

Theia/OZCAR Thesaurus: lessons learned on implementing the I-ADOPT framework, a new Research Data Alliance recommendation designed to facilitate interoperability between scientific variables from different controlled vocabularies.

Charly Cousot (1), Isabelle Braud (2), Véronique Chaffard (3), Brice Boudevillain (4), Sylvie Galle(3)

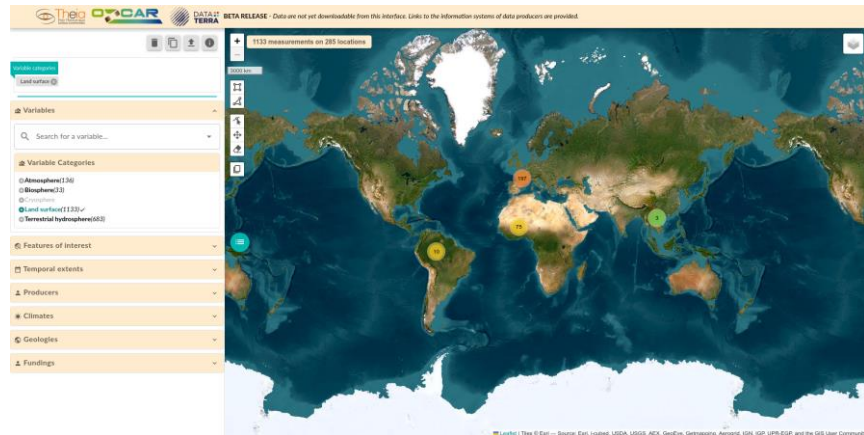
1. Université Grenoble Alpes, IRD, CNRS, Météo-France, INRAE, OSUG, 38000 Grenoble, France
2. INRAE, RiverLy, Villeurbanne, France
3. Université Grenoble Alpes, IRD, CNRS, Grenoble-INP, IGE, 38000 Grenoble, France
4. Université Grenoble Alpes, CNRS, IRD, Grenoble-INP, IGE, 38000 Grenoble, France

Data sharing in the context of OZCAR-RI and interdisciplinary science



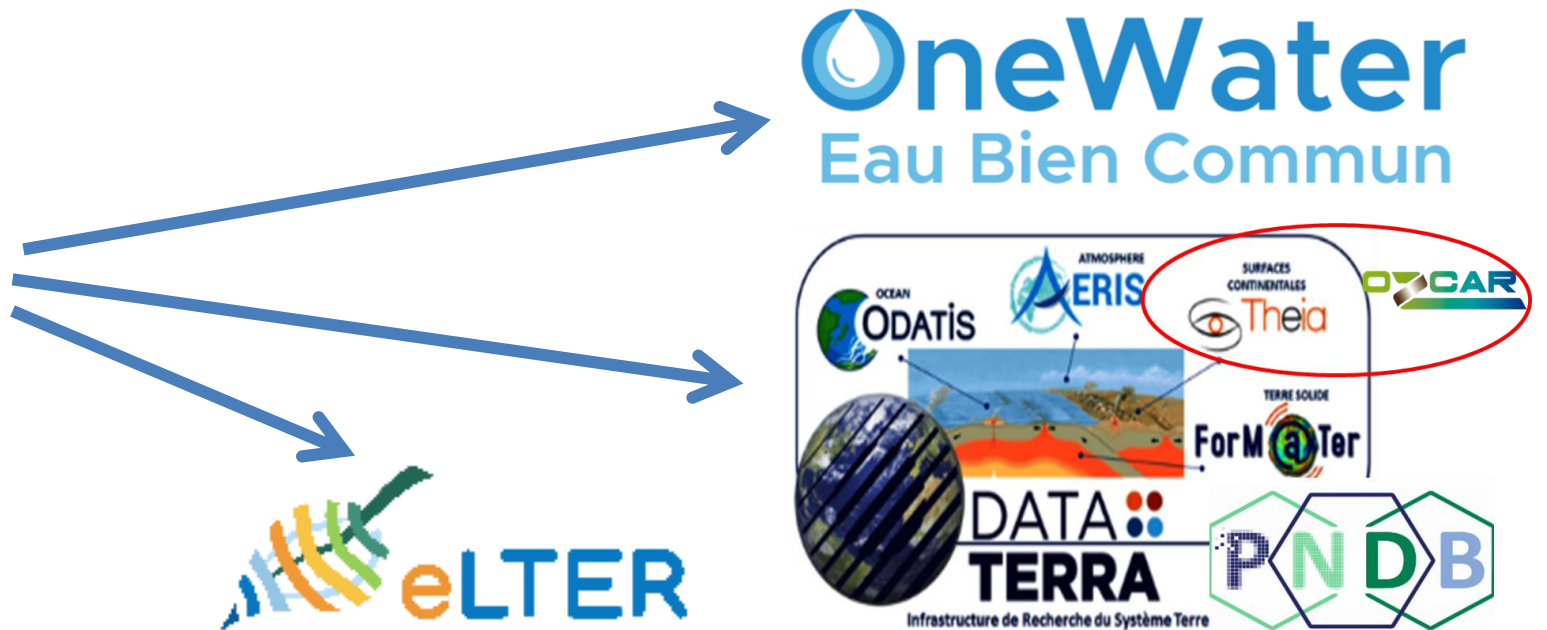
Data sources : 22 long term French critical zone observatories with a long history, with 22 pre-existing Information systems

Large heterogeneity in data (diversity of variables names), data management and practices



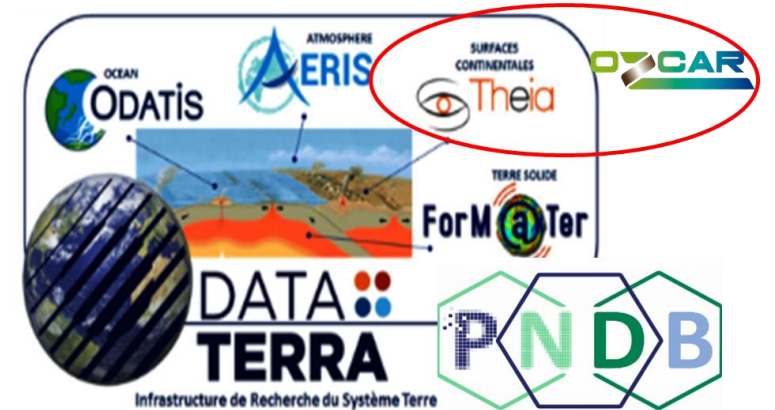
Theia/OZCAR information system

Data must be reused in a context of interdisciplinary research



OneWater
Eau Bien Commun

eLTER



A controlled vocabulary to support the objectives of the Theia/OZCAR IS

Objectives of Theia/OZCAR IS:

- Expose data from heterogeneous sources on a [single data portal](#) with homogenous metadata
- The search on **variable names** is a common need of the critical zone science community and more generally of the Earth Science community

Solutions provided by the creation of a controlled vocabulary of environmental variables : the [Theia/OZCAR thesaurus](#)

- ⇒ To facilitate the data discovery by providing research criteria using **harmonized variable names**
- ⇒ To enhance data reusability and interoperability by providing **rich information on variable** allowing for unambiguous interpretation of data by a wider user community than the one that produced it and also by machines

The screenshot shows the 'Theia/OZCAR thesaurus' interface. It features a search bar at the top right with a language dropdown set to 'English'. Below the search bar are three tabs: 'Alphabetical', 'Hierarchy', and 'Groups'. Under 'Alphabetical', there are two rows of letter-based navigation links: 'A B C D E F G H I K L M N O P' and 'R S T U V W Y Z 0-9'. A scrollable list of variables is displayed, including terms like 'Aboveground', 'Aboveground dry vegetation biomass', 'Absolute humidity', 'Absorbance', 'Accumulation', 'Acetochlor', 'Acoustic investigation variable', 'Air absolute humidity', 'Air constraint', 'Air pressure', 'Air pressure at 20 meters height', 'Air pressure at 5 meters height', 'Air relative humidity', 'Air relative humidity at 15 meters height', 'Air relative humidity at 18 meters height', 'Air relative humidity at 2 meters height', 'Air relative humidity at 20 meters height', 'Air relative humidity at 5 meters height', 'Air specific humidity', 'Air temperature', and 'Air temperature at 15 meters height'. To the right of the list is a 'Vocabulary information' section with the following details:

- TITLE:** Theia/OZCAR thesaurus
- DESCRIPTION:** Thesaurus for in situ data from Environmental and Critical Zone Sciences. Used by Theia/OZCAR information system : <https://in-situ.theia-land.fr/>
- CREATOR:** Charly Coussot <https://orcid.org/0000-0002-0544-4802>, Véronique Chaffard <https://orcid.org/0000-0003-2823-7117>, Isabelle Braud <https://orcid.org/0000-0001-9155-0056>, Sylvie Galle <https://orcid.org/0000-0002-3100-8510>
- LICENSE:** <http://creativecommons.org/licenses/by/4.0/>
- LANGUAGE:** <http://lexvo.org/id/iso639-3/eng>
- SOURCE:** GCMD Science Keywords: <https://earthdata.nasa.gov/about/gcmd/global-change-master-directory-gcmd-keywords>
- CREATED:** Monday, January 1, 2018 00:00:00
- LAST MODIFIED:** Friday, September 15, 2023 13:27:27
- DC-REQUIRES:** <http://purl.org/voc/cpm>, <https://w3id.org/iadopt/ont>
- TYPE:** <http://www.w3.org/2004/02/skos/core#ConceptScheme>
- URI:** <https://w3id.org/ozcar-theia>

At the bottom of the information section, there is a link: 'Download this vocabulary: RDF/XML TURTLE'.

Examples of expected details:

- "Precipitation amount": integration over which time step? is it solid or liquid precipitation (snow)?
- "Water level": to which object does this variable refer? surface or groundwater?

Why do we need a framework for naming variables ?

Problem !!

The more precise the variable concept, the more difficult it is to find similarity relationships with concepts in other vocabularies.

 **Interoperability** 

Why do we need a framework for naming variables ?

Problem !!

The more precise the variable concept, the more difficult it is to find similarity relationships with concepts in other vocabularies.

→ **Interoperability** 

Solution

Decompose the complex variable concept into simpler atomic concepts and define similarity relationships between these simpler concepts and concepts from other vocabularies.

→ **Interoperability** 

The I-ADOPT framework ontology

Decomposition of variables names into atomic concepts :

- Property
- Entity: roles [ObjectOfInterest, ContextObject, Matrix]
- Constraint (depth, temperature, wavelength, ...)

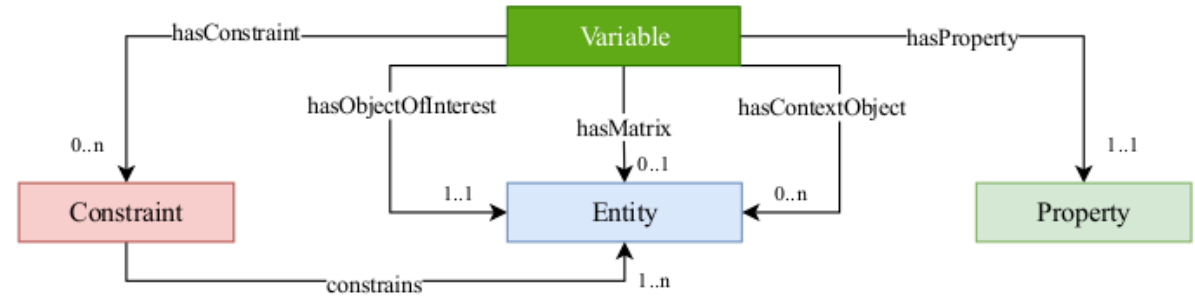
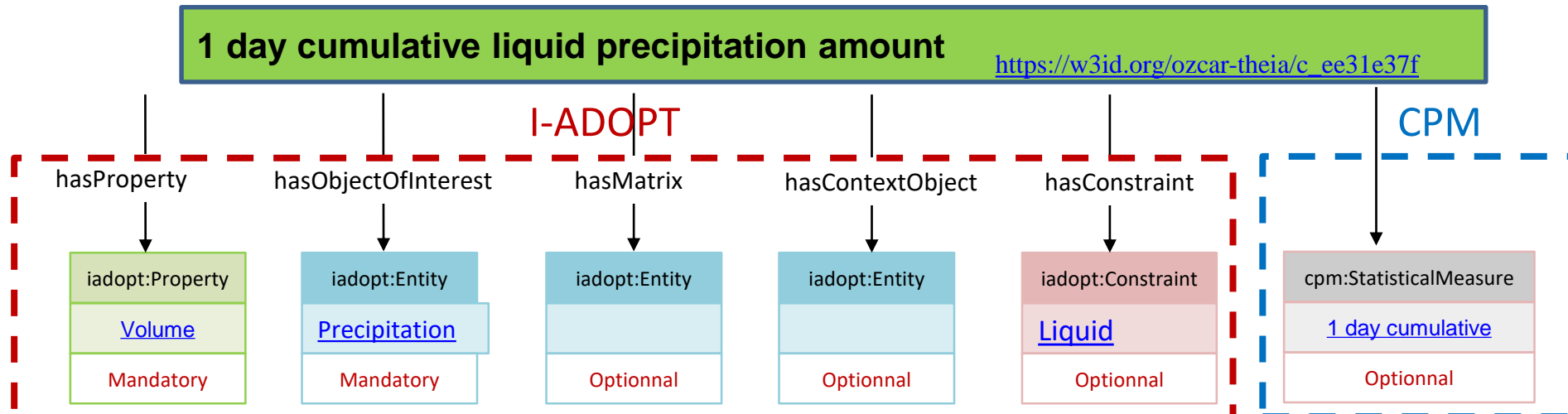


Figure 2: The I-ADOPT Framework.

Magnana et al., S4BioDiv2021, 2021



- Implementation facilitated by the [I-ADOPT patterns](#) provided on quantitative/qualitative variables
- Need to complement with the CPM (Complex Property Model) ontology for the notion of **temporal and spatial aggregation**



What worked well

- We manage to get grips on the framework with available online resources
- Generic enough to model all of our variables (atmosphere, biosphere, cryosphere, continental surface, hydrosphere, chemistry, geophysics ...)
- Enriched our vocabulary with new concepts that could be used to create newer dimension for data discovery : process, phenomenon, chemical entity, environmental entity (lake, river, ...) ...
- Combination with SKOS hierarchical relations to provide categorisation and to enhance data discovery
- Combination with CPM ontology to describe statistical aggregation
- Satisfy our semantic interoperability needs

... > Land surface variable > Soil variable > Soil geophysical variable > Acoustic investigation variable > Soil reflected acoustic wave median amplitude in all directions inside borehole

PREFERRED TERM	Soil reflected acoustic wave median amplitude in all directions inside borehole 
TYPE	Variable
BROADER CONCEPT	Acoustic investigation variable
STATISTICAL MEASURE	360° median
HASCONSTRAINT	Reflected wave
HASCONTEXTOBJECT	Borehole
HASMATRIX	Soil
HASOBJECTOFINTEREST	Acoustic wave
HASPROPERTY	Amplitude
SIMPLIFIED LABEL	Soil reflected acoustic wave amplitude
URI	https://w3id.org/ozcar-theia/c_1731d463 
DOWNLOAD THIS CONCEPT:	RDF/XML TURTLE JSON-LD Created 12/20/22, last modified 12/20/22

Limitations of I-ADOPT framework ontology

- Variable label often too complicated to be used directly for data discovery. We implemented our own “Simplified Label” for use on the web portal

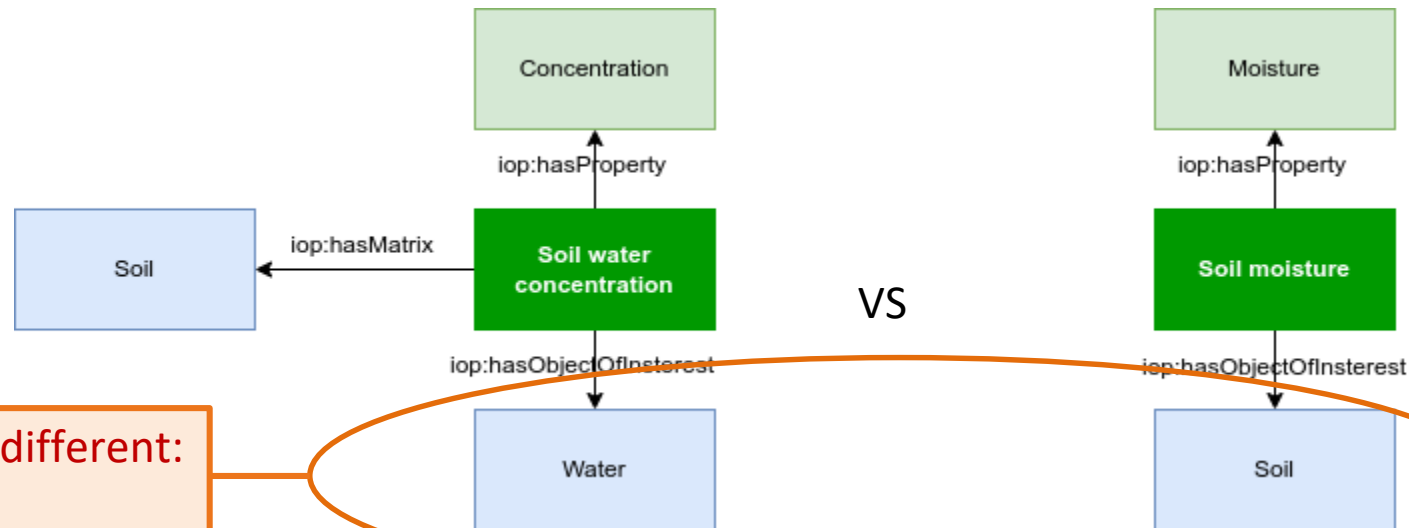
PREFERRED TERM

1 day cumulative liquid precipitation amount 

SIMPLIFIED LABEL

Precipitation amount

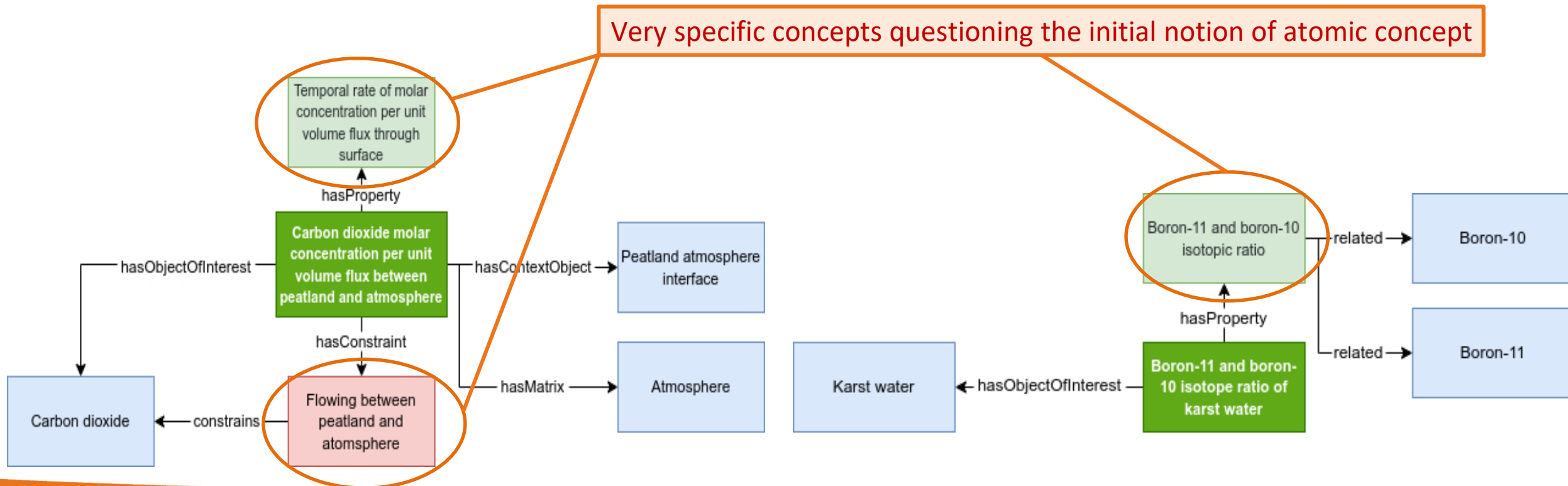
- For some variables, different implementations are possible. How to choose one or another ? How can we infer similarity relation between identical variables modelled differently ?



The object of interest is different:
water versus soil

Limitations of I-ADOPT framework ontology

- Difficulty for modelling some variables such as a flux between two compartments (ex: carbon dioxide flux between the land surface and the atmosphere).
- Some concepts need to be very specific to describe the variable. Which can cause the loose of atomicity notion of I-ADOPT concepts.



Benefits of this work



- ✓ Allows an unambiguous interpretation of data, thus a better reuse
- ✓ Promotes alignments between international thesauri in the field
- ✓ Promotes better semantic interoperability of data at the national/international level

-> interdisciplinary studies requiring cross-referencing of multi-source and multi-theme data

Enables to add precise information about the variables measured by a dataset, using the « keyword » fields of metadata standards that do not include a description of the variables measured.

The screenshot shows the eLTER data discovery interface. At the top, there are navigation links: "Discover eLTER Data", "Visualization Time Series Data", and "Discovery Service". Below this is a search bar with a "Search" button and a "Map" button. A "Back to search" button is also visible. The main content area displays search results for "Time series of type chemistry in Le Lez (Méditerranée) basin - MEDYCYSS observatory - KARST observatory network - OZCAR Critical Zone network Research Infrastructure". The results include "GEMET - INSPIRE themes, version 1.0" with a link to "Environmental monitoring facilities". A dashed green box highlights the "OZCAR-Theia variables thesaurus" section, which lists several variables with search icons: "Dissolved aluminium mass concentration per unit volume in karst water", "Dissolved arsenic mass concentration per unit volume in karst water", "Dissolved barium mass concentration per unit volume in karst water", and "Dissolved hydrogenocarbonate mass concentration per unit volume in karst water".

To learn more about the project:

Braud, I., Chaffard, V., Coussot, C., Galle, S., et al., 2020. Building the Information System of the French Critical Zone Observatories network: Theia/OZCAR-IS, Hydrological Sciences Journal, special issue "Data: opportunities and barriers",

<https://doi.org/10.1080/02626667.2020.1764568> .

Coussot et al., Putting in practice the I-ADOPT framework for the naming of environmental variables from continental surfaces, in preparation

To access the portal, the thesaurus and the project Github

Data portal : <https://in-situ.theia-land.fr/>

Thesaurus: <https://w3id.org/ozcar-theia/>

Cataloguing CSW webservice: <https://in-situ.theia-land.fr/geonetwork/srv/eng/csw?service=CSW&version=2.0.2&request=GetCapabilities>

GitHub : <https://github.com/theia-ozcar-is>

Contacts:

charly.coussot@univ-grenoble-alpes.fr

Veronique.Chaffard@univ-grenoble-alpes.fr

Isabelle.braud@inrae.fr

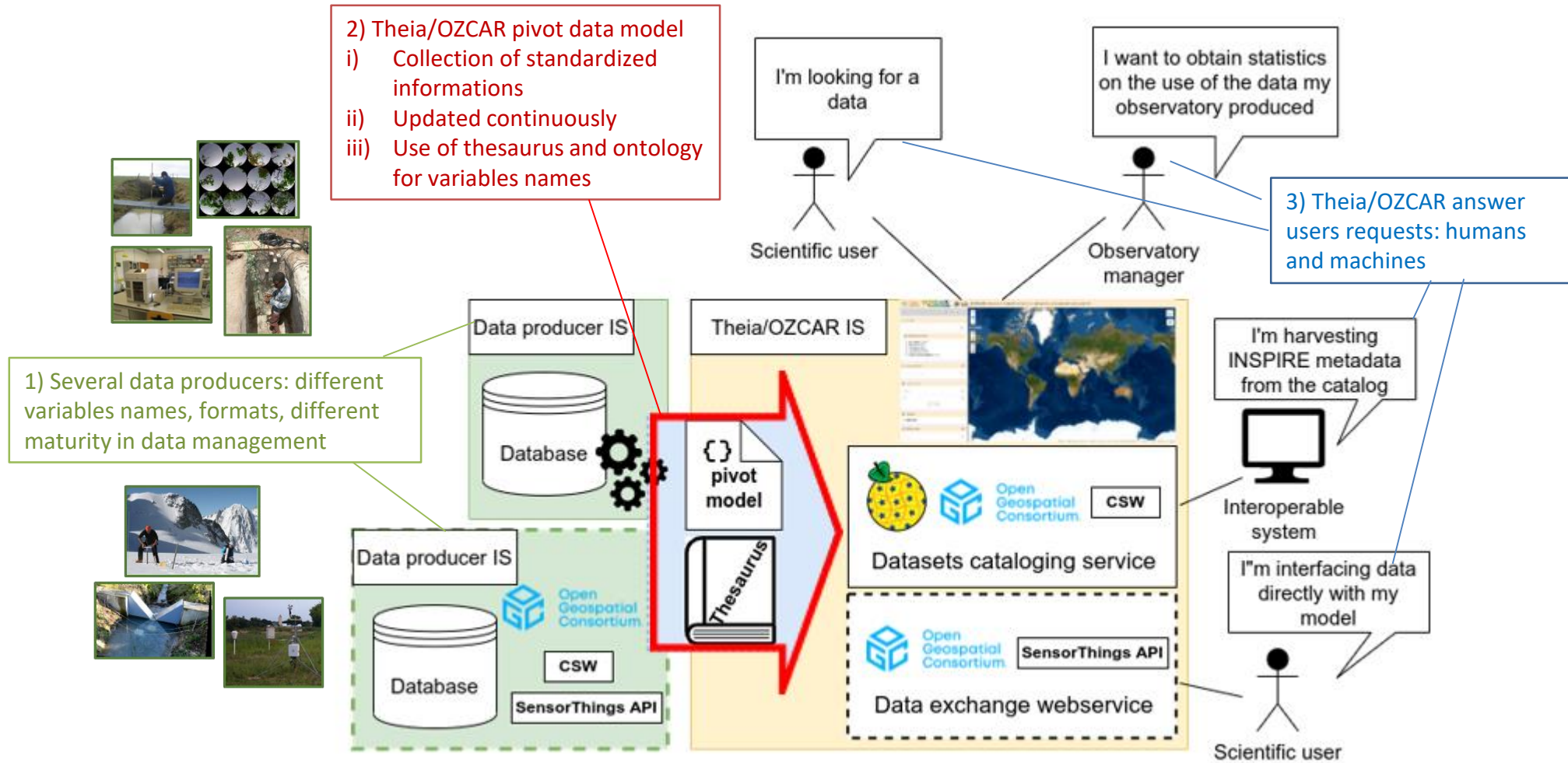
Sylvie.galle@ird.fr

Acknowledgements: The project contributes to the Theia, continental surface data pole and the Data Terra Research Infrastructure. It receives supports from the OZCAR Research Infrastructure, IRD, CNRS and was partly funded by the FAIRTois ANR Project (grant ANR-19-DATA).

Thank you for your attention:
Questions ?

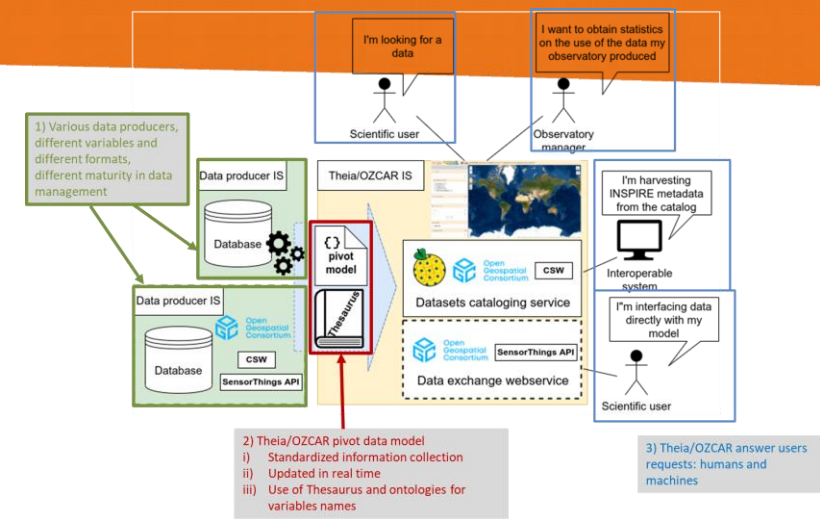


Data fluxes and services between data producers and users

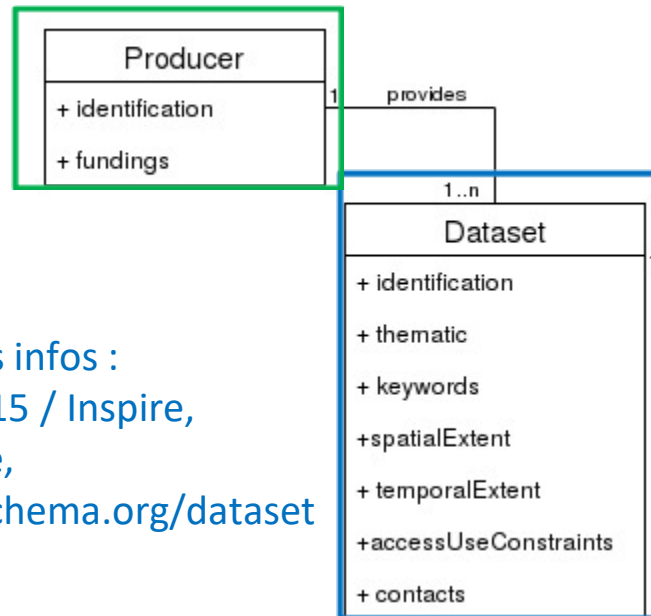


Pivot data model

- **Pivot data model** to harmonize data description, get the required information for the faceted search and set up data exchange web services
- Based on the mapping of different standards

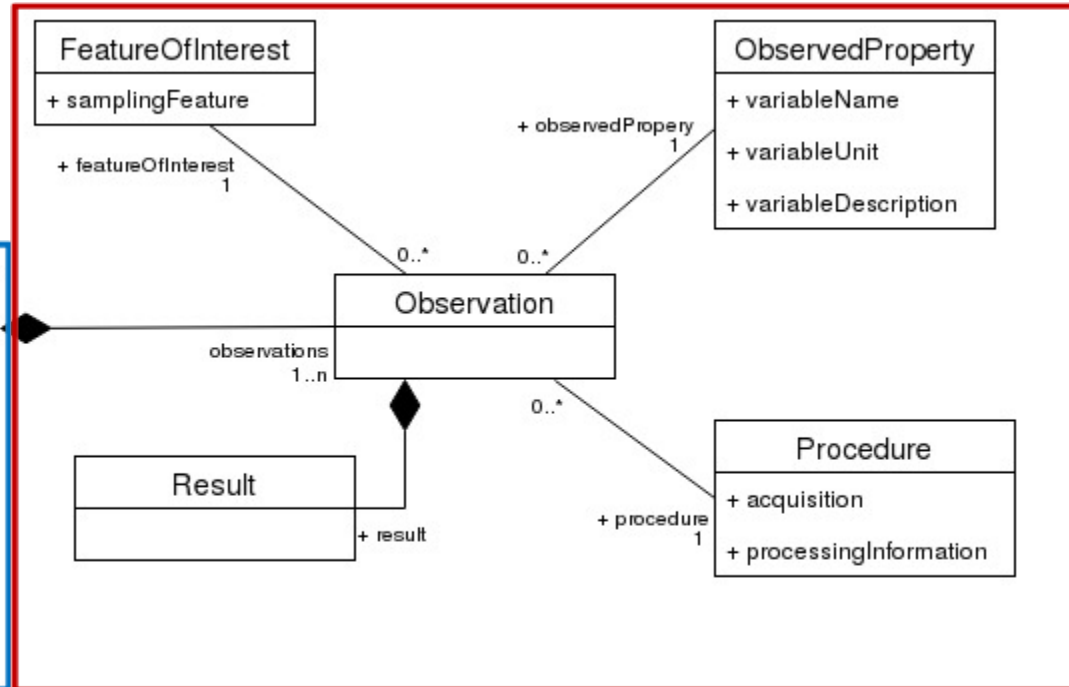


Producers info (DataCite, scanR API)



Datasets infos :
 ISO 19115 / Inspire,
 DataCite,
 DCAT, schema.org/dataset

Observations infos: O&M



<https://github.com/theia-ozcar-is/data-model-documentation>

A FAIR thesaurus FAIR : <https://w3id.org/ozcar-theia>

DOI : 10.17178/67b5a1d5-8c8c-4a94-a646-1cca1d0adf79

Variables

Objects of interest

OZCAR Theia in-situ thesaurus

OZCAR Theia in-situ thesaurus

Theia/OZCAR thesaurus English

Alphabetical | Hierarchy | Groups

- Constraint
- Instrument
- Method
- Observable property
- Phenomenon
- Physical entity
- Process
- Time
- Variable

Vocabulary information

TITLE	Theia/OZCAR thesaurus
DESCRIPTION	Thesaurus for in situ data from Environmental and Critical Zone Sciences. Used by Theia/OZCAR information system : https://in-situ.theia-land.fr/
CREATOR	Charly Coussot https://orcid.org/0000-0002-0544-4802 Véronique Chaffard https://orcid.org/0000-0003-2823-7117 Isabelle Braud https://orcid.org/0000-0001-9155-0056 Sylvie Galle https://orcid.org/0000-0002-3100-8510
LICENSE	http://creativecommons.org/licenses/by/4.0/
LANGUAGE	http://lexvo.org/id/iso639-3/eng
SOURCE	GCMD Science Keywords: https://earthdata.nasa.gov/about/gcmd/global-change-master-directory-gcmd-keywords
CREATED	Monday, January 1, 2018 00:00:00
LAST MODIFIED	Friday, July 1, 2022 13:45:37
DC:REQUIRES	http://purl.org/voc/cpm https://w3id.org/iadopt/ont
TYPE	http://www.w3.org/2004/02/skos/core#ConceptScheme

Alphabetical | Hierarchy | Groups

- Variable
 - Atmosphere variable
 - Biosphere variable
 - Cryosphere variable
 - Land surface variable
 - Terrestrial hydrosphere variable
 - Groundwater hydrology
 - Karst hydrology
 - Surface water hydrology
 - Surface water chemistry
 - Surface water microbiology
 - Surface water physic variable
 - Pond turbidity
 - River discharge
 - Surface water conductivity
 - Surface water pH
 - Surface water suspended sediment concentration
 - Surface water temperature
 - Water level

Alphabetical | Hierarchy | Groups

- Physical entity
 - Chemical entity
 - Environmental entity
 - Atmosphere
 - Biosphere
 - Cryosphere
 - Hydrosphere
 - Groundwater
 - Karst water
 - Water table
 - Surface water
 - Lake
 - Pond
 - River
 - Spring
 - Water
 - Cloud
 - Dew
 - Raindrop
 - Land surface
 - Grain
 - Rock
 - Sediment
 - Soil
 - Topography

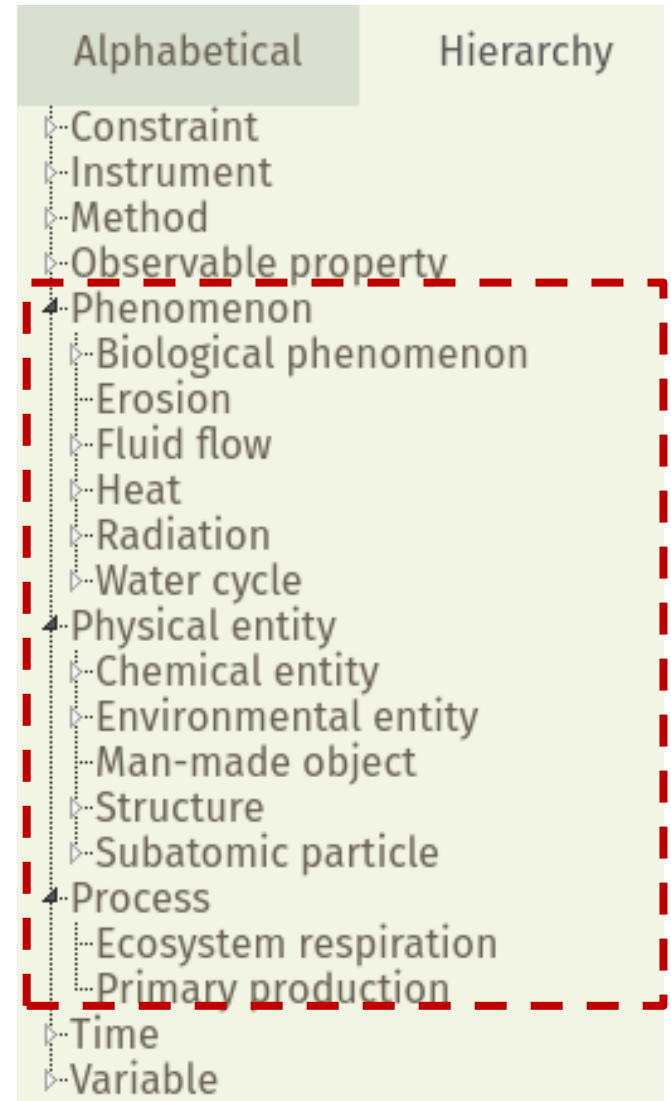
Benefits

Description of variables with rich and formal semantics (ad-hoc ontology)

- ✓ Enriched our thesaurus with new concepts

ObjectOfInterest : process, phenomenon, chemical entity,
environmental entity (lake, river, ...), ...

-> which would allow us to offer new search dimensions on our portal (in addition to the observed variable)



Benefits

- ✓ Promotes unambiguous interpretation of data and therefore better reuse

Variable

PREFERRED TERM	1 day cumulative liquid precipitation amount
TYPE	Variable
BROADER CONCEPT	Precipitation amount
STATISTICAL MEASURE	1 day cumulative
HASCONSTRAINT	Liquid
HASOBJECTOFINTEREST	Precipitation
HASPROPERTY	Volume
SIMPLIFIED LABEL	Precipitation amount
URI	https://w3id.org/ozcar-theia/c_ee31e37f
DOWNLOAD THIS CONCEPT:	RDF/XML TURTLE JSON-LD Created 2/7/22, last modified

ObjetOfInterest

PREFERRED TERM	Precipitation
TYPE	Entity
DEFINITION	[Wikipedia] In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravitational pull from clouds. The main forms of precipitation include drizzling, rain, sleet, snow, ice pellets, graupel and hail. Precipitation occurs when a portion of the atmosphere becomes saturated with water vapor (reaching 100% relative humidity), so that the water condenses and "precipitates" or falls.
BROADER CONCEPT	Water cycle
URI	https://w3id.org/ozcar-theia/c_2b48133e
DOWNLOAD THIS CONCEPT:	RDF/XML TURTLE JSON-LD Created 2/7/22, last modified 7/1/22

