of the decomposition of the variable measuring the isotope ratio of an entity. The *Property* is associated with the element of the ratio it is measuring using *skos: related* relationship

– "Variable", the concept under which all variables modeled during this work are organized.

Assigning definitions to each concept is a work in progress. *Entity* and *Property* concept definitions often rely on Wikipedia. Most of the chemical concepts are defined in ChEBI. The benefit of using Wikipedia lies in the existence of a structured and standardized version of its content in semantic web format known as DBpedia[19]. When DBpedia URIs are employed, navigation through these definitions is possible not only for humans but also for machines. Efforts are currently in progress within the Theia/OZCAR thesaurus to substitute all Wikipedia URLs with DBpedia URIs to take advantage of this. At the time of writing, no *Constraint* concept has been defined and scientists have not yet been involved to help define more specific concepts.

Variable names were created according to the information provided by the data producers. This work revealed that the levels of information are not homogeneous from one producer to another and within a given producer. Most of the variables from the Theia/OZCAR IS are related to time series data while only a few mention time steps of time series data in the variable name modeling. For complex variable modeling, validation by a scientist concerned with data interoperability seems more effective than validation by a scientific expert in the field alone.

Discussion

Modeling difficulties using the I-ADOPT framework

The detailed variable names described using the I-ADOPT modeling framework provide rich information that meets the interoperability and reusability needs of the Theia/OZCAR IS (Table 2). In the context of data discovery services, detailed variable names are often too complicated to be directly employed for data exploration. Therefore, each variable was associated with a generic term using the specially designed *simplifiedLabel* relationship. The simplified terms follow the design of the first version of the Theia/OZCAR vocabulary by limiting the precision of the term to an *Entity* and the *Property* with which it is observed, and if necessary specifying the critical zone compartment where the observation is made. The variable "dissolved carbon dioxide mass concentration in groundwater" is simplified to "groundwater carbon dioxide concentration".

The I-ADOPT framework modeling of the Theia/OZCAR variables was straightforward for most of the variables, and the tools proposed by the I-ADOPT working group (2023b) were useful in this regard (in particular the "unit to property" tool). However, some modeling tasks were more difficult and required special attention to document the variables accurately. We present below some difficulties encountered in modeling the *Property* and *Entity* respectively.

Difficulties related to the *Property* modeling

Particular attention was paid to the modeling of a *Variable* whose *Property* is directly related to an *Entity*. This case was generally encountered when modeling variables observing a ratio. A *Property* documenting a ratio must be specific enough to define both elements of the observed ratio. For example, porosity is the ratio between the volume of voids and the total volume of a material. The two elements of the fraction are the volume of voids and the total volume of material. When these elements are not intrinsically defined by the *Property*, it is necessary to create the specific *Property* that defines them. This is the case for the isotope ratio presented in Fig. 6 for which it is necessary to create a *Property* for each isotope. Both isotopes are linked to the isotopic ratio property using the *skos: related* property.

We can also expect to have to model very complex variables involving complex and uncommon properties. In this case, a problem identical to the one discussed in this article with variables would arise with *Properties* since complex properties would be too specific to be aligned with other

[19] https://www.dbpedia.org/.

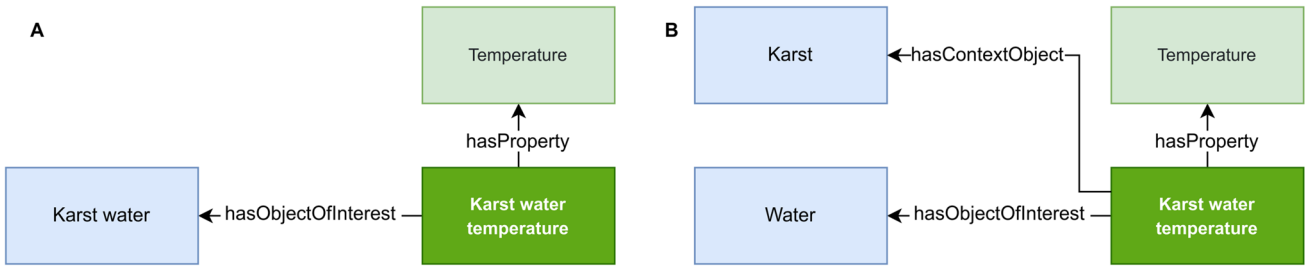

Fig. 7 Two possible modelings of the same *Variable*. In the solution retained in the Theia/OZCAR thesaurus (**A**), critical zone compartments are qualified as *Entity* to be used in *Variable* modeling. The second example (**B**) uses a more generic *ObjectOfInterest* and provides information on the critical zone compartment using the *ContextObject*

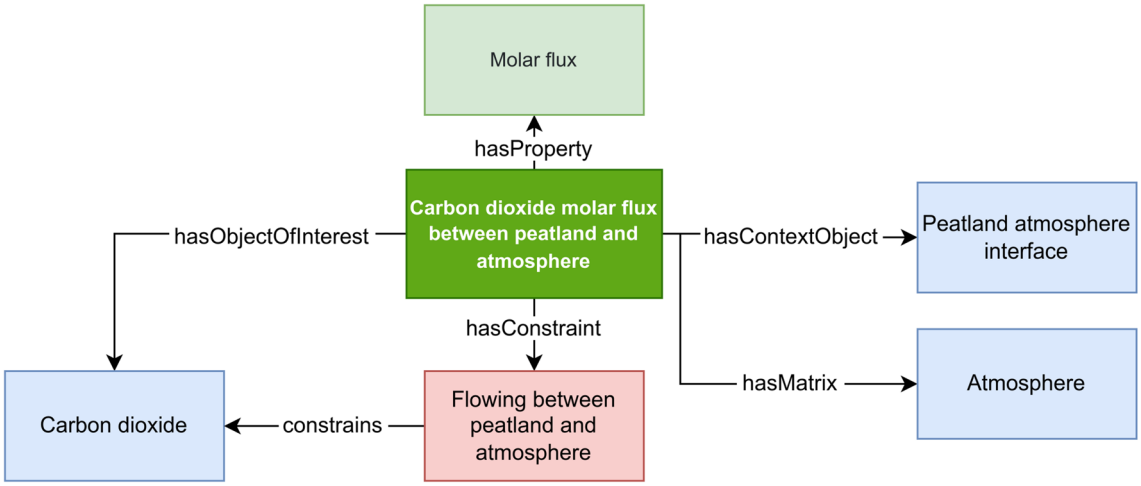

Fig. 8 Example of the decomposition of the variable measuring the flux of an entity at an interface. The matrix on which the variable is sampled is documented using the *hasMatrix* property, while the interface on which the observation is made is documented using the *hasContextObject* property. A *Constraint* is added to indicate that the *Entity* playing the role of *ObjectOfInterest* is flowing from one compartment to the other. The flow direction can also be indicated by the *Constraint*

vocabularies. Furthermore, it seems reasonable to think that identical *Properties* would be defined differently from one vocabulary to another. For example, the *Property* "speed" could be decomposed using its basic dimension elements and labeled as "distance per unit of time". So, further description of *Properties* using ontologies, such as the "Quantity, Unit, Dimension and Type" (QUDT) collection of ontologies[20], as suggested by Magagna (2021), could be evaluated to address this problem and facilitate *Property* identification and alignment. The QUDT ontology collection defines the base classes, properties, and restrictions used for modeling physical quantities, units of measurement, and their dimensions in various measurement systems. QUDT provides a unified model of measurable quantities, units of measurement for different kinds of quantities, numerical values of quantities in different units of measurement, and the data structures and data types used to store and manipulate these objects in software (FAIRsharing.org: QUDT, 2024). These classes and properties have been used by Simons et al. (2013) to model variables related to water quality. A similar approach could be evaluated to better characterize the modeling of a *Property* in the Theia/OZCAR thesaurus.

Difficulties related to the *Entity* modeling

Entity modeling can also vary depending on the solution chosen to model the *Variable*. In the context of the OZCAR-RI observatories, similar *Variables* are measured in surface water, groundwater, and karst water. Those three environmental compartments of the critical zone are qualified as *Entities* in the Theia/OZCAR vocabulary and are then used directly in the modeling of the *Variable* as shown in Fig. 7(A). However, the variables whose modeling involves those *Entities* could have been modeled by the "water' *Entity*

[20] Quantities, Units, Dimensions, and Types Ontology (https://www.qudt.org/).

and by using a *ContextObject* entity to describe the critical zone compartment. Figure 7(B) illustrates another possible modeling for the same *Variable*. The "karst water temperature" *Variable* could be described using "temperature" as the measured *Property*, "water" as the *ObjectOfInterest* entity and "karst" as the *Context*Object entity or the *Matrix* entity.

For consistency reasons, variables of the same type should follow the same design pattern in the vocabulary. Although the I-ADOPT working group does not provide recommendations on how to label the variables, it does define design patterns on how to model certain types of variables (quantitative and qualitative)[21]. Throughout the modeling of the Theia/OZCAR variables, an additional pattern could be identified for complex variables. This is the case for variables monitoring the flux of an entity between two compartments. Figure 8 shows the modeling of a "flux" variable and illustrates this pattern. For these variables, the *ObjectOfInterest* is an *Entity* that flows from one compartment to the other. We decided to limit the scope of the *Variable* by constraining the *ObjectOfInterest* using a *Constraint* that designates the two compartments and the direction of the flow. We also documented the interface of the two compartments as a *ContextObject* of the *Variable*. Similar patterns have also been identified in the context of applying the framework to Climate and Forecast Standard Names[22] (Pamment 2023).

Interoperability enhancement

Semantic consistency

The examples provided earlier in the section show that different modelings are possible for the same *Variable*. The preceding sections show that some modeling choices are subjective and guided by usage rather than by semantic accuracy. Beyond the consistency of the modeling of a given set of variables, this questions the way to automatically identify similarity links between variables coming from different information systems since the modeling depends on how one chooses to decompose one's variables. Identifying similarity links between variable decomposition from different vocabularies will require numerous semantic alignments between vocabularies from different domains since no vocabulary is likely to become mature enough to be used in a cross-domain manner. In particular, there is a need for semantic mediation between generic and specific properties that lends itself to reuse across multiple domains, while allowing users to use their preferred domain-specific terminology (Leadbetter and Vodden. 2016). Ontological analysis and alignment of the I-ADOPT framework with upper-ontologies could improve semantic consistency. They would help maintain a consistent interpretation of concepts within the vocabulary and between different systems. Ontological analysis is defined in this context according to Guarino (2008) as "the process of eliciting and discovering relevant distinctions and relationships bound to the very nature of the entities involved in a certain domain, for the practical purpose of disambiguating terms having different interpretations in different contexts" (Degbelo 2011). Applied to the I-ADOPT framework in the context of the Theia/OZCAR thesaurus, it would enable better characterization of the set of concepts and would help to constrain the modeling. It could then alleviate the modeling difficulties documented in the previous sections. Mapping the I-ADOPT framework to an upper-ontology would contribute to this process and would provide a common abstract foundation for sharing variable information between systems.

Semantic alignments

Semantic alignments of Theia/OZCAR thesaurus concepts have been defined with other vocabularies of the domain of environmental sciences (GCMD, EnvThes, AGROVOC, GEMET, AnaEE Thesaurus). There are several techniques for calculating similarities between ontologies: (i) terminological methods calculate similarity between text elements such as labels, descriptions, and definitions of the entities. (ii) Structural methods focus on the structural aspects of ontologies, aligning entities based on their positions in the ontology hierarchy. (iii) Semantic methods explore the semantics of entities, considering their meanings and relationships within ontologies by comparing ontologies according to a common context (i.e. an intermediate ontology). Considering that I-ADOPT framework is missing description logic specification and since none of the vocabulary mentioned in this section is aligned with a formal upper-ontology, semantic methods would require extra work to implement sophisticated reasoning mechanisms needed. Structural methods could facilitate the calculation of alignments with at least two vocabularies since the Theia/OZCAR thesaurus is inspired from AGROVOC and GCMD for the classification of environmental concepts and the hierarchization of variable categories respectively. However, for simplicity reasons, terminological methods using string comparisons of the labels are used to perform an initial sorting of the concepts to be aligned. The OnaGUI tool is used for this purpose. A manual check analyzing the positions of concepts in the vocabulary structures and their definitions is then carried out before validating the alignments. In the future, more advanced tools providing lexical matching using WordNet[23]

[21] https://github.com/i-adopt/patterns.

[22] http://vocab.nerc.ac.uk/collection/P07/current.

[23] https://wordnet.princeton.edu/.

synsets and combining the different approaches, as reviewed in Ardjani et al. (2015), could be evaluated to align Theia/OZCAR thesaurus with vocabularies having a more complex structure (using not only hierarchical relations) and already aligned to upper-ontology (e.g. Envo[24]).

Benefits of linked data in the scientific data sharing context

In the current context of open science, generic data repositories are being set up to meet the growing demand for data sharing. International metadata standards on which popular data repositories are based do not allow fine-grained description at the scale of observations composing datasets. Dublin Core (Weibel, S et al.1998) and ISO19115 (International Organisation for Standardisation 2014) are international standards that focus on describing datasets rather than observations. Dataverse (King 2007) and Zenodo[25] are based on Dublin Core, and Geonetwork (Ticheler and Hielkema 2007) is based on the ISO19115 metadata standard. One benefit of qualifying variable terms using models described in semantic web standards and published on the web is the possibility of integrating information describing observation into the metadata of a standard that is not designed to document information relating to the observation. By using variable terms linked to their web-accessible URIs and structured according to a conceptual model that defines variables, we can employ the keyword section of the metadata to document the measured variables in the dataset. This point can be illustrated by Theia/OZCAR datasets information harvested by the eLTER-DIP catalog where information about variables measured in a dataset appears in the keyword section[26]. This level of information must be accessible in data repository catalogs and is necessary to enable these datasets to be used in scientific contexts requiring the cross-referring of data from different sources and disciplines. In the same way, it could be useful to include in the metadata describing datasets, other information at the level of observation that is not described by the modeling of variable names (e.g. acquisition procedure, sensors, units, ultimate feature of interest.). The use of ontologies such as SOSA (Janowicz et al. 2019), which can describe the entire act of observation, would enable documenting this information. We could then imagine including the collection of observations composing the dataset in the metadata describing the dataset. Each observation would then be a keyword, resolvable with its URI, pointing to an *SOSA: Observation* class instance, modeling an act of observation. The variable name modeling work described in this paper is relevant in this context as it would bring precision to the description of the act of observation (Beretta et al. 2021).

Although further progress is needed to achieve interoperability between different vocabularies, the use of structured terms linked to each other and accessible through their URIs represents a first step towards the interoperability of information systems. Information systems use different data description models, the elements of which must be aligned if interoperability is to be achieved. The use of the I-ADOPT framework provides an initial semantic decomposition needed to implement some of the international standards encoding atomic observations. Data description model alignments are facilitated by the I-ADOPT WG, which provides alignments[27] to other ontologies referring to *Variables*.

Conclusion

To improve the use of data by a broad scientific community and in an interdisciplinary context, a clear understanding of the measured variables is required for both humans and machines. The use of ontologies describing variable names has been examined in this article. The I-ADOPT framework, which decomposes variable names into atomic elements, was tested within the context of the Theia/OZCAR IS, dedicated to continental surfaces and critical zone science and characterized by a large number and variety of observed variables. In this context, we proposed an implementation of the I-ADOPT framework. We showed that the I-ADOPT framework can be used effectively to describe complex variables with precision and that the framework, initially developed in the context of biodiversity variables, was flexible enough to be used in a wider context such as critical zone science. We have proposed a methodology for decomposing existing variables in the Theia/OZCAR IS. This allowed us to document variable names in detail while enabling them to be aligned with other ontologies or thesauri. We have encountered difficulties in using the framework for complex variables such as fluxes between different elements or ratios of physical quantities. We also showed that, for some variables, different decompositions were possible, which could make alignments with other ontologies and thesauri more difficult. The precision of the variable names proved inadequate for data discovery services and a non-standard label (*SimplifiedLabel*) had to be defined for this purpose. The I-ADOPT framework promotes the interoperability of information systems and will improve the use of data from different sources and disciplines in an open science perspective, as a user from another scientific community can more easily understand the meaning of variable names.

[24] https://sites.google.com/site/environmentontology/.

[25] https://zenodo.org/.

[26] https://dip.lter-europe.net/geonetwork/srv/eng/catalog.search#/metadata/TheiaOZCAR.KARS_DAT_MOSSON-6.

[27] https://github.com/i-adopt/supplementary/tree/master/alignments.

In the future, the Theia/OZCAR thesaurus will be complemented with new variable names, according to the needs of the scientific community it serves. The work presented in this paper also provides a solid basis for interoperability with other information systems at both national and international levels. We believe that the methodology proposed in this paper can be tested by other communities and we welcome feedback on this implementation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12145-024-01373-9.

Acknowledgements The work presented in this paper was funded by ANR, the French National Research Agency (FairTOIS project, contract ANR-19-DATA-0003). The work was conducted as part of the activities of OZCAR-RI and DATA TERRA RI, which are supported by the French Ministry of Research, French research institutions and universities. We would like to thank the open-source software and open science communities, without whom this work to improve the use of environmental data would not have been possible. We thank the four reviewers for their detailed reading of our manuscript and for the constructive comments they provided, which greatly improved the quality of the manuscript.

Author contributions CC, VC, IB, SG designed the study and discussed the implementation of the I-ADOPT framework. CC implemented the I-ADOPT framework in Theia/OZCAR thesaurus, with the help of VC. CC wrote the first version of the paper and designed the figures. All authors contributed to the paper and read and corrected the manuscript. SG led the FairTOIS project to which this paper contributes.

Funding The work presented in this paper was funded by ANR, the French National Research Agency (FairTOIS project, contract ANR-19-DATA-0003).

Data availability All the scientific variables described using the I-ADOPT framework during this work are included in the Theia/OZCAR controlled vocabulary, https://w3id.org/ozcar-theia.

Statement and declarations

The authors have no competing interests to declare that are relevant to the content of this article.

Competing interests The authors declare no competing interests.

References

Ardjani F, Bouchiha D, Malki M (2015) Ontology-Alignment Techniques: Survey and Analysis. International Journal of Modern Education and Computer Science 7:67–78. https://doi.org/10.5815/ijmecs.2015.11.08

Beretta V, Desconnets J-C, Mougenot I et al (2021) A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences. Comput Geosci 154:104807. https://doi.org/10.1016/j.cageo.2021.104807

Braud I, Chaffard V, Coussot C et al (2020) Building the information system of the French Critical Zone Observatories network: Theia/OZCAR-IS. Hydrol Sci J 67:2401–2419. https://doi.org/10.1080/02626667.2020.1764568

Bui EN (2016) Data-driven Critical Zone science: A new paradigm. Sci Total Environ 568:587–593. https://doi.org/10.1016/j.scitotenv.2016.01.202

Campos PMC, Reginato CC, Almeida JPA et al (2020) Finding reusable structured resources for the integration of environmental research data. Environ Model Softw 133:104813. https://doi.org/10.1016/j.envsoft.2020.104813

Cox SJD (2011) Geographic Information: Observations and Measurements OGC Abstract Specification Topic 20, OpenGIS® Abstract Specification OGC 10–004r3 and ISO 19156

Cox SJD, Gonzalez-Beltran AN, Magagna B, Marinescu M-C (2021) Ten simple rules for making a vocabulary FAIR. PLoS Comput Biol 17:e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Crutzen PJ (2002) Geology of mankind. Nature 415:23–23. https://doi.org/10.1038/415023a

David J, Euzenat J, Scharffe F, Trojahn Dos Santos C (2011) The Alignment API 4.0. Semantic Web 2:3–10. https://doi.org/10.3233/SW-2011-0028

Degbelo A (2011) An Ontological Analysis of Observation Collections. Semantic Web 1:5

Eclipse RDF4J (2021) Eclipse RDF4J - version 3.7.4. https://github.com/eclipse/rdf4j. Accessed 19 Jun 2023

EnvThes (2023) eLTER Vocabularies: EnvThes - Thesaurus for long term ecological research, monitoring and experiments. https://vocabs.lter-europe.net/envthes/en/. Accessed 19 Jun 2023

FAIRsharing.org: QUDT Quantities, Units, Dimensions and Types, https://doi.org/10.25504/FAIRsharing.d3pqw7, Last Edited: Friday, May 6th 2022

Finkel M, Baur A, Weber TKD et al (2020) Managing collaborative research data for integrated, interdisciplinary environmental research. Earth Sci Inform 13:641–654. https://doi.org/10.1007/s12145-020-00441-0

Gaillardet J, Braud I, Hankard F et al (2018) OZCAR: The French Network of Critical Zone Observatories. Vadose Zone Journal 17:180067. https://doi.org/10.2136/vzj2018.04.0067

Grellet S, Magagna B, schleidt K, et al (2021) How I-ADOPT complements ISO/OGC Observations and Measurements? Virtual SciDataCon 2021. https://hal.science/hal-04233789

Guarino N (2008) Ontologies and ontological analysis: an introduction. Tutorial at FOIS 2008

Huynh F, Baghdadi N, Diament M et al (2019) L'infrastructure de recherche « Pôle de données et services pour le système Terre », à la pointe des techniques d'imagerie et de cartographie numérique. Annales Des Mines - Responsabilité Et Environnement 94:8–13. https://doi.org/10.3917/re1.094.0008

I-ADOPT working group (2023a) I-Adopt Terminology Repository. https://i-adopt.github.io/terminologies. Accessed 21 Jun 2023

I-ADOPT working group (2023b) i-adopt/supplementary: Documents supplementing the I-ADOPT recommendations. https://github.

com/i-adopt/supplementary/tree/master/alignments. Accessed 19 Jun 2023
I-ADOPT working group (2021) InteroperAble Descriptions of Observable Property Terminology WG (I-ADOPT WG). In: RDA. https://www.rd-alliance.org/groups/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg. Accessed 30 Aug 2023
INSPIRE Maintenance and Implementation Group (MIG) (2016) Guidelines for the use of Observations & Measurements and Sensor Web Enablement-related standards in INSPIRE. Version 3.0. https://inspire.ec.europa.eu/id/document/tg/d2.9-o%26m-swe. Accessed 19 Jun 2023
International Organisation for Standardisation (2014) ISO 19115–1:2014, Geographic information - Metadata - Part 1: Fundamentals. https://www.iso.org/obp/ui/en/#iso:std:iso:19115:-1:ed-1:v1:en
Isaac A, Summers E (2009) SKOS Simple Knowledge Organization System Primer. W3C recommendation. https://www.w3.org/TR/skos-primer/
Janowicz K, Haller A, Cox SJD et al (2019) SOSA: A lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics 56:1–10. https://doi.org/10.1016/j.websem.2018.06.003
King G (2007) An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods & Research 36:173–199. https://doi.org/10.1177/0049124107306660
Lausch A, Schmidt A, Tischendorf L (2015) Data mining and linked open data – New perspectives for data analysis in environmental research. Ecol Model 295:5–17. https://doi.org/10.1016/j.ecolmodel.2014.09.018
Leadbetter AM, Vodden PN (2016) Semantic linking of complex properties, monitoring processes and facilities in web-based representations of the environment. International Journal of Digital Earth 9:300–324. https://doi.org/10.1080/17538947.2015.1033483
Madin J, Bowers S, Schildhauer M et al (2007) An ontology for describing and synthesizing ecological observation data. Eco Inform 2:279–296. https://doi.org/10.1016/j.ecoinf.2007.05.004
Magagna B, Moncoiffé G, Devaraju A, et al (2022) InteroperAble Descriptions of Observable Property Terminologies (I-ADOPT) WG Outputs and Recommendations. https://doi.org/10.15497/RDA00071
Magagna B, Rosati I, Stoica M, et al (2021) The I-ADOPT Interoperability Framework for FAIRer data descriptions of biodiversity. S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity, Bozen, Italy. https://arxiv.org/abs/2107.06547
Magagna B, Schindler S, Stoica M, et al (2023) I-ADOPT Framework ontology. https://i-adopt.github.io/index-en.html. Accessed 19 Jun 2023
Mazuel L (2023) OnAGUI - Ontology Alignment GUI - version 0.3.6. https://github.com/lmazuel/onagui. Accessed 19 Jun 2023
McDowell WH (2015) NEON and STREON: opportunities and challenges for the aquatic sciences. Freshwater Science 34:386–391. https://doi.org/10.1086/679489
Miles A, Bechhofer S (2009) SKOS simple knowledge organization system reference. W3C recommendation. https://www.w3.org/TR/skos-reference/
Mosconi G, Li Q, Randall D, et al (2019) Three Gaps in Opening Science. Computer Supported Cooperative Work (CSCW) 28:. https://doi.org/10.1007/s10606-019-09354-z
Mougin C, Azam D, Caquet T et al (2015) A coordinated set of ecosystem research platforms open to international research in ecotoxicology, AnaEE-France. Environ Sci Pollut Res Int 22:16215–16228. https://doi.org/10.1007/s11356-015-5233-9
National Research Council (2001) Basic Research Opportunities in Earth Science. National Academies Press, Washington, D.C.
Pamment A (2023) Practical implementations of the I-ADOPT framework and future directions - Implementation with the CF Standard Names. Research Data Alliance 20th Plenary Meeting, Sweden
Parsons M, Godøy Ø, LeDrew E et al (2011) A conceptual framework for managing very diverse data for complex, interdisciplinary science. J Information Science 37:555–569. https://doi.org/10.1177/0165551511412705
Peckham SD, Hutton EWH, Norris B (2013) A component-based approach to integrated modeling in the geosciences: The design of CSDMS. Comput Geosci 53:3–12. https://doi.org/10.1016/j.cageo.2012.04.002
Pichot C, Maurice D, Monet G et al (2021) Semantic Management of Data from Biodiversity and Ecosystem Studies: Toward an Integrated Workflow from Collection to Publication. Application to Plankton Data from Lake Geneva. S4BioDiv, 3rd International Workshop on Semantics for Biodiversity. Bozen, Italy
Rittel HWJ, Webber MM (1973) Dilemmas in a general theory of planning. Policy Sci 4:155–169. https://doi.org/10.1007/BF01405730
Schildhauer M, Jones MB, Bowers S, et al (2016) OBOE: The Extensible Observation Ontology - version 1.2. https://github.com/NCEAS/oboe. Accessed 19 Jun 2023
Simons B, Yu J (2013) Defining a water quality vocabulary using QUDT and ChEBI
Skosmos (2023) Skosmos. https://skosmos.org/. Accessed 21 Jun 2023
Stellato A, Fiorelli M, Lorenzetti T, Turbati A (2021) Collaborative Maintenance of EDOAL Alignments in VocBench. pp 243-254
Stellato A, Fiorelli M, Turbati A, et al (2020) VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. SW 11:855–881. https://doi.org/10.3233/SW-200370
Stoica M, Peckham SD (2019) The Scientific Variables Ontology: A Blueprint for Custom Manual and Automated Creation and Alignment of Machine-Interpretable Qualitative and Quantitative Variable Concepts. https://api.semanticscholar.org/CorpusID:250521087
Suominen O, Ylikotila H, Pessala S, et al (2015) Publishing SKOS vocabularies with Skosmos. https://seco.cs.aalto.fi/publications/2016/suominen-et-al-skosmos.pdf
Theia/OZCAR Thesaurus (2018) Theia/OZCAR thesaurus: Thesaurus for in situ data from Environmental and Critical Zone Sciences. http://doi.osug.fr/r/67b5a1d5-8c8c-4a94-a646-1cca1d0adf79. Accessed 19 Jun 2023
Ticheler J, Hielkema JU (2007) Geonetwork opensource internationally standardized distributed spatial information management. OSGeo Journal 2:1. https://svn.osgeo.org/osgeo/journal/volume_2/en-us/final_pdfs/OSGeoJournal_vol2_GeoNetwork.pdf
VocBench (2023) VocBench: A Collaborative Management System for OWL ontologies, SKOS(/XL) thesauri, Ontolex-lemon lexicons and generic RDF datasets. https://vocbench.uniroma2.it/doc/sys/#separate_triple_store. Accessed 21 Jun 2023
Weibel S, Kunze J, Lagoze C, Wolf M (1998) RFC2413: Dublin Core Metadata for Resource Discovery. https://www.rfc-editor.org/rfc/rfc2413
Weinberger D (2002) Small pieces loosely joined: a unified theory of the Web. Perseus, Cambridge, MA
Wilkinson MD, Dumontier M, IjJ A et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18
Wohner C, Peterseil J, Klug H (2022) Designing and implementing a data model for describing environmental monitoring and research sites. Eco Inform 70:101708. https://doi.org/10.1016/j.ecoinf.2022.101708