Semantic Management of Data from Biodiversity and Ecosystem Studies: Toward an Integrated Workflow from Collection to Publication. Application to Plankton Data from Lake Geneva

Christian Pichot [1] , Damien Maurice [2], Ghislaine Monet [3], Rachid Yahiaoui [4], Philippe Clastre [1] and Benjamin Jaillet [1]

[1] INRAE, URFM, 228 route de l'Aérodrome, 84914 Avignon, France
[2] INRAE, UMR SILVA, route d'Amance, 54280 Champenoux, France
[3] INRAE, UMR CARRTEL, 75 avenue de Corzent, 74200 Thonon-les-bains, France
[4] INRAE, US INFOSOL, 2163 avenue de la Pomme de Pin, 45075 Orléans, France

Abstract

Biodiversity is a key player in ecosystem characteristics and dynamics. Acting as a driver, it also results from ecosystem functioning. Understanding this complex interplay between biological and physical components is one of the main current challenges in the context of land use changes and climate warming. The acquisition of knowledge on biodiversity requires multidisciplinary approaches and mobilises numerous research teams. Data are collected or computed in large quantity but are most often poorly standardised and therefore heterogeneous. In this context the development of semantic interoperability is a major challenge for the sharing and reuse of these data. This objective is implemented within the framework of the AnaEE (Analysis and Experimentation on Ecosystems) Research Infrastructure dedicated to experimentation on ecosystems and biodiversity. A distributed Information System (IS) is developed, based on the semantic interoperability of its components using common vocabularies (AnaeeThes thesaurus and OBOE-based ontology extended for disciplinary needs) for modelling the studied system. This modelling covers the measured variables including biodiversity, as well as the different components of the experimental or observational context, from sensor to plot and network. Driven by the ontology, the approach relies on the atomic decomposition of each of the components into observed entities, their characteristics and qualifiers, their units or naming standards. The modelling of the system allows the semantic annotation of relational databases or flat files for the production of URIs based graph databases. A first pipeline automates the annotation process and the production of the semantic data. A second pipeline is devoted to the exploitation of these semantic data by generating i) metadata records formatted according to the geospatial extension for the Data Catalog Vocabulary standard and the ISO 19139 standard, and ii) Network Common Data Form data files. The implementation of this integrated semantic management of data is presented here for phyto- and zoo-plankton data collected from water columns in Lake Geneva over a 30 years period, as well as for environmental data about water temperature and nutrients. The work carried out contributes to the development and use of semantic vocabularies within the biodiversity and ecology research community, leading to semantically enriched metadata records and interoperable data sets. The genericity of the tools make them usable in different contexts of data production, management and ontologies involved in semantic modelling.

Keywords

interoperability, biodiversity, plankton, ontology, modelling, pipeline, entity property, FAIR data

[1]S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity, held at JOWO 2021: Episode VII The Bolzano Summer of Knowledge, September 11–18, 2021, Bolzano, Italy.
EMAIL: christian.pichot@inrae.fr (A. 1); damien.maurice@inrae.fr (A. 2); ghislaine.monet@inrae.fr (A. 3); rachid.yahiaoui@inrae.fr (A. 4); philippe.clastre@inrae.fr (A. 5); benjamin.jaillet@inrae.fr (A. 6)
ORCID: 0000-0003-1636-9438 (A. 1)

1. Introduction

The knowledge of ecosystem structure and functioning is more than never a prerequisite to tackle the global challenges we are now facing (global warming, food supply, biodiversity preservation...). Actually we have to anticipate the middle to long term trajectory of our living planet according to the socio-economic-political choices that could be made. Lessons from the past and empirical knowledge are of course to be taken into account but the complexity of the system in which multiple interactions take place and, above all, the unprecedented environmental context produced by global warming, make it essential to increase scientific knowledge and share it across disciplines. Acting as a driver as well as resulting from ecosystem functioning biodiversity takes a central place in the game. In addition to data collection, their FAIRification for common understanding, sharing and re-use is the challenge.

In this general context, the AnaEE Research Infrastructure develops services dedicated to the study of continental terrestrial and aquatic ecosystems. Its thematic scope concerns the biological diversity and functioning of grassland, crop, forest and lake ecosystems. The services offered include open-air and closed experimental platforms, analytical platforms and digital resources for data management and data-model coupling [1]. According to the studied ecosystem, various qualitative or quantitative features of biodiversity are analysed such as: flora and resident soil organisms in grassland; soil microbial diversity; phytoplankton or fishes in freshwater ecosystems.

Produced by distributed platforms, most of the collected data are initially poorly standardised and managed in different information systems from flat files to relational databases. With the objective of ensuring technical and semantic interoperability a workflow is developed for data annotation and exploitation. We present here the general strategy deployed and provide examples of its implementation for biodiversity (zooplankton and phytoplankton) and environment data (water temperature and phosphorus concentration) collected from Lake Geneva, at a single sampling point referred to as SHL2, the deepest and pelagic part of the lake [2]. Data were collected once a month in winter and twice a month in other seasons.

2. Strategy and Workflow

The publication of open access data (F<u>A</u>IR) is nowadays easy. It is also becoming easier to make people aware of their existence (<u>F</u>AIR) thanks to the catalogues of the data repositories and the interoperability of the metadata standards they used. Nevertheless, their fine description, specific to the relevant thematic field and, even more so, their semantic interoperability (FA<u>IR</u>) are often weaker because requiring a significant investment and the use of shared vocabularies. To be reusable the data set must be described not only for the variables it contents but also for all the surrounding factors that influence variables values. The investment required is all the more important as this work is carried out late, i.e. at data publication step and not during the data life cycle, from acquisition, curation, processing...This analysis motivates the implementation of a workflow operating as far upstream as possible and whose genericity allows it to be used widely.

The strategy developed (**Figure 1**) is based on the modelling of the whole system (the experimental design in AnaEE) using an ontology as described in 3.1. The modeled system is used by a first pipeline [3] that automates the annotation process of relational database or flat files and generates the rdf triples (data lifting). A second pipeline is devoted to the exploitation of these semantic data by generating i) metadata records formatted according to the geospatial extension for the Data Catalog Vocabulary (GeoDCAT) standard and the ISO 19139 standard, and ii) Network Common Data Form (NetCDF) data files. It also offers a data publication service presently implemented for Dataverse repositories. The content of the dataset is determined by the user of 'pipeline 2' according to criteria defining the perimeter of interest and currently based on variables and variable categories, years, experimental platforms and networks, ecosystems. In addition to the general information on user and date, the defined perimeter is also used for the feeding of the metadata fields of the generated GeoDCAT record (abstract as dcterms:description, keyword and thesaurus as dcat:theme, dcat:contactPoint, dcterms:spatial and dcterms:temporal).

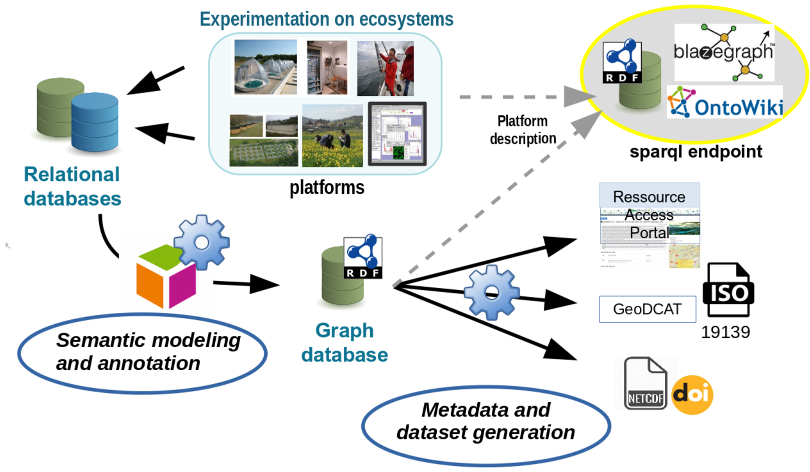

Figure 1: Semantic workflow for the management and valorisation of the data produced by the AnaEE platforms

3. Semantic Modelling of the Experimental/Observational System

3.1. Ontology Based Modelling

The choice of the reference ontology was determined by two main objectives: i) the need to model the whole observation or experimentation system and ii) the wish to achieve an atomic of "entity-quality" type. As a consequence we adopted the Extensible Observation Ontology (OBOE) [4] a formal ontology for capturing the semantics of scientific observation and measurement and developed in the frame of ecology [5].

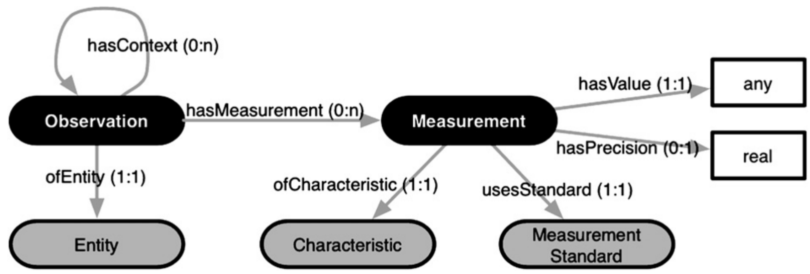

Figure 2: The core classes and properties of the Extensible Observation Ontology (OBOE), from Madin et al. 2007[5]).

In addition to the modelling of the observed variable (e.g: water temperature) OBOE can characterize, using the hasContext property, the context of an observation (e.g., space and time), as well as dependencies such as nested experimental observations. The main concepts in OBOE (**Figure 2**) include: - Observation: an event in which one or more measurements are taken - Measurement: the measured value of a property for a specific object or phenomenon (e.g., 12.5) - Entity: an object or phenomenon on which measurements are made (e.g., water) - Characteristic: the property being measured (e.g., Temperature - Standard: units and controlled vocabularies for interpreting measured values (e.g., degree celcius) - Protocol: the procedures followed to obtain measurements – Qualifier: statistical process (e.g., maximum; half-hourly-average) .

Our modelling covers the measured variables, the different components of the experimental context, from the sensor to the plot and the network, through the atomic decomposition of the observed entities, their characteristics and qualifiers, the units and the naming standards. In order to cover the whole system, OBOE extensions were realised mostly on experimental entity (network, site, plot, treatment), characteristics and standard naming for the experimental components (e.g. site names) and the observable properties (presently 350 variables, e.g dissolved orthophosphorus mass concentration).

3.2. Modelling of the Observable Property

Measured variables are often named succinctly in databases, in a common form, such as in column headers of flat files. They can be insufficient for a proper understanding of what is measured and often gather information of different natures. In order to solve these ambiguities, a semantic decomposition work is carried out for each variable, driven by the model and the terms of the ontology used. In our case, the principle is based on the identification of a characteristic measured according to a standard (a unit) for the observation of an entity, itself potentially contextualised by one or several observations of other entities.

As an illustration, the variable named "Ammonia nitrogen" in a physical chemistry database on lake data is decomposed as shown in **Figure 3**.

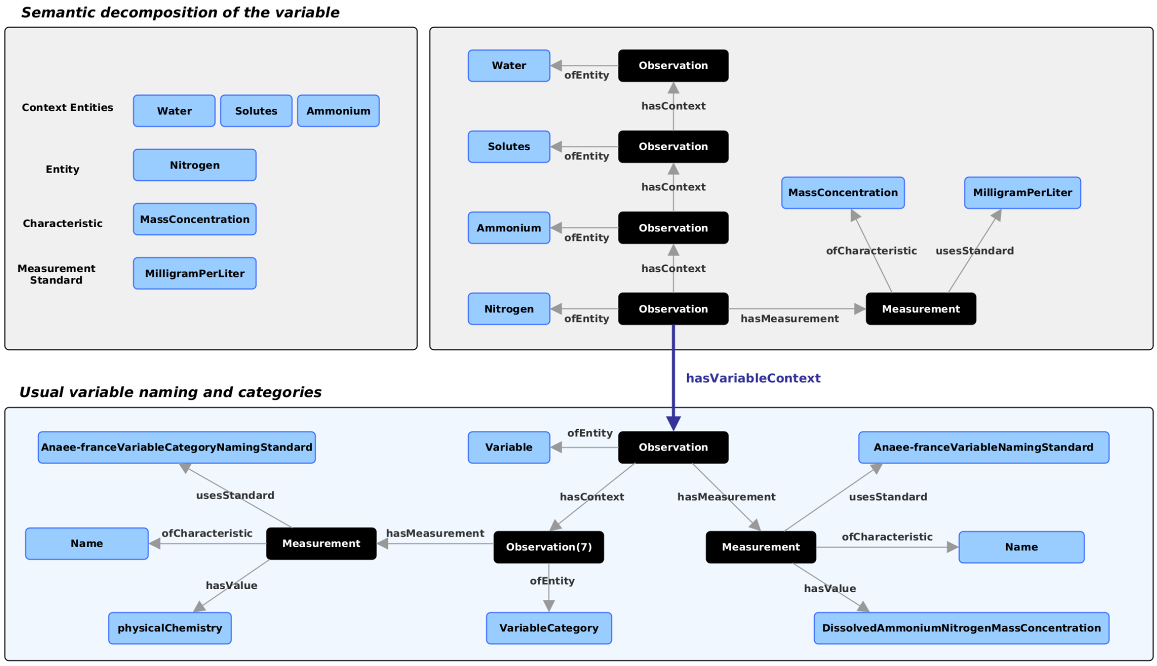

Figure 3: semantic decomposition of “Ammonia nitrogen” variable according to the ontology model and terms (upper left) with corresponding representation as a graph (upper right). The graph also provides the correspondence to the usual variable and category(ies) names (bottom box).

In addition to this semantic decomposition that will be used for annotation, the usual name of each variable is supplied to facilitate communication and to easily respond to certain uses such providing column headings of data files, metadata fields or drop-down lists of available variables. To do this, two naming standards have been added to the ontology as classes, one for the usual variable names (Anaee-France variableNamingStandard), the other for variable categories (Anaee-France variable category naming standard). Each of these classes contains a list of items as individuals. In the previous example of the variable "Ammonia nitrogen", the standardised usual name retained is "dissolved ammonium nitrogen mass concentration" and it is associated with the standard category "physical chemistry". A dedicated property named "hasVariableContext" has also been added to the ontology and is used to associate the result of the previous semantic decomposition with the standardised usual name of the variable and with one or several standard categories (**Figure 3**).

3.3. Generic Graph Models

The complete modelling of a graph for a measured variable implies adding other contextual elements such as spatial, temporal or location information through the reflexive property "hasContext" of the ontology (**Figure 4**). This results in a rich graph, more or less complex depending on the variables and the use cases.

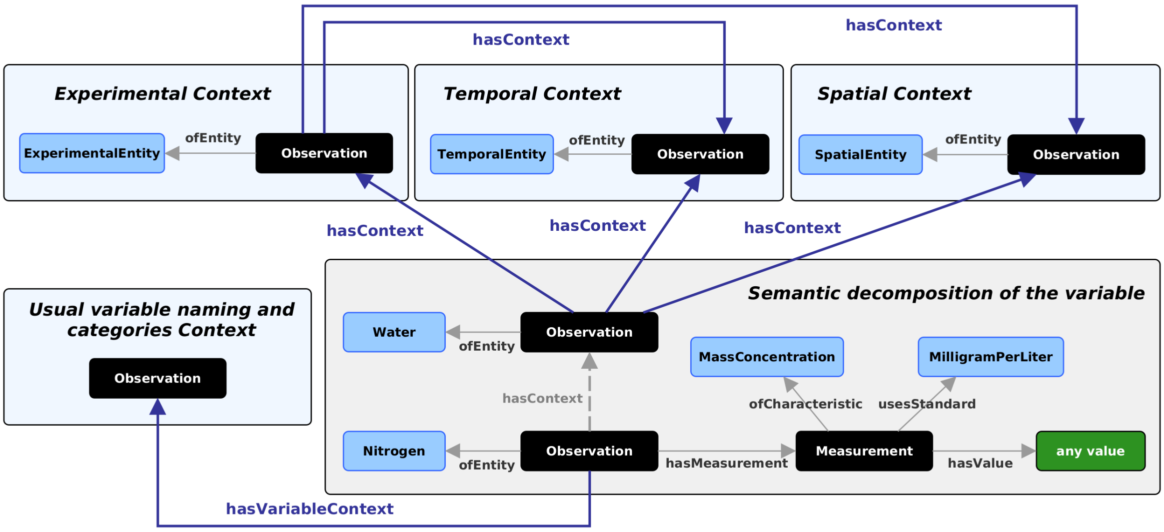

Figure 4 : Complete graph modelling overview for the "Ammonia nitrogen" example.

The modelling of a graph for one variable is most often suitable to other variables. Indeed if the values of the modelled information are different from one variable to another, their natures and their structures are common to several variables. Sets of variables can thus share the same graph structure defining a common pattern called "graph model". Thus, the graph initially constructed for one variable can be generalised to multiple ones by introducing a set of elements whose values vary according to the variable being processed. These so-called "dynamic elements" are the usual name of the variable, the entities observed, the characteristics, the measurement standards, the entities of the near context observations (e.g. matrix) or the thematic categories of the variables. The resulting graph model is then instantiated as a specific graph for each of the related variables. Whatever the variables, the graphs systematically contain similar graph structures from the semantic decomposition of the variables and their standardised naming. The common part shared by all the graph models is presented in **Figure 5**. The dynamic elements used for the instantiation of a graph for a given variable can be single (standardised variable name) or multiple (entities og the context observations and categories). Their values, resulting from the semantic analysis of the variables, are provided in a dedicated file, csv format, where each line corresponds to a single variable (Table 1). The instantiation of the graphs, for each variable, is then delegated to the first pipeline, responsible for the semantic annotation.

Thus, this generalisation of graphs through patterns called "graph models" and the use of an input parameters csv file makes it possible (i) to make the modelling effort more generic and (ii) to automate the instantiation of graphs for each variable to be processed.

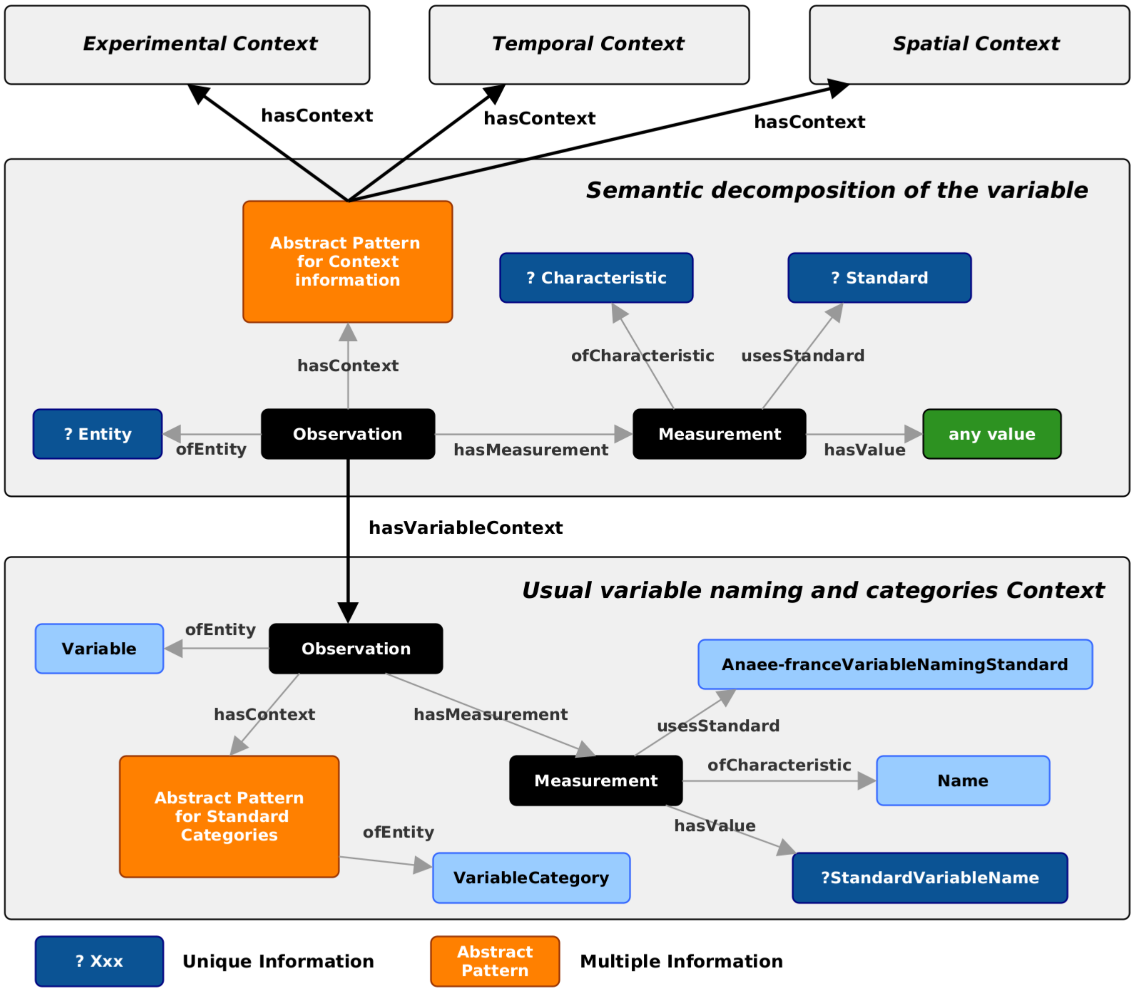

Figure 5: Graph model resulting from the generalisation, applicable to all the variables.

Table 1

Semantic decomposition and standard naming of variables. In blue, the unique elements and in orange, the potentially multiple elements.

Standard Variable Name	Category (ies)	Context(s)	Entity	Characteristic	Standard Measurement
Dissolved Ammonium Nitrogen Mass Concentration	Physical Chemistry	Water, Solutes, Ammonium	Nitrogen	Mass Concentration	Milligram Per Liter
Calcium Mass Concentration	Physical Chemistry	Water	Calcium	Mass Concentration	Milligram Per Liter
WaterPH	Physical Chemistry		Water	pH	pH unit
Biovolume	Biodiversity	Water	Zooplankton	Biovolume	MicroSquare Meter Per Millilitre
...	...	...	...	...	...

4. Implementation for Planktonic Biodiversity Data from Lakes

4.1. Variable Description

An application of the workflow described in section 2 was conducted on the Observatory on LAkes (OLA) database [2] by mobilising the total phytoplankton (“biovolume”) and the zooplankton per taxon (“sedimented volume”) as biodiversity data, and “water temperature” and “dissolved orthophosphorus” as environmental data. These variables were processed for the Lake Geneva data and the period 1974-2004. The semantic decomposition of the different variables involved was carried out with the domain experts and produced the file of Table 2.

Table 2
Semantic decomposition and standard naming of chosen variables for planktonic biodiversity data.

Standard Variable Name	Category(ies)	Context(s)	Entity	Characteristic	Standard Measurement
Dissolved Orthophosphorus Mass Concentration	Physical Chemistry, Phosphorus Cycle	Water, Solutes	Ortho phosphorus	Mass Concentration	Milligram Per Liter
Water Temperature	Physical Chemistry		Water	Temperature	Degree Celsius
Sedimented Volume	Biodiversity	Water	Zooplankton	Sedimented volume	Millilitre Per Square Meter
Biovolume	Biodiversity	Water	Phytoplankton	Biovolume	MicroSquare Meter Per Millilitre

4.2. Variable Modelling

Following the modelling principles described in section 3, two models generated for this implementation are illustrated below. The first (**Figure 6**) concerns the physico-chemical environment variable orthophosphorus and is an instantiated graph for this variable, automatically generated by a pipeline from a graph model. The second (**Figure 7**) is a graph model applicable to biodiversity data and used in this implementation for phyto and zooplankton data.

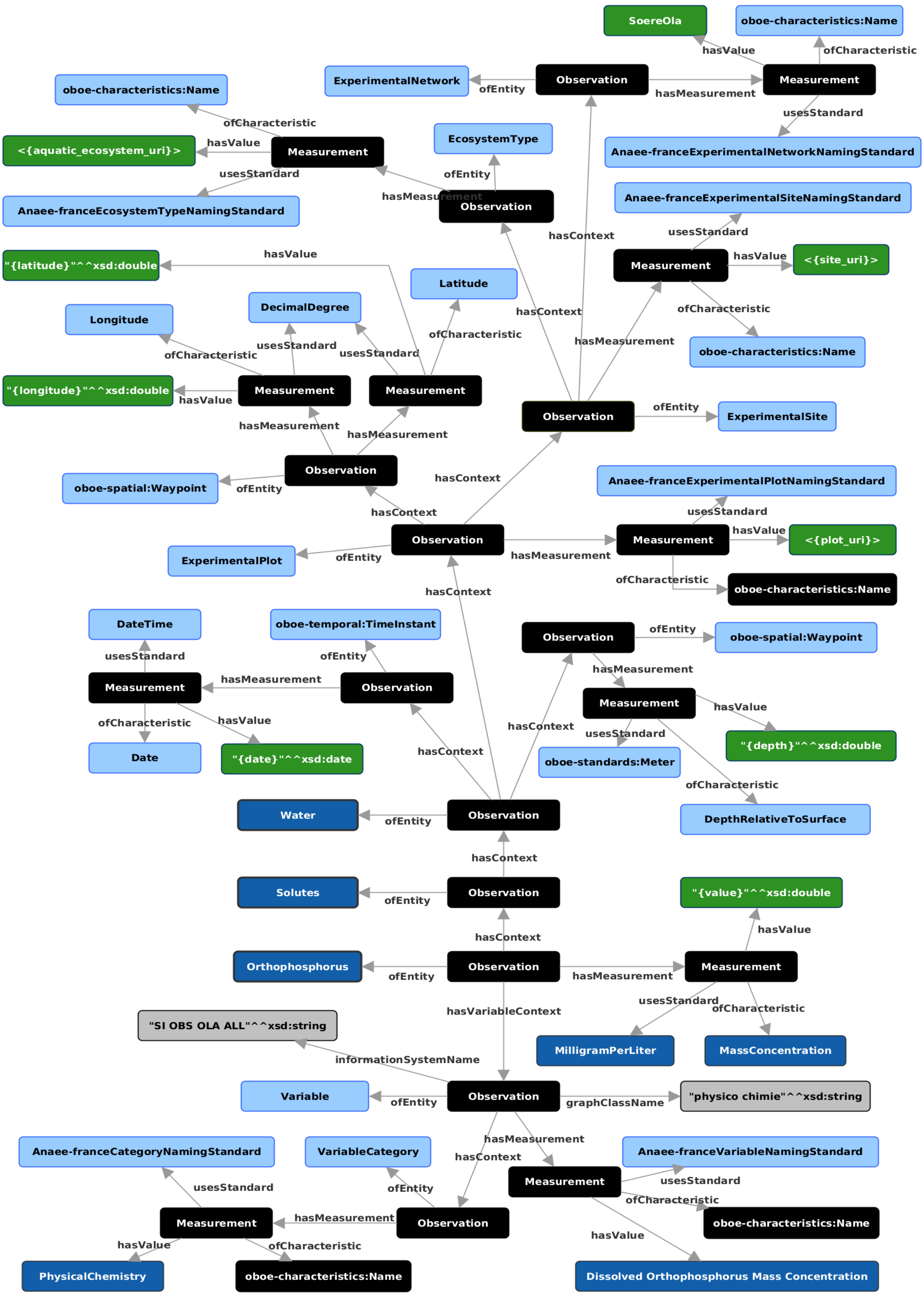

Figure 6: Instance of the physico-chemical data graph model for the variable dissolved orthophosphorus mass concentration.

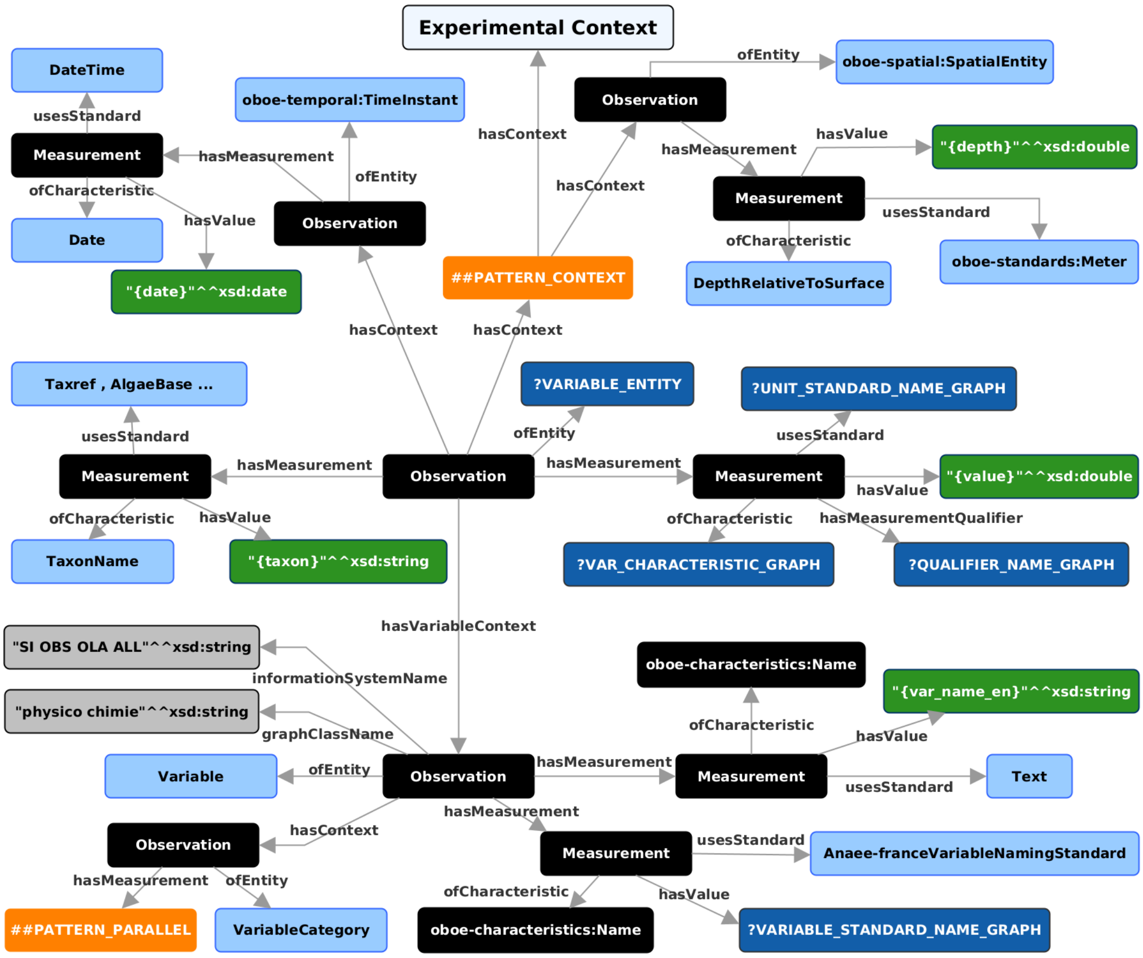

Figure 7: Graph model for plankton variables. The part specific to the experimental context is not detailed here.

4.3. Data Sets and Metadata Records

Defined on the perimeter 'Lake Geneva x [water temperature, orthophosphorus, zooplankton, phytoplankton] x [1974-2004]', a dataset and a metadata record were generated and published by the AnaEE workflow (doi:10.15454/XZWVM8). The discovery and exploitation metadata respectively present in the GeoDCAT file (Box 1) and the NetCDF file header (Box 2) were automatically filled in with data from the database and semantic information from the graph model and from the ontology.

Box 1

Extract from the GeoDCAT metadata record.

```
      <dcterms:description xml:lang="en">The data set was produced from experimentation(s) from the network(s) SOERE OLA on
the site(s) Lake Geneva in the ecosystem(s) lake.
Measurements are about the following variables:
dissolved orthophosphorus mass concentration,
water temperature,
zooplankton biovolume,
phytoplankton biovolume
....to be completed</dcterms:description>

      <dcat:theme>
         <rdf:Description>
            <skos:prefLabel xml:lang="en">water temperature</skos:prefLabel>
            <skos:inScheme>
               <skos:ConceptScheme rdf:about="http://opendata.inra.fr/anaeeThes/">
                  <dcterms:title xml:lang="en">anaeeThes</dcterms:title>
                  <dcterms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2017-06-13</dcterms:issued>

      <dcat:contactPoint>
         <vcard:Individual>
            <vcard:fn xml:lang="en">Frederic Rimet</vcard:fn>
            <vcard:organization-name xml:lang="en">INRAE</vcard:organization-name>
            <vcard:hasTelephone rdf:resource="tel:+33450267874"/>
            <vcard:hasEmail rdf:resource="mailto:frederic.rimet@inrae.fr"/>
```

```
            <vcard:hasAddress>
               <vcard:Address>
                  <vcard:street-address>UMR CARRTEL 75bis Avenue de Corzent</vcard:street-address>
                  <vcard:locality>Thonon-les-Bains</vcard:locality>
                  <vcard:postal-code>74200</vcard:postal-code>
                  <vcard:country-name>France</vcard:country-name>
               </vcard:Address>
```

Box 2

Extract from the NetCDF file header for water temperature and orthophosphorus concentration. Water temperature (‘Var0’) is expressed in degree Celcius and provides data collected in one experimental plot (‘Shl2Platform’, 46.45°N, 6.59°E) at 25 depths (Dim0) expressed in meter and for 559 dates (Dim1)

```
dimensions:
        Var0Dim0 = 25 ;
        Var0Dim1 = 559 ;
        Var1Dim0 = 25 ;
        Var1Dim1 = 573 ;
variables:
        string Var0Dim0(Var0Dim0) ;
                Var0Dim0:characteristic = "http://opendata.inra.fr/anaeeOnto#DepthRelativeToSurface" ;
                Var0Dim0:entity = "http://ecoinformatics.org/oboe/oboe.1.2/oboe-spatial.owl#Waypoint" ;
                Var0Dim0:standard = "http://ecoinformatics.org/oboe/oboe.1.2/oboe-standards.owl#Meter" ;
                Var0Dim0:description = "depth relative to surface of Waypoint in meter" ;
        string Var0Dim1(Var0Dim1) ;
                Var0Dim1:characteristic = "http://opendata.inra.fr/anaeeOnto#Date" ;

.../…

                Var1Dim1:description = "date of time instant in ISO 8601 DateTime" ;
        double Var0(Var0Dim0, Var0Dim1) ;
                Var0:characteristic = "http://ecoinformatics.org/oboe/oboe.1.2/oboe-characteristics.owl#Temperature" ;
                Var0:qualifier = "" ;
                Var0:entity = "http://opendata.inra.fr/anaeeOnto#Water" ;
                Var0:standard = "http://ecoinformatics.org/oboe/oboe.1.2/oboe-standards.owl#Celsius" ;
                Var0:description = " temperature of water in degree Celsius" ;
                Var0:name_of_experimental_site_in_text = "Léman" ;
                Var0:name_of_variable_in_text = "Temperature" ;
                Var0:name_of_ecosystem_type_in_Anaee-France_ecosystem_type_naming_standard = "http://opendata.inra.fr/anaeeOnto#Lake" ;
                Var0:name_of_experimental_network_in_Anaee-France_experimental_network_naming_standard = "http://opendata.inra.fr/anaeeOnto#SoereOla" ;
                Var0:name_of_experimental_plot_in_Anaee-France_experimental_plot_naming_standard = "http://opendata.inra.fr/anaeeOnto#Shl2Platform" ;
                Var0:name_of_experimental_site_in_Anaee-France_experimental_site_naming_standard = "http://opendata.inra.fr/anaeeOnto#LakeGeneva" ;
                Var0:name_of_variable_in_Anaee-France_variable_naming_standard = "http://opendata.inra.fr/anaeeOnto#WaterTemperature" ;
                Var0:latitude_of_Waypoint_in_decimal_degree = "46.453457" ;
                Var0:longitude_of_Waypoint_in_decimal_degree = "6.5942335" ;
        double Var1(Var1Dim0, Var1Dim1) ;

../..

// global attributes:
                :lineage = "The dataset was generated, formatted and published using AnaEE semantic services and vocabularies" ;
data:

 Var0Dim0 = "0.0", "10.0", "100.0", "15.0", "150.0", "2.5", "20.0", "200.0",
    "225.0", "25.0", "250.0", "275.0", "280.0", "285.0", "290.0", "295.0",
    "30.0", "300.0", "305.0", "309.0", "35.0", "40.0", "5.0", "50.0", "7.5" ;

 Var0Dim1 = "1974-01-14", "1974-02-18", "1974-03-18", "1974-04-22",
    "1974-05-13", "1974-06-17", "1974-07-15", "1974-08-19", "1974-09-16",
```

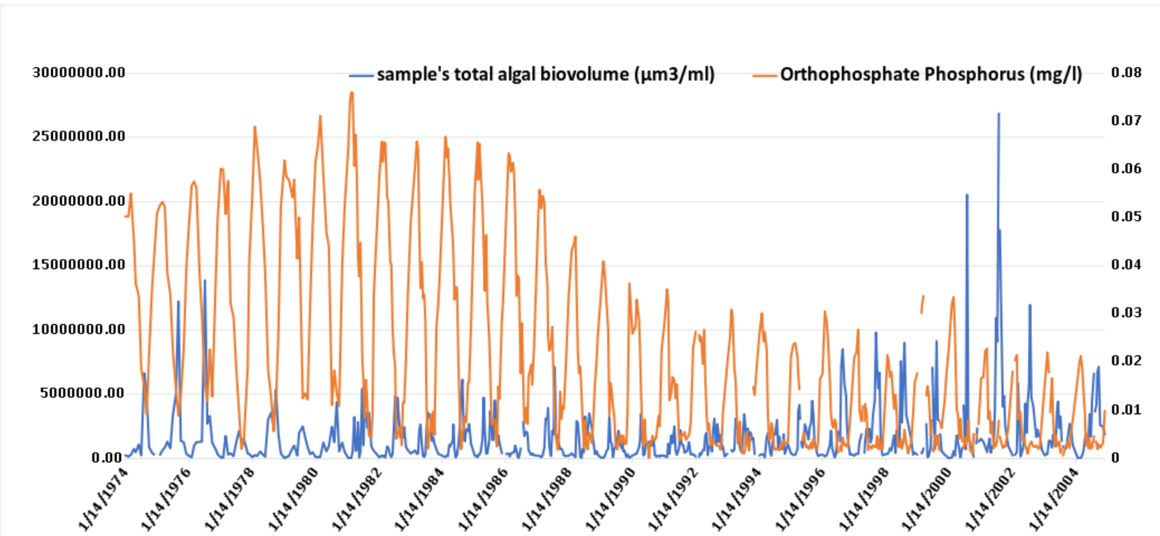

Figure 8: Relationship between orthophosphorus and phytoplankton biovolume (all species) in Lake Geneva, based on semantic data produced by the workflow described in present document.

Figure 8 illustrates the seasonality of phosphorus and also the drop in concentrations induced by restoration measures aimed at limiting P inputs to the lake. The decrease in zooplankton means less grazing pressure on phytoplankton and therefore less effective regulation [6].

5. Conclusion and Perspectives

Although of main importance for interoperability, the implementation of the semantic characterisation of data and metadata is a difficult task to implement as it is very costly in terms of shared technical and semantic resources and in terms of annotation of the data and their acquisition contexts.

The work presented and carried out in the framework of AnaEE and the ENVRI-plus/FAIR projects aims to facilitate this implementation. The genericity of OBOE allows a modelling of the whole system. Its Entity-Observation-Characteristics model also allows an alignment with other ontologies based on entity property relationships (O&Ḿ, SOSA...). The convergence of these models is the subject of ongoing work in the framework of the RDA I-Adopt WG. The challenge here is not only to develop interoperability within the theoretical communities but also between the domains.

The enrichment of OBOE modelling, for components concerning people and sensors via specific ontologies (FOAF, SOS, etc.) is a short-term prospect. The PROV ontology (PROV-O) will be used to model provenance elements from data acquisition to dataset publication, triples being generated by the developed pipelines, from a dedicated graph.

Although the deployment of the AnaEE workflow is currently limited to French experimental platforms and does not cover all variables, the strategy being pursued is to implement it as systematically as possible, based on the experience gained.

6. Acknowledgements

The work was co-funded by AnaEE-France, French program “Investissements d’Avenir” (ANR-11-INBS-0001), the European Horizon 2020 ENVRIplus project (No 654182) and ENVRI-FAIR project (No 824068) and the D2KAB project (ANR-18-CE23-0017). It notably relies the contribution of the AnaEE semantic working group (*) who develops the vocabularies: i) the anaeeThes thesaurus (http://agroportal.lirmm.fr/ontologies/ANAEETHES) ii) the anaee ontology extension of OBOE (ongoing publication). We are also grateful to the INRAE-CARRTEL technical and scientific team that collected and provided the data from the Geneva Lake Observatory.

(*) Pichot C. (1), Callou C. (2), Chanzy A. (3), Clastre P. (1), Clavreul A. (3), El-Hamadry M. (4), Evtimova M. (1), Jaillet B. (1), Lafolie F. (3), Le Gaillard J.-F. (5), Martin C. (2), Massol F. (5), Maurice D. (6), Moitrier N. (3), Monet G. (7), Raynal H. (4), Schellenberger A. (8), Yahiaoui R. (8), Aïvayan E. (3), Beudez N. (3), Léturgie A. (1)

1. INRAE URFM, 2. CNRS UMS BBEES, 3. INRAE UMR EMMAH, 4. INRAE MIAT, 5. UMS CEREEP-Ecotron, 6. INRAE UMR SILVA, 7. INRAE UMR CARRTEL, 8. INRAE INFOSOL.

7. References

[1] J. Clobert, A. Chanzy, J.F. Le Galliard, A. Chabbi, L. Greiveldinger, T. Caquet, M. Loreau, C. Mougin, C. Pichot, J. Roy, L. Saint-André, How to Integrate Experimental Research Approaches in Ecological and Environmental Studies: AnaEE France as an Example. Front. Ecol. Evol. 6:43 (2018). doi: 10.3389/fevo.2018.00043, hal-01773144.

[2] F. Rimet, O. Anneville, D. Barbet, C. Chardon, L. Crepin, et al., The Observatory on LAkes (OLA) database: Sixty years of environmental data accessible to the public. Journal of Limnology, 79-2 (2020): 164-178. doi:10.4081/jlimnol.2020.1944, hal-02916312.

[3] C. Pichot, D. Maurice, P. Clastre, B. Jaillet, R. Yahiaoui, Pipelines for semantic annotation and FAIR data production. RDA 17th Plenary Meeting, Apr 2021, Edinburgh, United Kingdom. hal-03234314.

[4] M. Schildhauer, M. B. Jones, S. Bowers, J. Madin, S. Krivov, D. Pennington, F. Villa, B. Leinfelder, C. Jones, M. O'Brien, OBOE: the Extensible Observation Ontology, version 1.2. KNB Data Repository (2016). doi:10.5063/F1125R0F

[5] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, F. Villa, An ontology for describing and synthesizing ecological observation data, Ecological Informatics 2 (2007): 279–296. doi:10.1016/j.ecoinf.2007.05.004

[6] O. Anneville, C. Chang, G. Dur, S. Souissi, F. Rimet, C. Hsieh, The paradox of re-oligotrophication: the role of bottom–up versus top–down controls on the phytoplankton community, Oikos 128-11 (2019): 1666-1677.