Developing semantic interoperability in ecology and ecosystem studies : semantic modeling and annotation for FAIR data production

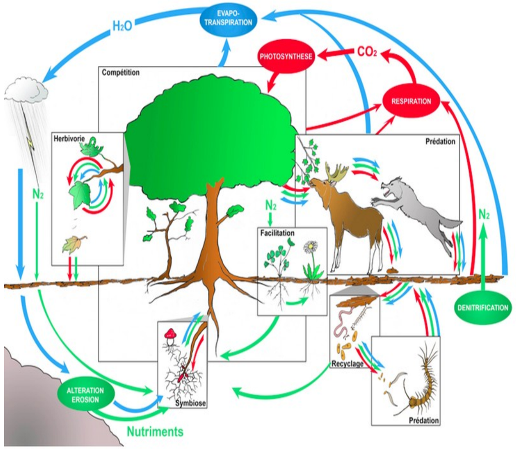
Christian PICHOT,
Philippe CLASTRE, Benjamin JAILLET,
Damien MAURICE, Ghislaine MONET, Rachid YAHIAOUI.

Callou C., Chanzy A., Clavreul A., El-Hamadry M., Evtimova M., Lafolie F., Le Gaillard J.-F.,
Martin C., Massol F., Moitrier N., Raynal H. , Schellenberger A., Aïvayan E., Beudez N., Léturgie A.

Séminaire INRAE Semantic Linked Data
11 au 14 octobre 2021 - Domaine du Lazaret - Sète

Biodiversity and ecosystem studies

Rationale



Ecosystem study requires complex research and deals with heterogeneous, varied and widespread data.

The proper understanding and interoperability of
the information sources remains one of the greatest challenges

A Research Infrastructure for experimentation on ecosystems

Experimentation platforms

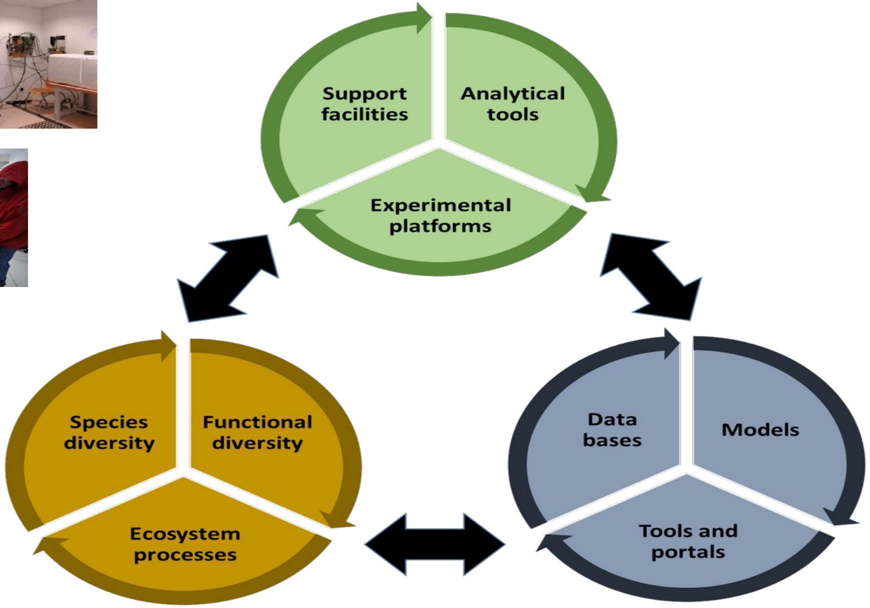


Plate-forme de modélisation et de simulation des agro-écosystèmes

Mobile instruments and analytical plaforms

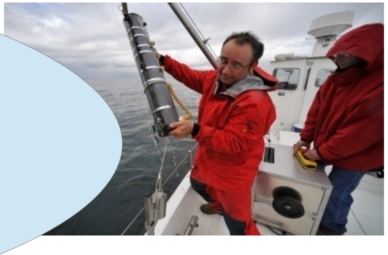
Data and models

AnaEE
Analysis and Experimentation on Ecosystems

How to deal with data heterogeneity?

plus ENVRI

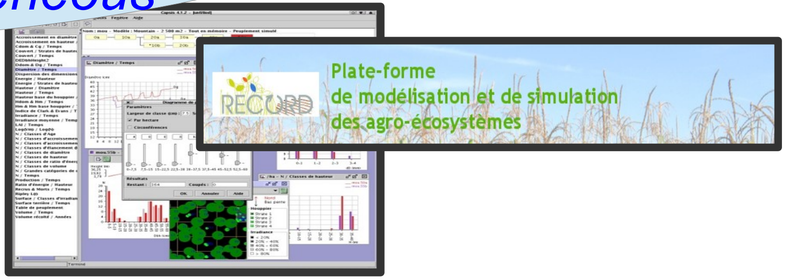
AnaEE
Analysis and Experimentation on Ecosystems
France
Laboratoire Mobile
Mesure de CO$_2$, CH$_4$, N$_2$O et H$_2$O



Managing data for:

☛ discovery
☛ access to resources

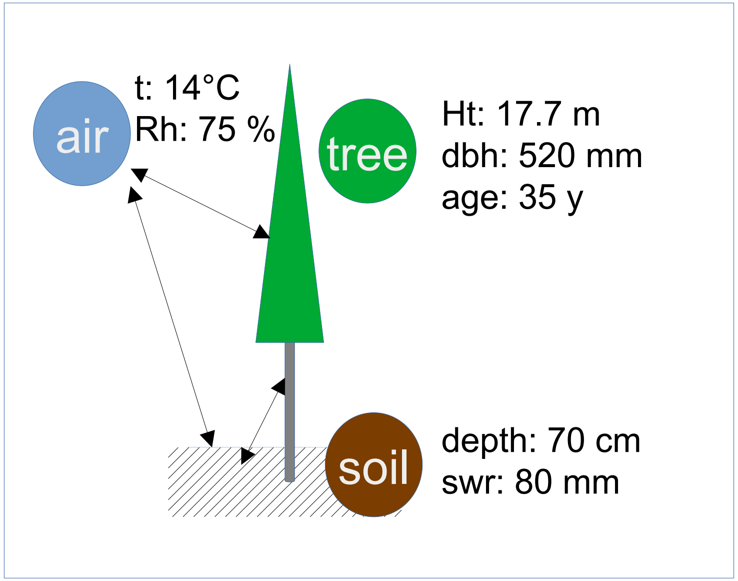
...distributed and heterogeneous



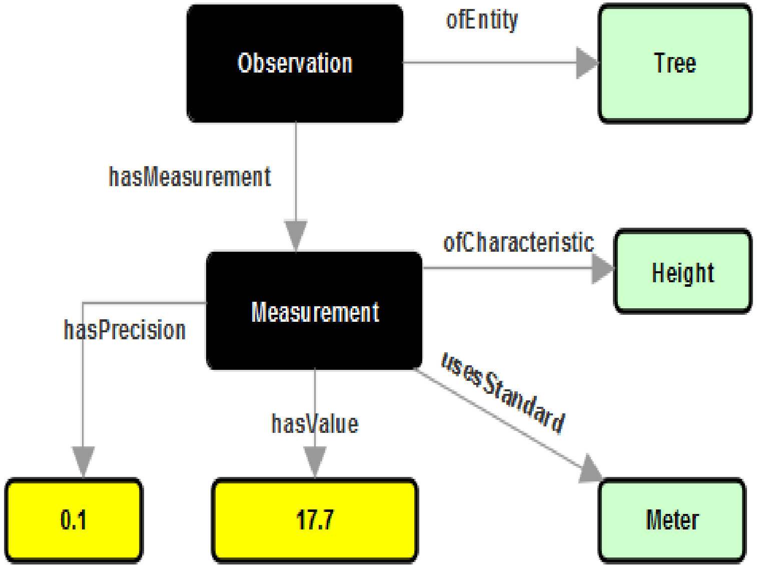
Developing semantic interoperability

Method

1) Identify
- the components of the system
- and their relationships



2) Model the system
using semantic vocabularies

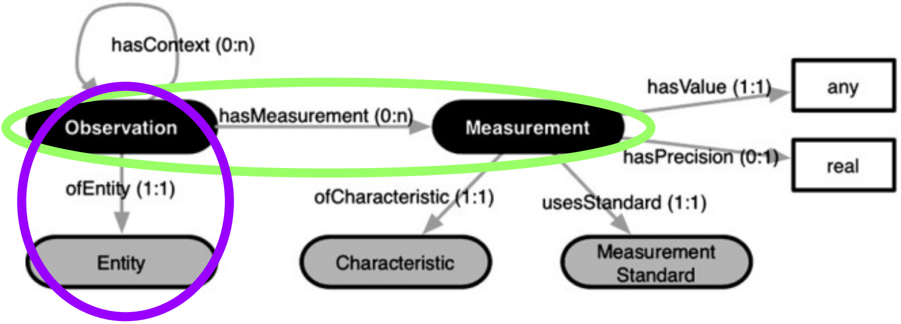


Developing semantic interoperability

Implementation

AnaEE* RI as scientific context:
The Research Infrastructure offers services for experimentation on continental ecosystems

OBOE* as ontological framework:
The ontology provides the atomic elements for modeling observations

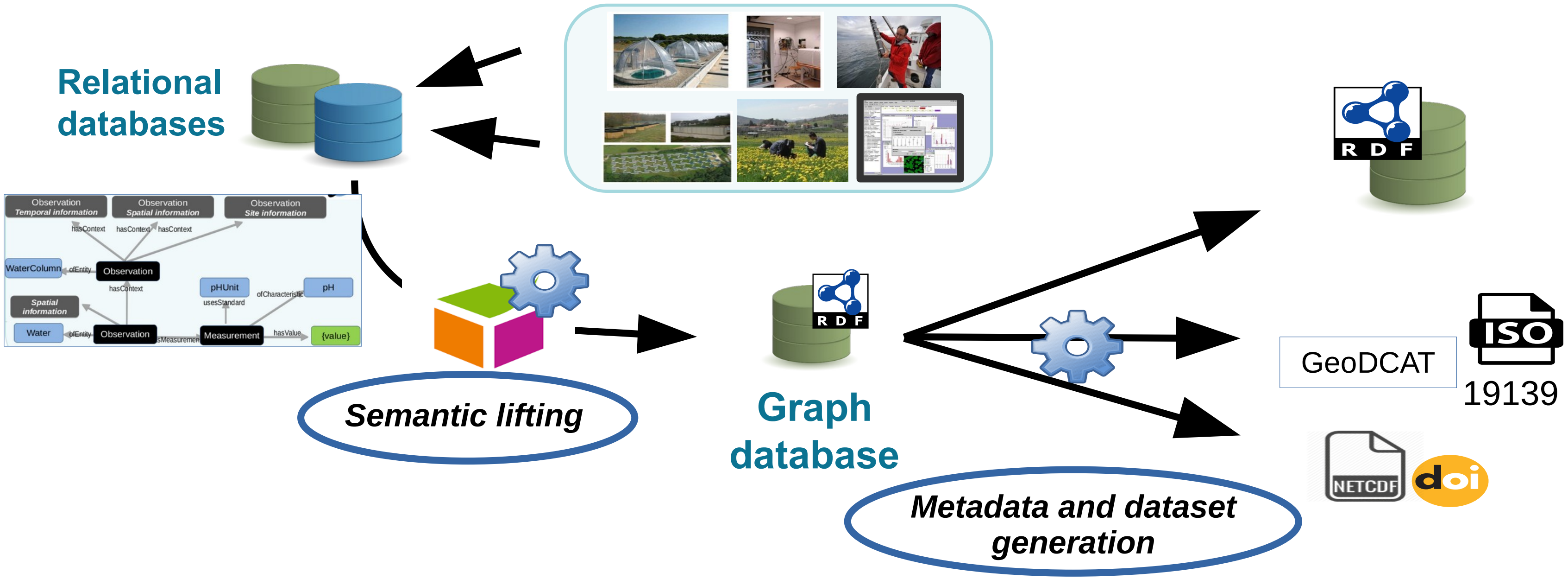


*Mark Schildhauer, Matthew B. Jones, Shawn Bowers, Joshua Madin, Sergeui Krivov, Deana Pennington, Ferdinando Villa, Benjamin Leinfelder, Christopher Jones, and Margaret O'Brien. 2016. OBOE: the Extensible Observation Ontology, version 1.2. KNB Data Repository. doi:10.5063/F1125R0F

Developing semantic interoperability

Semantic lifting and data exploitation

Graph patterns and variable semantic descriptions are processed by a pipeline for semantic lifting of the data before their exploitation



Relational databases

Semantic lifting

Graph database

Metadata and dataset generation

GeoDCAT

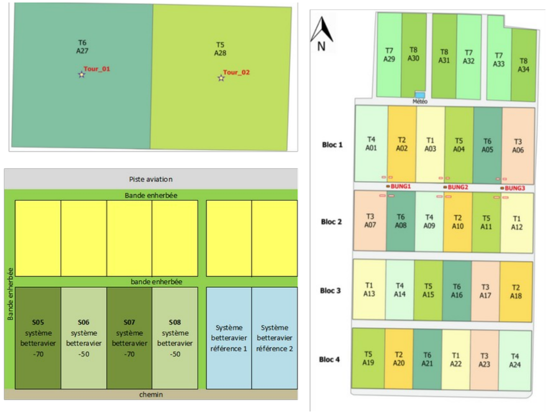
19139

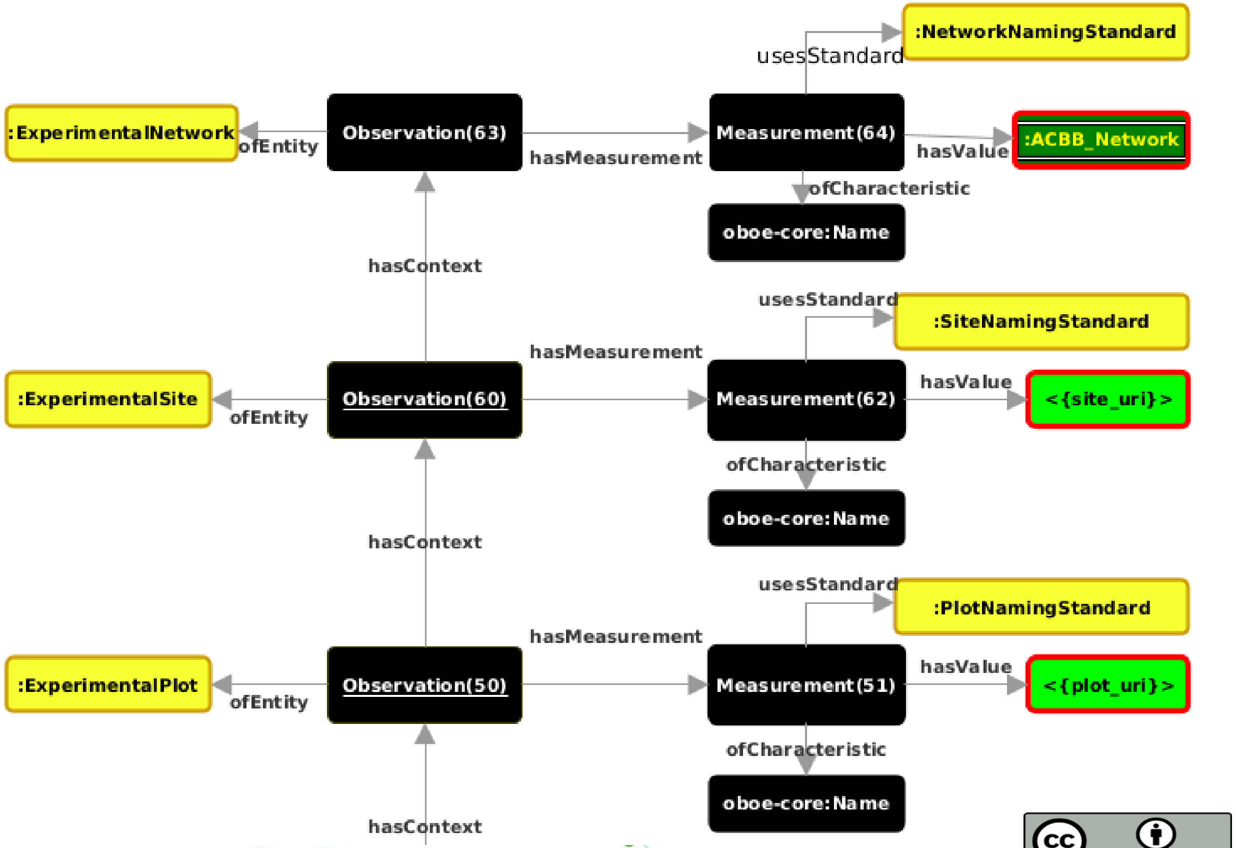
Developing semantic interoperability

Implementation

17th Plenary

Modeling the experimental context





:NetworkNamingStandard

usesStandard

:ExperimentalNetwork

ofEntity

Observation(63)

hasMeasurement

Measurement(64)

hasValue

:ACBB_Network

ofCharacteristic

oboe-core:Name

hasContext

usesStandard

:SiteNamingStandard

hasMeasurement

:ExperimentalSite

ofEntity

Observation(60)

Measurement(62)

hasValue

<{site_uri}>

ofCharacteristic

oboe-core:Name

hasContext

usesStandard

:PlotNamingStandard

hasMeasurement

:ExperimentalPlot

ofEntity

Observation(50)

Measurement(51)

hasValue

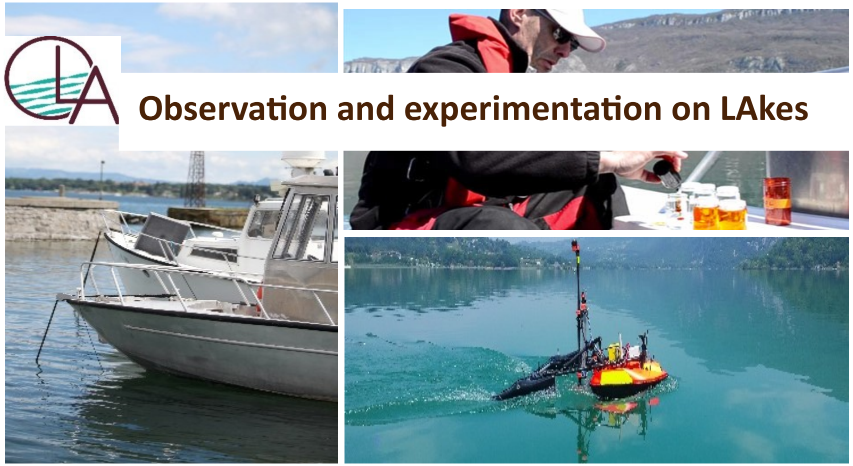
<{plot_uri}>

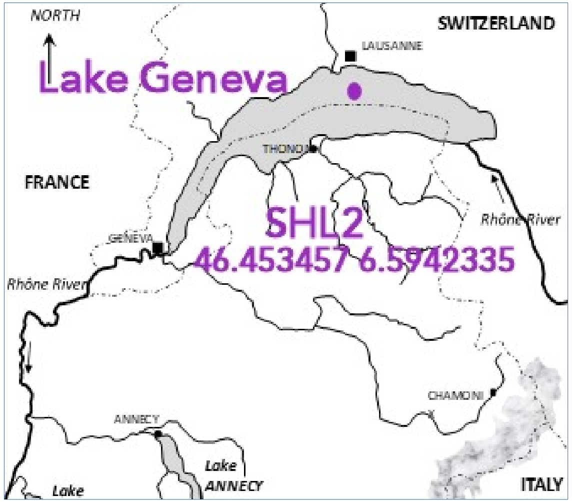
ofCharacteristic

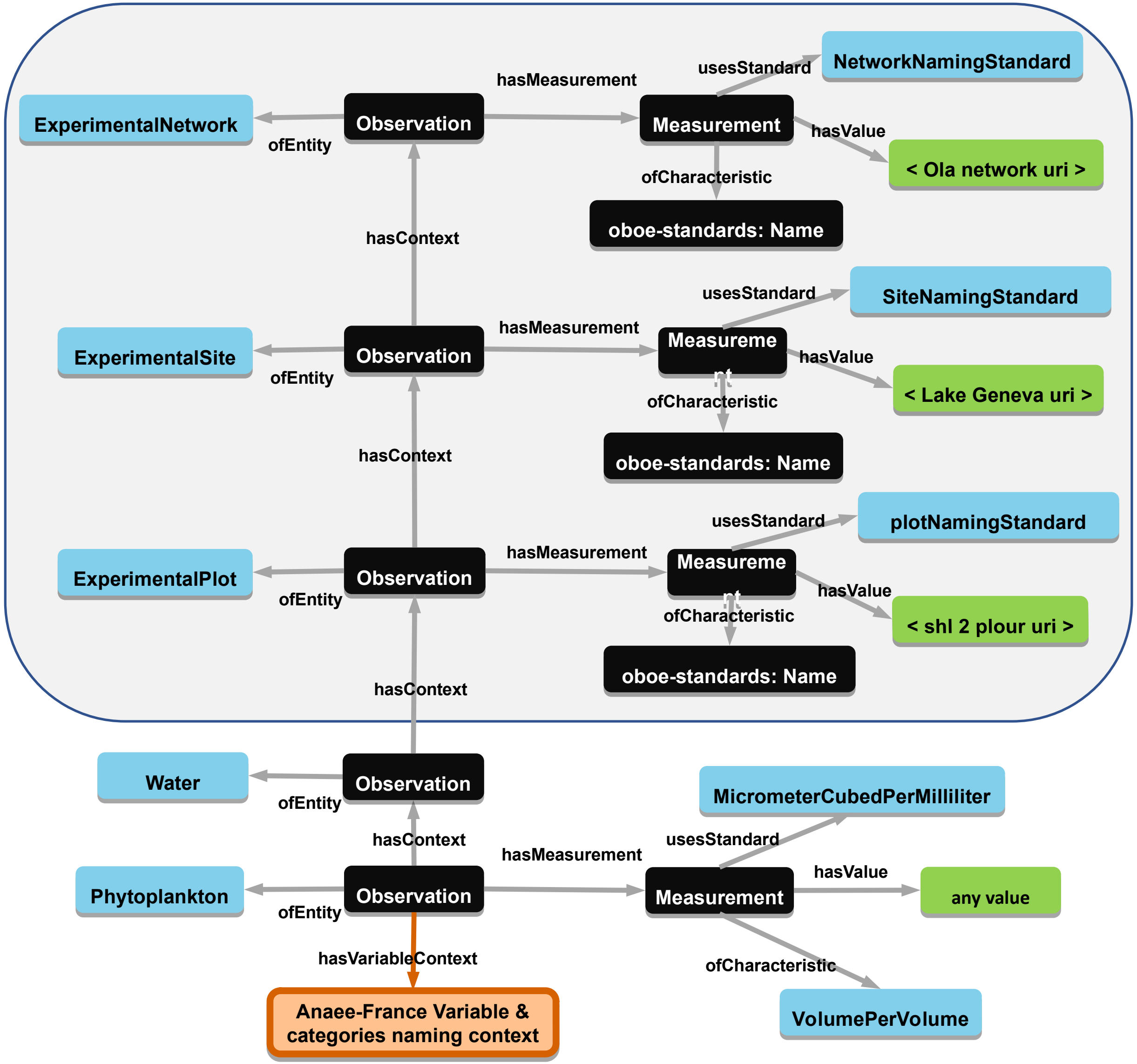
oboe-core:Name

hasContext

Modelling the experimental context







ExperimentalNetwork ← ofEntity ← Observation → hasMeasurement → Measurement

Measurement → usesStandard → NetworkNamingStandard

Measurement → hasValue → < Ola network uri >

Measurement → ofCharacteristic → oboe-standards: Name

Observation → hasContext → Observation

ExperimentalSite ← ofEntity ← Observation → hasMeasurement → Measureme nt

Measureme nt → usesStandard → SiteNamingStandard

Measureme nt → hasValue → < Lake Geneva uri >

Measureme nt → ofCharacteristic → oboe-standards: Name

Observation → hasContext → Observation

ExperimentalPlot ← ofEntity ← Observation → hasMeasurement → Measureme nt

Measureme nt → usesStandard → plotNamingStandard

Measureme nt → hasValue → < shl 2 plour uri >

Measureme nt → ofCharacteristic → oboe-standards: Name

Observation → hasContext → Observation

Water ← ofEntity ← Observation

Observation → hasContext → Observation

Phytoplankton ← ofEntity ← Observation → hasMeasurement → Measurement

Measurement → usesStandard → MicrometerCubedPerMilliliter

Measurement → hasValue → any value

Measurement → ofCharacteristic → VolumePerVolume

Observation → hasVariableContext → Anaee-France Variable & categories naming context

Modelling the measured variable

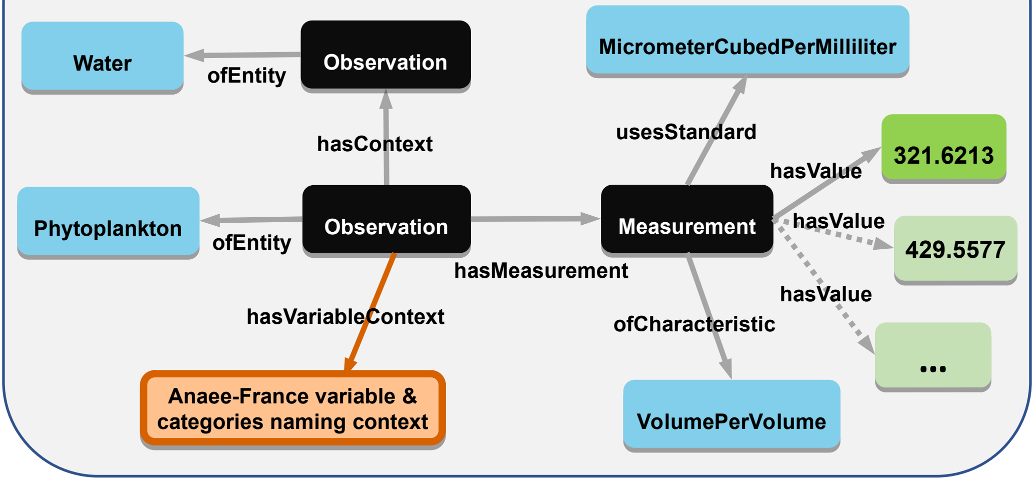
1. data from flat files or database with ambiguities

site	plot	date	species	vol.	prof min	prof max
Lake geneva	shl2	11/01/2012	Nitzschia sp.	321.6213	0	18
Lake geneva	shl2	11/01/2012	Ankyra judayi	429.5577	0	18
Lake geneva	shl2	11/01/2012	Cyclotella costei	1519.8612	0	18
Lake geneva	shl2	11/01/2012	Bicoeca ovata	12641.2	0	18
...	...	...	...	...	...	...

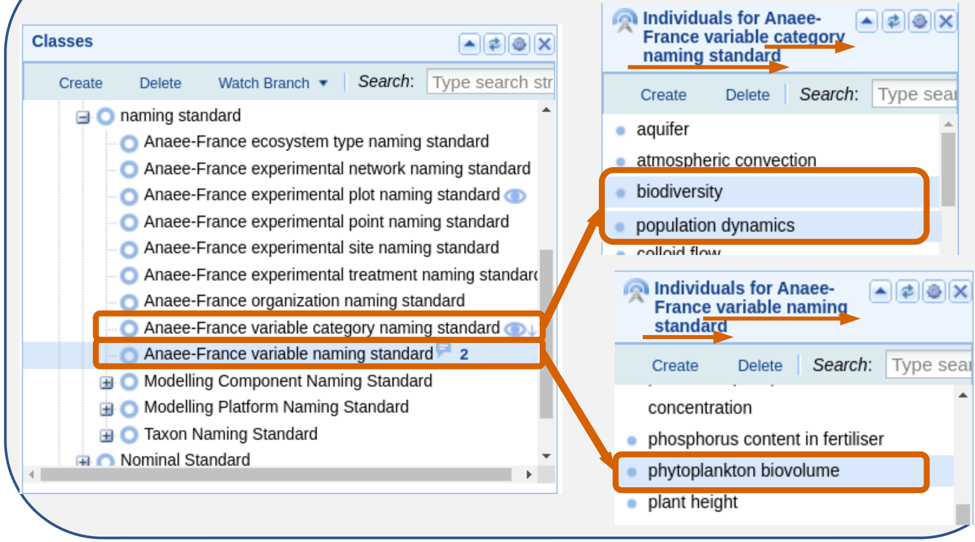
2. variable semantic description driven by OBOE ontology

Context Entity	Water
Entity	Phytoplankton
Characteristic	VolumePerVolume
Measurement Standard	MicrometerCubedPerMilliliter

3. variable semantic graph driven by OBOE model



4. OBOE extension for variable usual names and categories


Classes
Create
Delete
Watch Branch ▾
Search:
Type search str
Individuals for Anaee-France variable category naming standard
Create
Delete
Search:
Type sear
Individuals for Anaee-France variable naming standard
Create
Delete
Search:
Type sear

- naming standard
  - Anaee-France ecosystem type naming standard
  - Anaee-France experimental network naming standard
  - Anaee-France experimental plot naming standard
  - Anaee-France experimental point naming standard
  - Anaee-France experimental site naming standard
  - Anaee-France experimental treatment naming standard
  - Anaee-France organization naming standard
  - Anaee-France variable category naming standard
  - Anaee-France variable naming standard 2
- Modelling Component Naming Standard
- Modelling Platform Naming Standard
- Taxon Naming Standard
- Nominal Standard

Individuals for Anaee-France variable category naming standard:

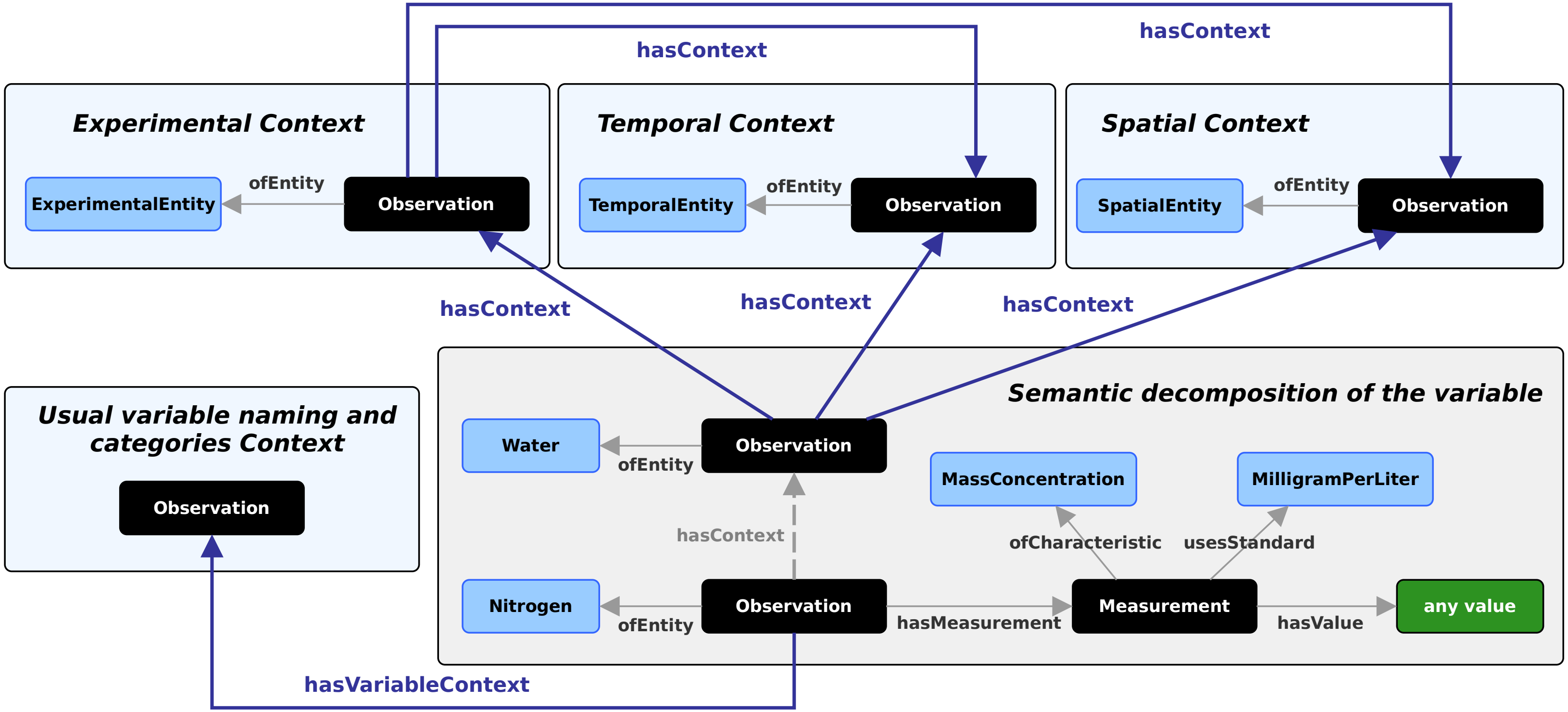
- aquifer
- atmospheric convection
- biodiversity
- population dynamics
- colloid flow

Individuals for Anaee-France variable naming standard:

- concentration
- phosphorus content in fertiliser
- phytoplankton biovolume
- plant height

Generic graph models

complete graph overview for **ONE** variable : phytoplankton biovolume



how not to duplicate this work for each variables?

Generic graph models for several variables

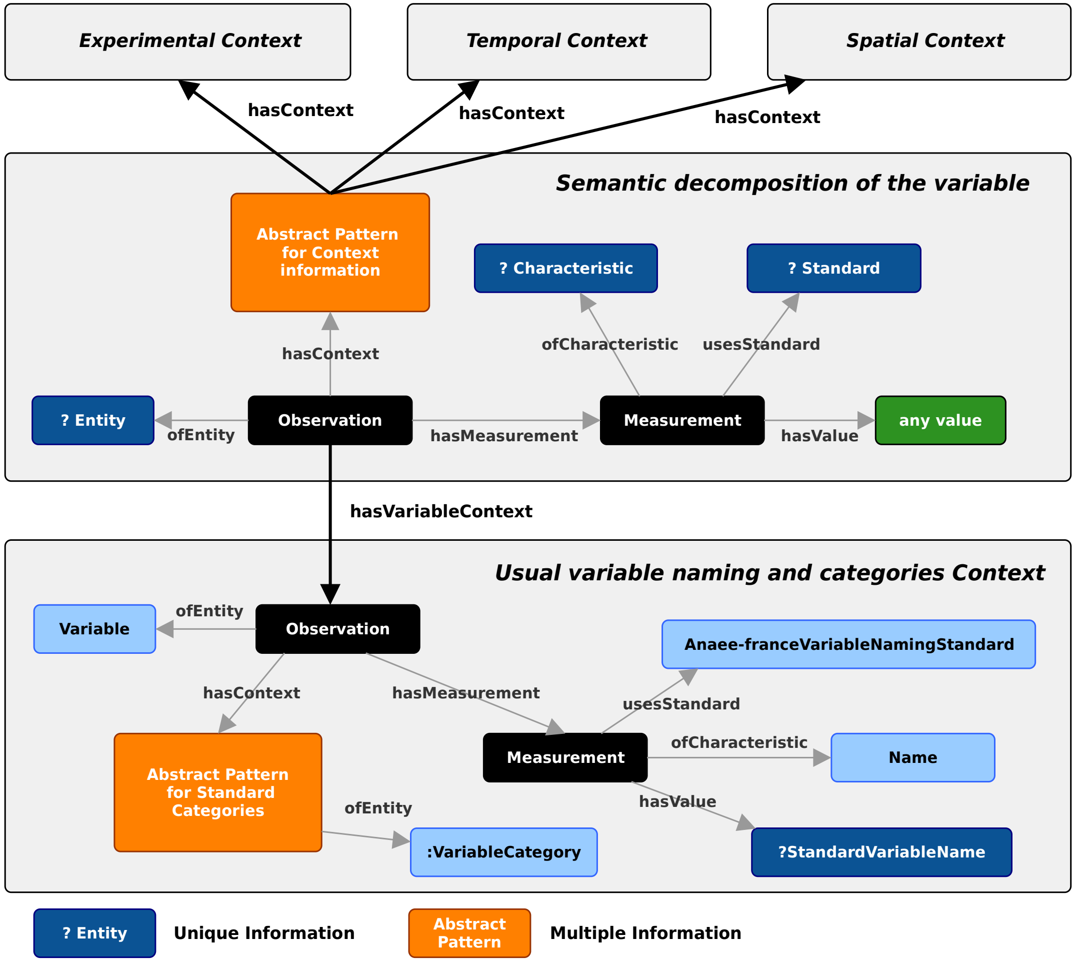
→ dynamically instantiated annotation patterns on several variables

→ depends on the relational database model : data from several variables must be managed in a similar way

Identification of the common structure of graphs for several variables, composed of:

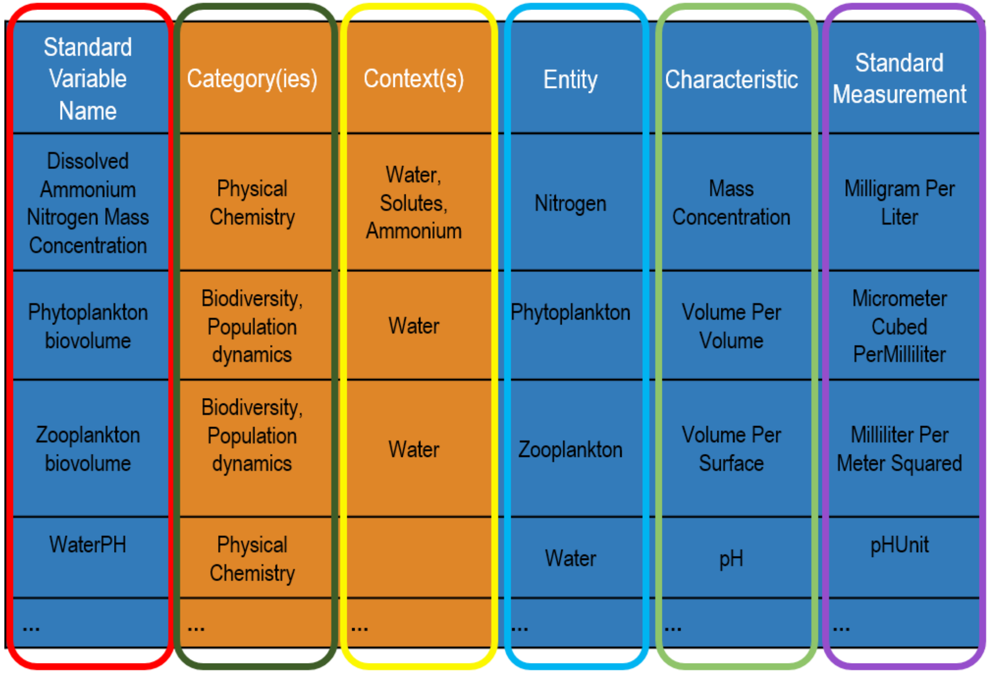
- single-value nodes dynamically instantiated per variable (blue)

- optional nodes, single or multiple according to the variable and whose values are dynamic, forming portions of the final graph (orange)



Generic graph models

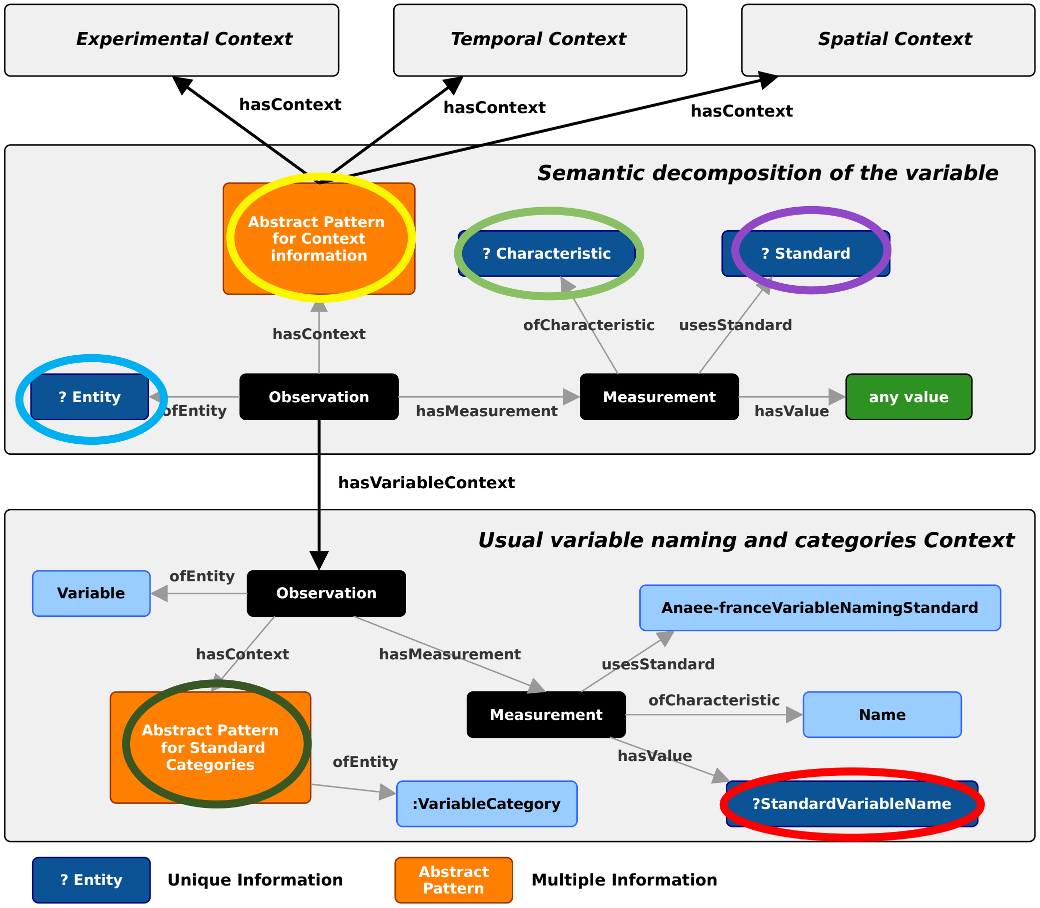
Semantic description and standard naming of variables, (1 line per variable)



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
Phytoplankton biovolume	Biodiversity, Population dynamics	Water	Phytoplankton	Volume Per Volume	Micrometer Cubed PerMilliliter
Zooplankton biovolume	Biodiversity, Population dynamics	Water	Zooplankton	Volume Per Surface	Milliliter Per Meter Squared
WaterPH	Physical Chemistry		Water	pH	pHUnit
...	...	...	...	...	...

+

Semantic annotation model



Application for planktonic biodiversity data from lakes

- in da dark blue d ora orange ynamic elements from the semantic description of the variables

- in green values stored in the relational database

+

rules for uri

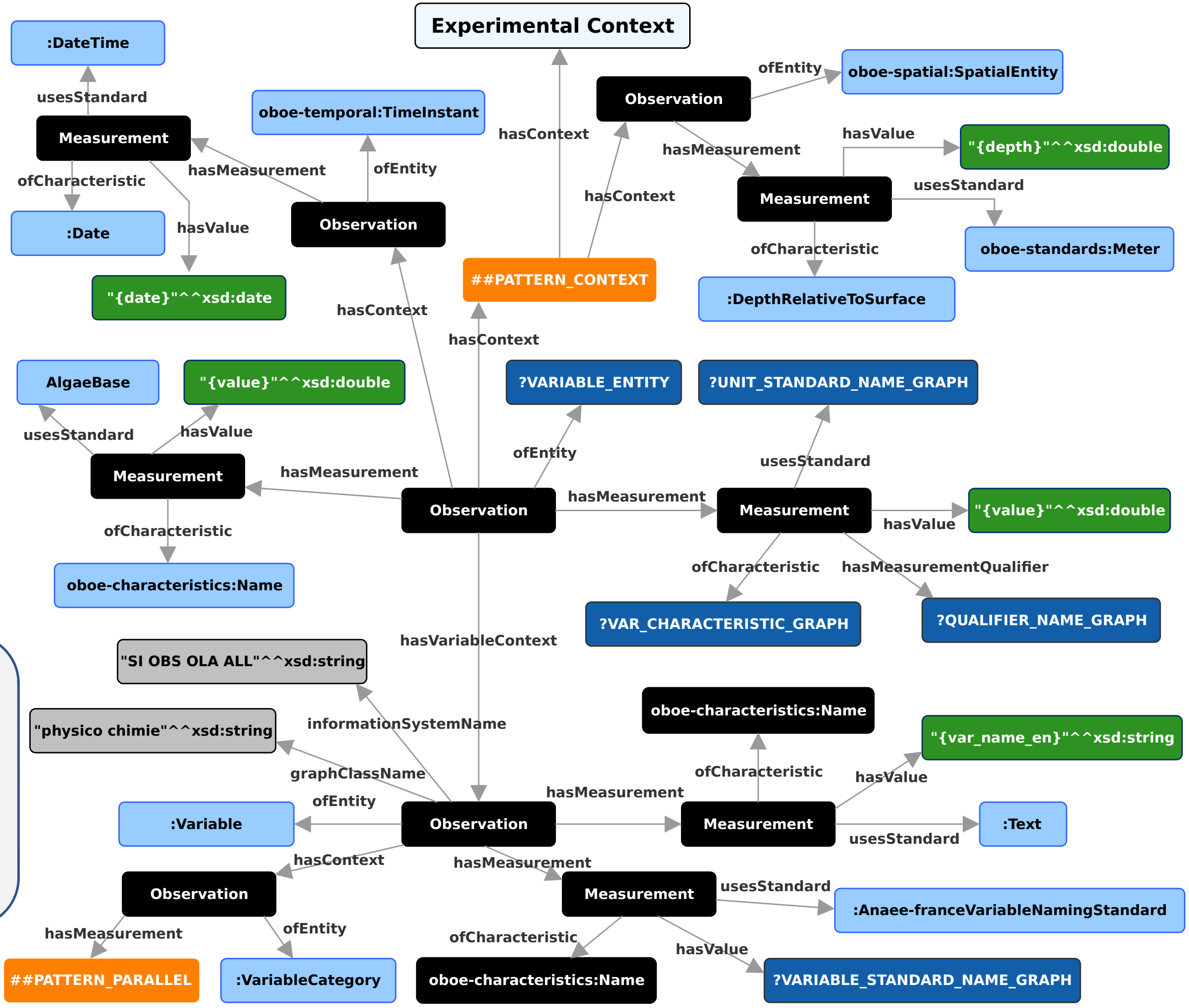
- naming pattern for dynamic uri (non-terminal nodes)

ex: http://anaee/ola/observation/water{**measure_id**}

+

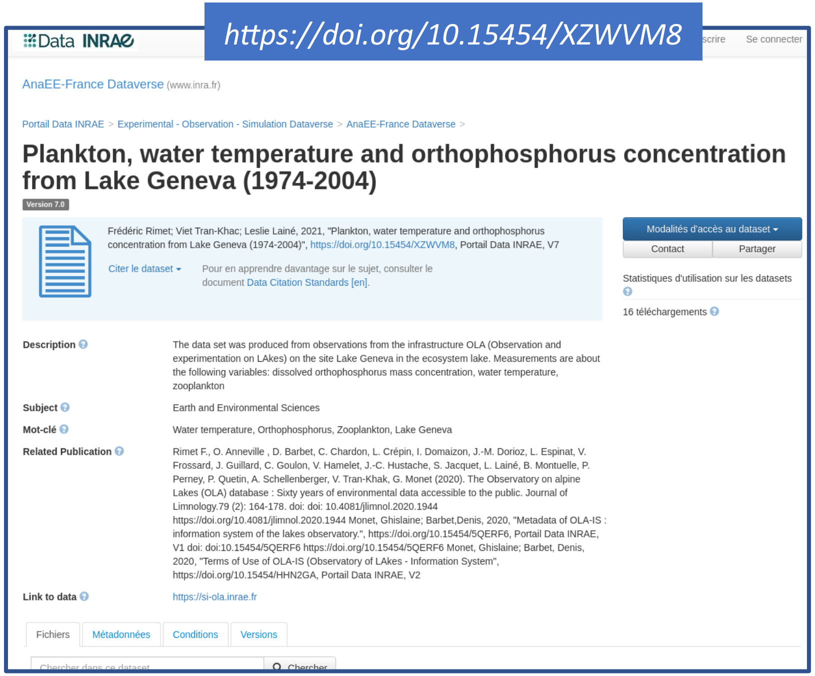
- corresponding SQL queries

ex: SELECT **measure_id**, value FROM measure



Application for planktonic biodiversity data from lakes

https://doi.org/10.15454/XZWVM8



AnaEE-France Dataverse (www.inra.fr)

Portail Data INRAE > Experimental - Observation - Simulation Dataverse > AnaEE-France Dataverse >

Plankton, water temperature and orthophosphorus concentration from Lake Geneva (1974-2004)

Version 7.0

Frédéric Rimet; Viet Tran-Khac; Leslie Lainé, 2021, "Plankton, water temperature and orthophosphorus concentration from Lake Geneva (1974-2004)", https://doi.org/10.15454/XZWVM8, Portail Data INRAE, V7

Citer le dataset ▾

Pour en apprendre davantage sur le sujet, consulter le document Data Citation Standards [en].

Statistiques d'utilisation sur les datasets

16 téléchargements

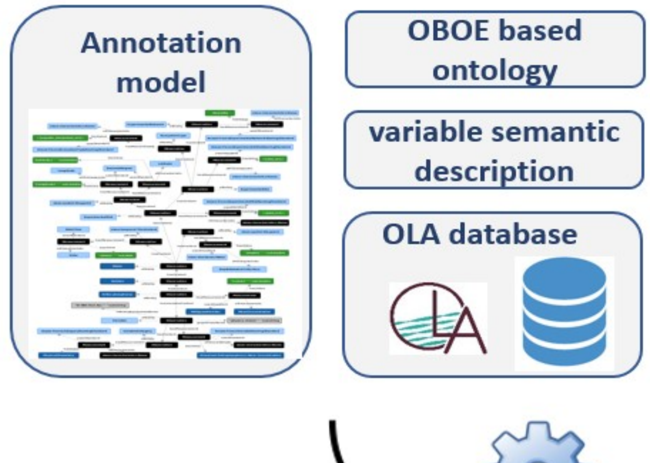
Description The data set was produced from observations from the infrastructure OLA (Observation and experimentation on LAkes) on the site Lake Geneva in the ecosystem lake. Measurements are about the following variables: dissolved orthophosphorus mass concentration, water temperature, zooplankton

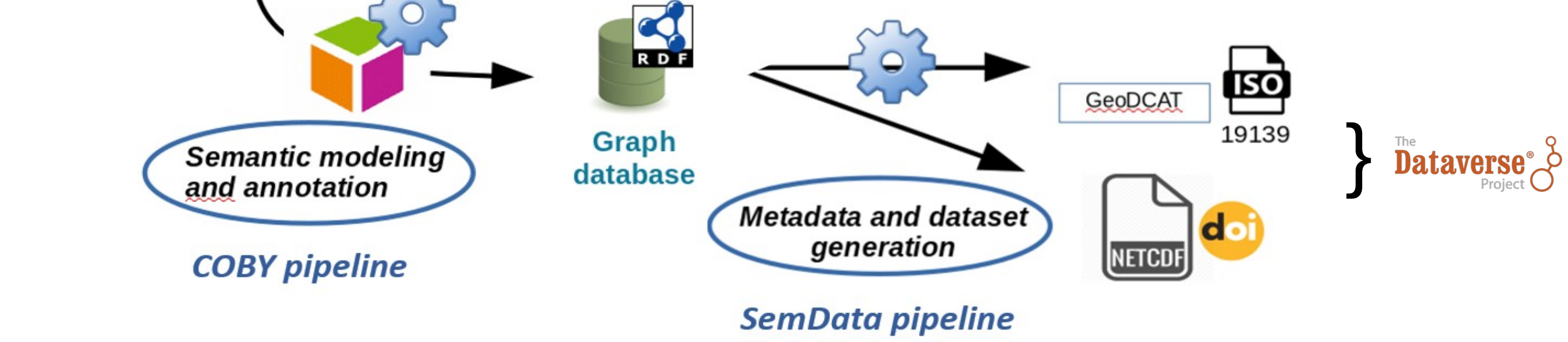
Subject Earth and Environmental Sciences

Mot-clé Water temperature, Orthophosphorus, Zooplankton, Lake Geneva

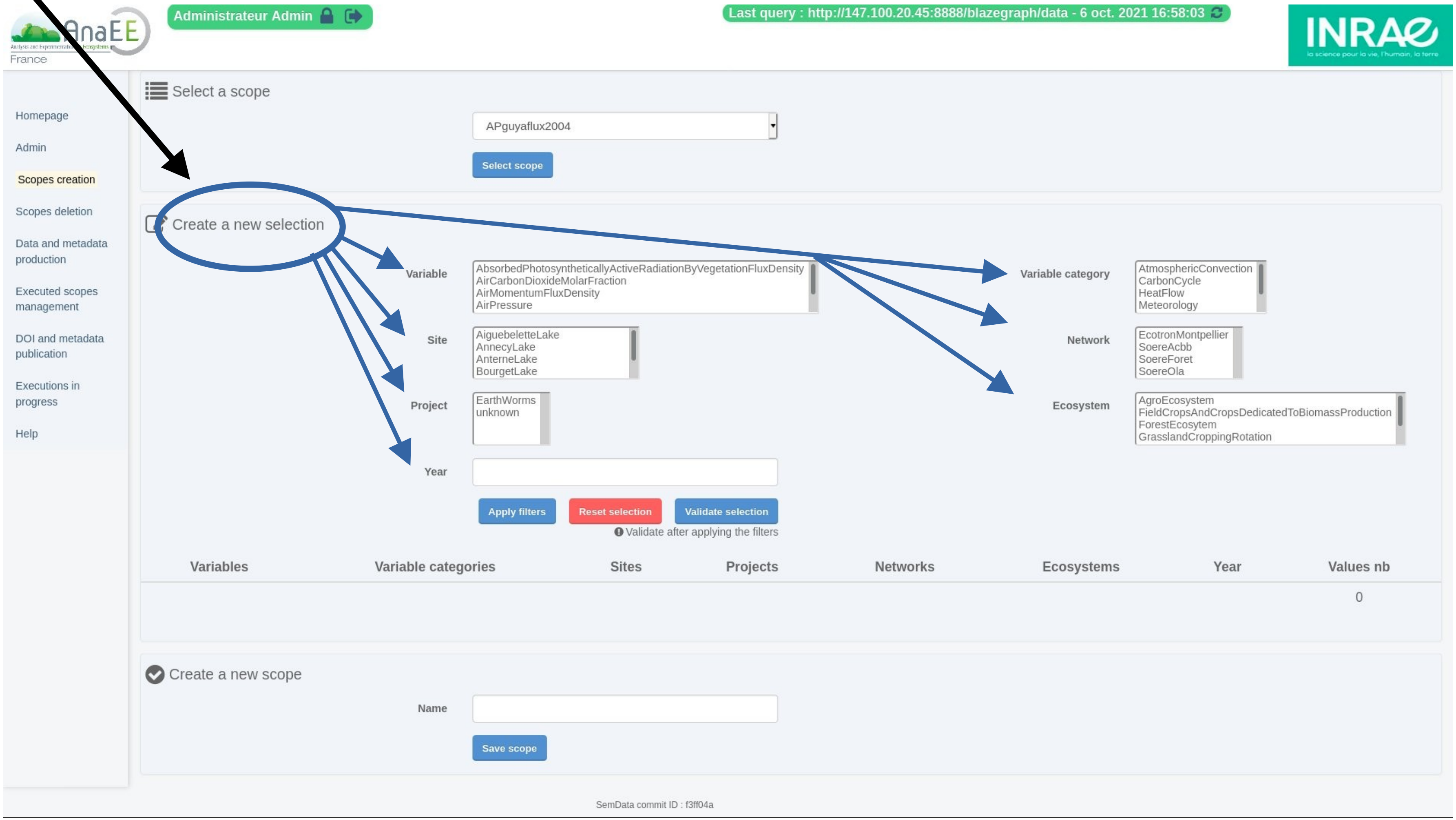
Related Publication Rimet F., O. Anneville , D. Barbet, C. Chardon, L. Crépin, I. Domaizon, J.-M. Dorioz, L. Espinat, V. Frossard, J. Guillard, C. Goulon, V. Hamelet, J.-C. Hustache, S. Jacquet, L. Lainé, B. Montuelle, P. Perney, P. Quetin, A. Schellenberger, V. Tran-Khak, G. Monet (2020). The Observatory on alpine Lakes (OLA) database : Sixty years of environmental data accessible to the public. Journal of Limnology.79 (2): 164-178. doi: doi: 10.4081/jlimnol.2020.1944 https://doi.org/10.4081/jlimnol.2020.1944 Monet, Ghislaine; Barbet,Denis, 2020, "Metadata of OLA-IS : information system of the lakes observatory.", https://doi.org/10.15454/5QERF6, Portail Data INRAE, V1 doi: doi:10.15454/5QERF6 https://doi.org/10.15454/5QERF6 Monet, Ghislaine; Barbet, Denis, 2020, "Terms of Use of OLA-IS (Observatory of LAkes - Information System", https://doi.org/10.15454/HHN2GA, Portail Data INRAE, V2

Link to data https://si-ola.inrae.fr





Metadata and dataset generation



AnaEE France

Administrateur Admin

Last query : http://147.100.20.45:8888/blazegraph/data - 6 oct. 2021 16:58:03

INRAE la science pour la vie, l'humain, la terre

- Homepage
- Admin
- Scopes creation
- Scopes deletion
- Data and metadata production
- Executed scopes management
- DOI and metadata publication
- Executions in progress
- Help

Select a scope

APguyaflux2004

Select scope

Create a new selection

Variable: AbsorbedPhotosyntheticallyActiveRadiationByVegetationFluxDensity
AirCarbonDioxideMolarFraction
AirMomentumFluxDensity
AirPressure

Variable category: AtmosphericConvection
CarbonCycle
HeatFlow
Meteorology

Site: AiguebeletteLake
AnnecyLake
AnterneLake
BourgetLake

Network: EcotronMontpellier
SoereAcbb
SoereForet
SoereOla

Project: EarthWorms
unknown

Ecosystem: AgroEcosystem
FieldCropsAndCropsDedicatedToBiomassProduction
ForestEcosystem
GrasslandCroppingRotation

Year:

Apply filters | Reset selection | Validate selection

Validate after applying the filters

Variables	Variable categories	Sites	Projects	Networks	Ecosystems	Year	Values nb
							0

Create a new scope

Name:

Save scope

SemData commit ID : f3ff04a

NetCDF header: Photyplankton biovolume

```
dimensions:
        Var0Dim0 = 2 ;      Var0Dim1 = 569 ;      Var0Dim2 = 425 ;
variables:
     string Var0Dim0(Var0Dim0) ;
            Var0Dim0:characteristic = "http://opendata.inra.fr/anaeeOnto#LowerDepthRelativeToSurface" ;
                [...]
     string Var0Dim1(Var0Dim1) ;
            Var0Dim1:characteristic = "http://opendata.inra.fr/anaeeOnto#Date" ;
                [...]
     string Var0Dim2(Var0Dim2) ;
            Var0Dim2:characteristic = "http://opendata.inra.fr/anaeeOnto#TaxonName" ;
            Var0Dim2:entity = "http://opendata.inra.fr/anaeeOnto#Phytoplankton" ;
            Var0Dim2:standard = "https://www.algaebase.org" ;
     double Var0(Var0Dim0, Var0Dim2, Var0Dim1) ;
            Var0:characteristic = "http://ecoinformatics.org/oboe/oboe.1.2/oboe-characteristics.owl#VolumePerVolume" ;
            Var0:entity = "http://opendata.inra.fr/anaeeOnto#Phytoplankton" ;
            Var0:standard = "http://opendata.inra.fr/anaeeOnto#MicrometerCubedPerMilliliter" ;
            Var0:name_of_experimental_network_in_Anaee-France_experimental_network_naming_standard=
http://opendata.inra.fr/anaeeOnto#OLAInfrastructure
            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#PhytoplanktonBiovolume            Var0:latitude_of_Waypoint_in_decimal_degree = "46.453457" ;
            Var0:longitude_of_Waypoint_in_decimal_degree = "6.5942335" ;
data:
 Var0Dim0 = "10.0", "18.0" ;
 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",  "1974-10-14 ».
```

No. of identified species

No. of dates

infos about species taxonomy

infos on the variable and linked contexts

Data section

NetCDF header: Photyplankton biovolume

infos on the variable and linked contexts

```
      double Var0(Var0Dim0, Var0Dim2, Var0Dim1) ;
             Var0:characteristic = "http://ecoinformatics.org/oboe/oboe.1.2/oboe-characteristics.owl#VolumePerVolume" ;
             Var0:entity = "http://opendata.inra.fr/anaeeOnto#Phytoplankton" ;
             Var0:standard = "http://opendata.inra.fr/anaeeOnto#MicrometerCubedPerMilliliter" ;
             Var0:name_of_experimental_network_in_Anaee-France_experimental_network_naming_standard=
http://opendata.inra.fr/anaeeOnto#OLAInfrastructure
             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#PhytoplanktonBiovolume            Var0:latitude_of_Waypoint_in_decimal_degree = "46.453457" ;
             Var0:longitude_of_Waypoint_in_decimal_degree = "6.5942335" ;

data:
 Var0Dim0 = "10.0", "18.0" ;
 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",  "1974-10-14 »,
 "1974-11-18", "1974-12-09", "1975-02-17", "1975-03-17",
 [...]
 Var0Dim2 = "Achnanthes catenata", "Achnanthes conspicua",  "Achnanthes exilis", "Achnanthes flexella", "Achnanthes
minutissima", "Achnanthes sp.", "Achroonema articulatum", "Actinastrum hantzschii", "Amphidinium sp.", "Amphipleura pellucida"
"Amphora ovalis",  "Amphora pediculus", "Amphora sp.
 [...]
Var0 =  NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, 399969, 222499,
328451, 603926, 111200, 31800, 74200, 0, 0, 10600, 0, NaN, 26500,
[...]
```

Data section

Semantics in ecology and ecosystem studies

Lessons from this work

- The OBOE generic ‘observation model’ allows for atomic modeling of the components of the system and of their nested or crossed relationships.

- In addition to the provided OBOE extensions (characteristics, spatial, temporal, standards), new classes and individuals are defined for the experimental modeling, especially for Entity (e.g experimental entities) and Standards (e.g lists of variable names or of experimental facilities)

- A graph pattern approach for the modeling of the variables leads to a more efficient investment at greatly reduced cost, allowing massive semantic processing of the data

- The generic pipelines developed can be re-used in other contexts and for other ontologies

- The whole process produces syntactically and semantically interoperable data, contributing to FAIR sharing and data reuse

Semantics in ecology and ecosystem studies

and some perspectives...

- In addition to the interoperability of data annotated with the same ontology (e.g OBOE), semantic interoperability between data annotated with different ontologies is needed.

  => alignment among semantics resources is of main importance

- As (most of) ontologies are domain specifics, the description of a broad perimeter has to rely on several ontologies

  => future enrichment of the description of experimentation on ecosystems using SSN, FOAF, PROV.. and use of existing controlled vocabularies, e.g for scientific name of taxon.

- The metadata generated by the workflow feeds trans-RI knowledge bases on the datasets and experimental sites

  => contribution the trans-domain linked data

- The NetCDF format is not well adapted to all types of data set

  => future generation of other “table type” formats

GEDEOP

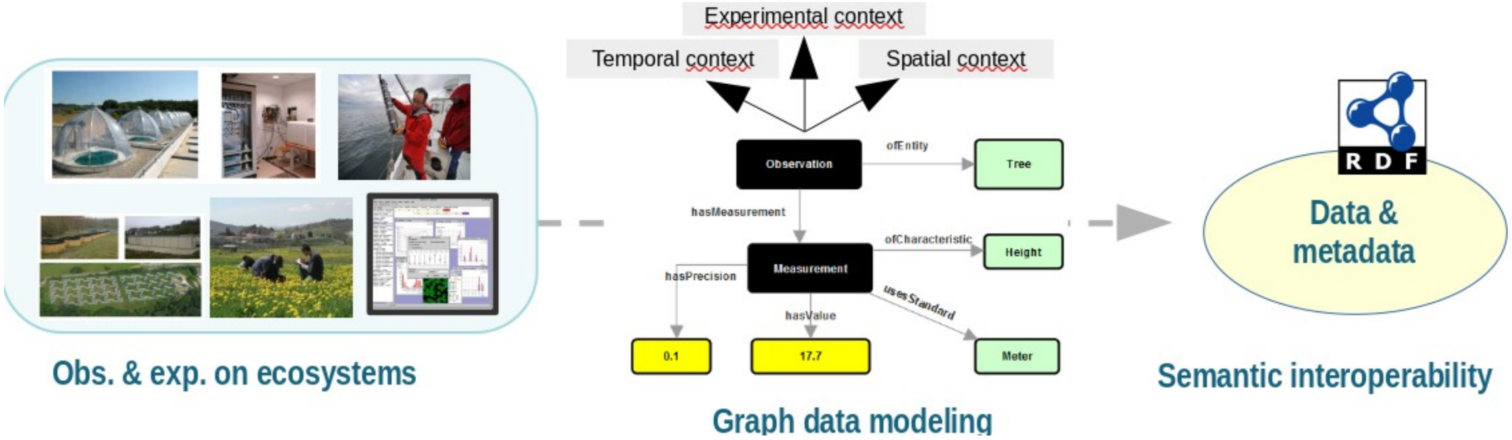
AnaEE
Analysis and Experimentation on Ecosystems

plus ENVRI

D2KAB

Thanks to all colleagues who contributed or are linked to this work

ENVRI FAIR



Obs. & exp. on ecosystems

Graph data modeling

Semantic interoperability