Towards a Core Ontology for Scientific Research Activities

Patricia M. C. Campos, Cássio C. Reginato and João Paulo A. Almeida

Ontology & Conceptual Modeling Research Group (NEMO),
Federal University of Espírito Santo (UFES),
Av. Fernando Ferrari, 514, Goiabeiras, 29075-910, Vitória, ES, Brazil
patricia.carnelli@aluno.ufes.br,
cassio.reginato@inf.ufes.br, jpalmeida@ieee.org

Abstract. The increasing volume and complexity of scientific research data associated with its semantic heterogeneity demands strategies to enable data integrated reuse. This is essential to improve global collaborations, in what has been called e-Science. A way to promote data integration is through the use of ontologies. Ontologies can play the role of a shared conceptualization, providing a common semantic background for data interpretation. In the case of scientific research, particularly empirical research, there are many concepts related to research activities that are general, despite any specific domain in which they may occur. Thus, they can be represented by means of a core ontology. In this paper, we propose the design of a core ontology to deal with research activities (e.g., sampling and measurement). As the concepts used are neutral with respect to different application domains, they can be reused to build ontologies for specific research domains, speeding up the development process. To illustrate this, we present an environmental research ontology developed based on this core ontology. The proposed core ontology is grounded in the Unified Foundational Ontology (UFO), which provides a solid basis for its key elements.

Keywords: Core Ontology, Scientific Research, Research Activity, Unified Foundational Ontology.

1 Introduction

In the last decades, scientific research has undergone major changes mainly due to the increasing volume and complexity of data produced and the need to share such data to improve global collaborations, in what has been called e-Science [1]. This new paradigm of scientific research aims to promote the development of science and technology through the use of methods that enable more powerful and synthetic data analyses from integrated data reuse. However, because of the variety of actors involved and the interdisciplinary nature of scientific research, scientific data are often available disconnected, using incompatible language and heterogeneous terminology. This brings up serious semantic interoperability concerns [2].

The FAIR Data initiative [3] presents a set of principles that must be regarded to address this problem. These principles cover aspects of how data should be semantically annotated with metadata in a way that it can be read, interpreted, and reused for humans and machines. One of the key principles establishes that metadata must meet domain-relevant community standards. This means that data produced must be annotated with metadata that, in turn, must reference domain-relevant community standards, such as ontologies. As presented in [4], ontologies can be used, among other possibilities, as global (or shared) conceptualization for data integration. In this sense, ontologies can promote data interoperability by providing a common semantic background for data interpretation, reducing conceptual ambiguities and inconsistencies, and supporting meaning negotiation.

In the case of scientific research [5], particularly empirical research, where evidence is gathered through experimentation or observation, there are many concepts related to research activities that are general, despite any specific domain in which they may occur. Thus, they can be represented by means of a core ontology. Core ontologies provide a precise definition of structural knowledge in a specific field that spans across different application domains in this field [6]. They can be reused and extended to incorporate particularities of the domains of interest, that is, for the construction of domain ontologies. So, in addition to providing a shared conceptualization, they enable the speeding up of the domain ontology development process.

In this paper, we propose the design of a core ontology to deal with the different types of research activities performed in empirical research, encompassing (physical) sampling, sample preparation and measurement. As the concepts used are neutral in relation to the various domains, they can be reused by a given domain. To illustrate this, we present an environmental research ontology developed on the basis of this core ontology. It is worth mentioning that the explicit modeling of research activities shows that provenance information (e.g., participation of actors, participation of devices, methods used, etc.), usually present in *metadata*, are actually properties of real-world events. The proposed core ontology is grounded in the Unified Foundational Ontology (UFO) [7][8], from which basic notions of object, relation, property, event, and others are adopted.

This paper is structured as follows. Section 2 presents the background of the paper, which includes UFO concepts relevant to ground the core ontology development. Section 3 presents the *Core Ontology for Scientific Research Activities*. Section 4 illustrates the use of the core ontology to build a domain ontology for environmental research. Section 5 discusses related work, and section 6 presents our final considerations.

2 Background

In developing a core ontology, it is desirable to use a solid modeling base given by a foundational ontology. Concepts and relationships defined in a core ontology should be aligned to the basic categories of a foundational ontology [6]. For the field of scientific research, we need the general concept of events to represent research activities.

Also, the basic concept of object is necessary to deal with devices, procedures and physical samples. The concept of agent is necessary to represent people and organizations involved in research activities. In addition, to approach measurements, we need to speak of properties (qualities) and their quantification.

As the Unified Foundational Ontology (UFO) [7][8] provides these basic concepts, we have used UFO to ground the construction of the *Core Ontology for Scientific Research Activities*. UFO has been developed based on theories from Formal Ontology, Philosophical Logics, Philosophy of Language, Linguistics and Cognitive Psychology. UFO consists of three main modules: UFO-A, an ontology of endurants (objects); UFO-B, an ontology of perdurants (events); and UFO-C, an ontology of social entities built up on UFO-A and UFO-B.

Fig. 1 shows a fragment of UFO containing concepts from UFO-A, UFO-B and UFO-C. The root concept is *Entity*, which is specialized into *Universal* and *Individual*. Universals are patterns of features that can be realized in a number of different individuals [7]. Individuals can be *concrete* (e.g., a particular person, an explosion) or *abstract* (e.g., sets, numbers, and propositions). *Concrete Individuals* are divided into *Endurants* and *Events*. Endurants are individuals that are wholly present whenever they are present (e.g., a house, a person, an amount of sand, etc.). Events are individuals that may have temporal parts. They happen in time in the sense that they extend in time and accumulate temporal parts (e.g., a soccer match). Whenever an event is present, it is not the case that all its temporal parts are present [8].

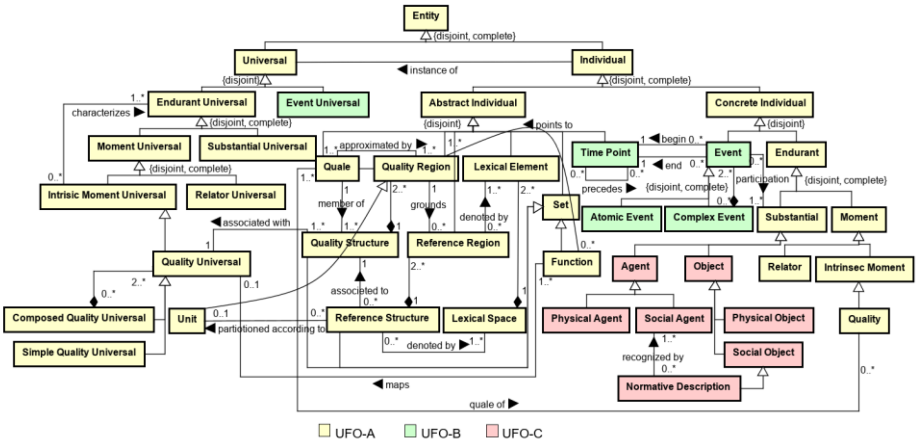

Fig. 1. A fragment of UFO-A, UFO-B and UFO-C.

The category of endurants can be further specialized into *Substantial* and *Moment*. Substantials are existentially-independent individuals (e.g., a house, a person). Moments are individuals that can only exist in other individuals, and, thus, they are existentially-dependent on their bearers (e.g., a color, an electric charge, a social commitment). *Intrinsic Moments* are moments that are dependent on one single individual (e.g., a color, a temperature). *Relators*, in turn, are moments that existentially depend on a plurality of individuals (e.g., an employment, a business process) [8].

Concerning the substantial hierarchy, a basic distinction is between agentive and non-agentive individuals, termed *Agents* and *Objects*, respectively. Agents can be divided into *Physical Agents* (e.g., a person) and *Social Agents* (e.g., an organization, a society). Objects can also be further categorized in *Physical Objects* and *Social Objects*. Physical objects include a book, a car, among others; social objects include money, language, etc. A *Normative Description* is a type of social object that defines one or more rules/norms recognized by at least one social agent. Examples of normative descriptions include contracts in general, but also sets of directives on how to perform actions within an organization [8].

Events can be atomic or complex. *Atomic Events* have no proper parts. *Complex Events* are aggregations of at least two disjoint events. Events are ontologically dependent entities in the sense that they depend on substantial participation to exist. Take for instance the event of measuring the height of a person. In this event, we have the participation of the measured person, the person that performs the measurement and the instrument used to measure the height. This event is composed of the individual participation of each of these entities and depends on them to exist. Besides that, each event is associated with two *Time Points*: a begin and an end time point. Time points are abstract individuals strictly ordered by a precedes relation [8].

Universals can be classified in *Endurant Universals* and *Event Universals.* Endurant universals are patterns of features of endurants. Event universals instead are patterns of features of events. *Substantial Universal* and *Moment Universal* are endurant universals whose instances are substantials and moments, respectively. Moment Universal is divided into *Intrinsic Moment Universal* and *Relator Universal* [7].

Regarding the intrinsic moment universal hierarchy, *Quality Universals* refer to the properties that characterize universals (e.g., weight, height). They are always associated with values spaces or *Quality Structures* that can be understood as the set of all possible regions (*Quality Regions*) that delimits the space of values that can be associated to a quality universal [7]. For example, height is associated with one-dimensional structure with a zero point isomorphic to the half-line of nonnegative numbers. Other properties such as color are represented by multidimensional structures. Quality universals associated with one-dimensional structures are called *Simple Quality Universals*. Quality universals associated with multidimensional structures are called *Composed Quality Universals*. The perception or conception of an intrinsic moment can be represented as a point in a quality structure. This point is named *Quale*. Quality regions and qualia are abstract entities. *Function* is a specialization of set that maps instances of a quality universal to points in a quality structure [9].

In order to allow quale communication, it is necessary to use *Lexical Elements* (e.g., 1.86 can be the lexical element used to communicate the height of a person) associated to *Reference Regions* and *Reference Structures*. A reference region is an abstract entity based on a quality region that acts as a bridge between that region and the lexical elements used to communicate the quale. A reference structure is associated to a quality structure and is a set of reference regions grounded in quality regions of that quality structure. When the 'value' of a particular quality is being referred by lexical elements (e.g., 1.86), what is actually being referred is a quality region that most approximates the quale. Reference structures associated to quality structures related to measurement act like scales grounded by quality structures. They can be partitioned in spaces with the same magnitude according to a *Unit* [9].

3 The Core Ontology for Scientific Research Activities

In this section, we present the *Core Ontology for Scientific Research Activities*, which deals with the different types of research activities performed in scientific research. We have identified that some characteristics are common to all types of research activities, such as temporal and spatial properties, actors involved in their execution, responsible actors, among others. They are related to provenance information and are generally addressed by the metadata domain, but the modeling of research activity shows that they are properties of events.

We have created a subontology to represent these properties: the *Research Activity Ontology*. This subontology must be specialized to handle the intrinsic characteristics of each type of research activity. We have specialized it in the following subontologies: *Sampling Ontology*, *Preparation Ontology* and *Measurement Ontology*. However, new specializations can be made to deal with other types of research activities, such as observations (e.g., an observation of the taxon of a beetle), assays, etc. **Fig. 2** shows the *Core Ontology for Scientific Research Activities* and the UFO concepts used to ground it. Next, we explain each of these subontologies.

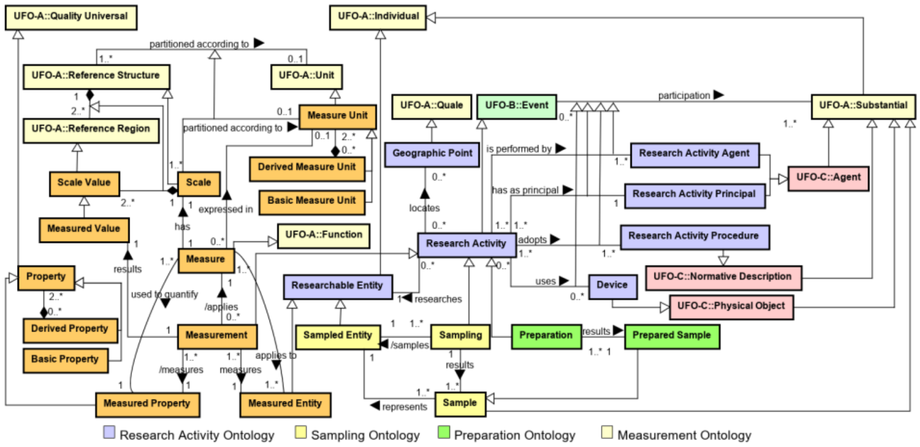

Fig. 2. The Core Ontology for Scientific Research Activities.

The *Research Activity Ontology* comprises concepts that are common to the different types of research activities. *Research Activity* is a UFO-B event used to generalize these types. Research activities are characterized by temporal and spatial properties, as well as the researched entity. Regarding temporal properties, research activities inherit begin and end time points from UFO-B. In relation to spatial properties, *Geographic Point* is a UFO-A quale that represents the coordinates corresponding to the spatial location of a research activity. *Researchable Entity* is a specialization of UFO-

A individual because it can be a substantial (e.g., a river, a city) or an event (such as a process). A research activity is also characterized by the procedure adopted and the device employed. *Research Activity Procedure* is a UFO-C normative description that defines the rules to be followed for the execution of a research activity. *Device* is a specialization of UFO-C physical object. Examples of devices are: collectors, sensors, etc. In order to capture provenance, the *Agents* involved in the execution and the agent responsible for a research activity (the so-called *Principal*) are identified. They are specializations of UFO-C agent and can be physical (such as researches) or social agents (governmental agencies, research institutions, laboratories, etc.).

The *Sampling Ontology* deals with concepts related to the sampling activity. Sampling is the collection of samples for in situ and/or laboratory analysis. *Sampling* is a specialization of research activity, inheriting concepts related to research activity. *Sampled Entity* is a specialization of researchable entity and represents the target research entity. *Sample* represents a portion of a sampled entity that must be analyzed with the ultimate goal of characterizing the sampled entity. Sample is a specialization of UFO-A substantial. For instance, in the case of a water quality research of a river, a sample of water or sediment can be collected to verify the river water quality.

The *Preparation Ontology* address concepts related to the sample preparation activity. It refers to the ways in which a sample is treated before being analyzed. *Preparation* is a specialization of research activity. *Prepared Sample* represents a sample that has been prepared for analysis. Not all samples need to be prepared before they are analyzed.

The *Measurement Ontology* provides concepts related to the measurement activity. Most of the concepts presented here were extracted from the measurement core ontology presented in [9], which was developed in alignment with UFO. Measurement can be defined as a set of actions aiming to characterize an entity by attributing values to its properties. *Measurement* is a specialization of research activity. *Measured Entity* is a specialization of researchable entity. It represents an entity that has one or more measured properties, such as a person, a water sample, etc. *Property* is a UFO-A quality universal that deals with qualities of entities. It specializes in basic and derived property. *Basic Property* is a UFO-A simple quality universal that does not depend on other properties to be measured (e.g., weight and height). *Derived Property* is a UFO-A composed quality universal that depends on others to be measured (for example, Body-Mass Index). *Measured Property* represents a property that is measured. *Measures* are used for quantifying measured properties. Measure is a UFO-A function in the sense that it maps an instance of measured property to a measured value. Measures have *Scales* composed by all possible values (*Scale Value*) to be associated to a measured property. Scale is a specialization of UFO-A reference structure and scale value is a specialization of UFO-A reference region. Measures can be expressed in *Units* (e.g., meter, kilogram). A measure unit in which a measure is expressed partitions its scale.

4 Using the Core Ontology for building an Environmental Research Ontology

In this section, we present how the proposed core ontology can be used as the basis for the development of a domain ontology. In this case, an environmental research ontology focused on water quality. This domain deals with the water quality assessment at monitoring points along rivers, lakes, sea, etc. This assessment is performed by analyzing measurements of physical, chemical and biological properties of water and sediment samples and ecotoxicological assays. We have used two other core ontologies to build this ontology: *Spatial Location Ontology* and *Material Entity Ontology*. The domain ontology, called *Environmental Research Ontology*, is divided into *Water Quality Ontology* and *Environmental Monitoring Ontology* (see **Fig. 3**).

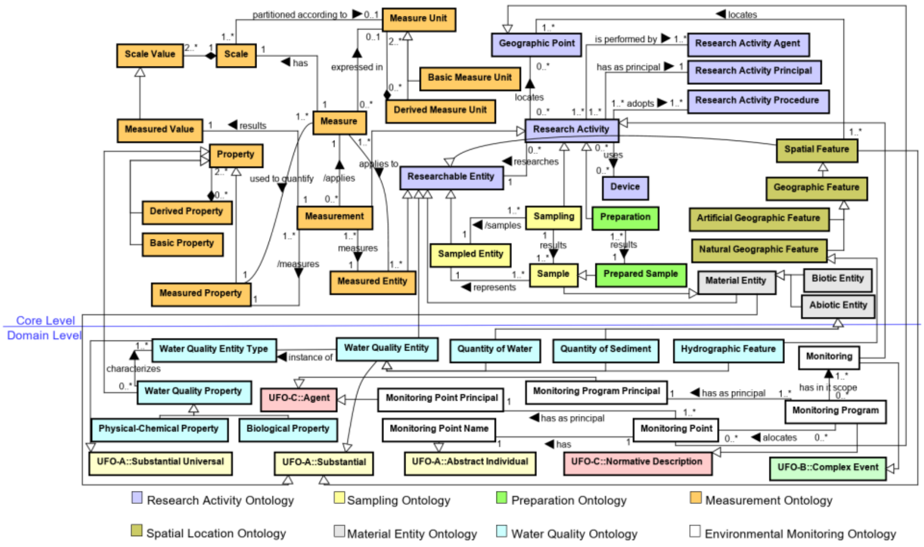

Fig. 3. The Environmental Research Ontology.

At the core level, the *Material Entity Ontology* comprises concepts for dealing with the existing types of material entities. The main concept is *Material Entity*, a UFO-A substantial that specializes into *Abiotic Entity* (non-living parts of an environment) and *Biotic Entity* (living parts of an environment).

The *Spatial Location Ontology* provides concepts related to spatial features (anything with spatial extent, such as a country, a river, etc.). *Spatial Feature* is a UFO-A substantial. It is specialized into *Geographic Feature*. Geographic features can be naturally-created (e.g., a river, a mountain) or artificially-created spatial features (e.g., a city, a water treatment plant). Spatial features are located in geographic points.

These other core ontologies are connected to the proposed core ontology by the researchable entity that is specialized in spatial feature and material entity. Also, sample is a specialization of material entity.

At the domain level, the *Water Quality Ontology* comprises concepts about water quality entities and properties. A *Water Quality Entity*, a UFO-A substantial, can be a *Hydrographic Feature*, a *Quantity of Water*, a *Quantity of Sediment*, among others. Hydrographic feature is a specialization of natural geographic feature and represents rivers, lakes, hydrographic basins, seas, etc. *Water Quality Property,* a UFO-A quality universal, refers to properties that are used to characterize water quality entities, encompassing both *Physical-Chemical* (e.g., temperature, dichloroethene concentration) and *Biological Properties* (e.g., concentration of coliforms, algae).

Finally, *the Environmental Monitoring Ontology* defines concepts related to environmental monitoring, monitoring points and monitoring programs. *Monitoring* consists of a set of research activities, performed periodically, for environmental quality control. Monitoring is a UFO-B complex event because it is composed of other research activities, such as sampling and measurement. *Monitoring Point* is a specialization of geographic point used to represent named geographic points. *Monitoring Point Name* is used to describe the location of the monitoring point. *Monitoring Programs* are UFO-C normative descriptions that have in their scope monitoring activities and allocate monitoring points to perform them. *Monitoring Point Principal* and *Monitoring Program Principal* are used to represent the agents responsible for monitoring points and monitoring programs, respectively.

The domain ontologies are connected to the proposed core ontology through researchable entity that specializes in water quality entity and property that specializes in water quality property. Monitoring is a specialization of research activity. It was not modeled as a core activity because it depends on particularities of the application domain. For example, monitoring program and monitoring point are concepts from the environmental domain, but not in every other scientific research domains.

5 Related Work

There are some models [10][11][12][13][14] related to scientific research based on the Observations and Measurements conceptual model from ISO 19156 [15]. This model defines an observation as an activity, the result of which is an estimate of the value of a property of the feature of interest, obtained using a specified procedure. Specializations of the observation have been classified by the result-type. For example, a measurement is an observation whose result is a scaled quantity, and a truth observation is an observation whose result is a Boolean value. As well as in [15], some ontologies [10][11] do not represent the sampling activity; they represent only the sampling features. A sampling feature is used to support the observation process and may or may not have a persistent physical expression. Physical samples are modeled as the sampling feature specimen. In [10] and [15], sample preparation is implemented using an association class with specimen. As sampling is not modeled as an activity, sampling properties need to be assigned to other entities. Specimen has properties related to sampling time, sampling location, etc. Observation has phenomenon time and result time to differentiate the moment of the sampling from the time of the ex-situ measurement of a sample, respectively. Thus, events and objects concepts are mixed. This shows the importance of developing core and domain ontologies based on a foundational ontology, characteristic not presented by these models.

The Semantic Sensor Network (SSN) ontology [12] describes sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators. In SSN, the sampling activity is modeled. Sampling is used to represent both sampling and preparation activities. Location is not addressed. It is suggested that other models must be used to deal with location. Agents and devices involved in observations are treated by the same sensor entity. The Extensible Observation Ontology (OBOE) [13] is a formal ontology for capturing the semantics of scientific observation and measurement. OBOE does not handle other research activities. The Observation Data Model (ODM) [14] is an information model and supporting software ecosystem for feature-based earth observations, designed to facilitate interoperability across scientific disciplines and domain cyberinfrastructures. It models observation results, sample properties, monitoring locations, but does not model the research activities themselves, which we have shown key to capturing provenance information.

6 Conclusions

Improving scientific research based on data reuse requires adequate support for data semantics. We have addressed this challenge developing the *Core Ontology for Scientific Research Activities*. From this core ontology, we can develop domain ontologies that form the basis of mechanisms for finding, publishing and querying heterogeneous scientific research data. It promotes the application of FAIR principles in the scientific research field. As an example, we have developed a domain ontology for the environmental research. It was also possible to verify how the reuse of the core ontology facilitates the domain ontology development process.

The use of UFO as a foundational ontology supports the correct classification of the different concepts and relations about research activities, leveraging key notions that are domain independent. Activities are modeling as events, actors as agents, devices as objects, their participations in events revealed, and so on. The use of foundational ontology is a key feature of the proposed core ontology when contrasted with related work. By not adhering to a foundational ontology, some misconceptions arise, e.g., with event properties assigned to objects.

Besides that, the explicit modeling of research activity reveals that provenance information, usually present in the metadata domain, are actually properties of events, including the participation of agents and non-agentive objects in those events. In the case of scientific research, the modeling of these concepts is fundamental to support the integrated data reuse. Otherwise, there is a risk that such data will be misused. For example, data produced by incompatible methods can be compared, leading to inconsistent analysis; incorrect providers can be assigned to data since original data can be reprocessed by different agents; and so on.

As future work, other types of research activities should be modeled to broaden the scope of the core ontology. In addition, other aspects of scientific research, as well as research activities, can be incorporated. Some examples are types of scientific research, scientific research purpose, scientific research planning, etc.

Acknowledgements

This work is partly supported by CNPq (407235/2017-5 and 312123/2017-5), CAPES (23038.028816/2016-41) and FAPES (69382549).

7 References

1. Hey, T., Trefethen, A.: The Data Deluge: An e-Science Perspective. In: Grid Computing - Making the Global Infrastructure a Reality. Berman, F., Fox, G. C., Hey, T. (eds.), pp. 809-824. Wiley and Sons, Ltd, Chichester, UK (2003).
2. Lenzerini, M.: Data Integration: A Theoretical Perspective. In: Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '02, pp. 233-246. ACM, New York, NY, USA (2002).
3. Wilkinson, M. D., et al.: Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1) (2016).
4. Cruz, I. F., Xiao, H.: The Role of Ontologies in Data Integration. Journal of Engineering Intelligent Systems, 13(4), 245-252 (2005).
5. Çaparlar, C. Ö., Dönmez, A.: What is scientific research and how can it be done? Turk J Anaesthesiol Reanim, 44(4), 212-218 (2016).
6. Scherp, A., Saathoff, C., Franz, T., Staab, S.: Designing core ontologies. Appl. Ontol., 6(3), 177-221 (2011).
7. Guizzardi, G.: Ontological foundations for structural conceptual models. CTIT PhD Thesis Ser (2005).
8. Guizzardi, G., Falbo, R., Guizzardi, R. S. S.: Grounding Software Domain Ontologies in the Unified Foundational Ontology (UFO): The case of the ODE Software Process Ontology. In: CIbSE, pp. 127-140 (2008).
9. Barcellos, M., Falbo, R., Frauches, V.: Towards a measurement ontology pattern language. In: Proc. 1st Joint Workshop ONTO.COM/ODISE on Ontologies in Conceptual Modeling and Information Systems Eng. of CEUR Workshop Proc., vol. 1301 (2014).
10. Cox, S. J. D.: An explicit OWL representation of ISO/OGC observations and measurements. In: Proceedings of the 6th International Conference on Semantic Sensor Networks, vol. 1063, pp. 1-18 (2013).
11. Cox, S. J. D.: Ontology for observations and sampling features, with alignments to existing models. Semantic Web, 8(3), 453-470 (2016).
12. Haller, A., et al.: The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. Semantic Web, 10(1), 9-32 (2018).
13. Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F.: An ontology for describing and synthesizing ecological observation data. Ecol. Inform., 2(3), 279-296 (2007).
14. Horsburgh, J. S., Tarboton, D. G., Piasecki, M., et al.: An integrated system for publishing environmental observations data. Environ. Model. Softw., 24(8), 879-888 (2009).
15. ISO 19156:2011: Geographic information - Observations and measurements. International Standard (2011).