Theia/OZCAR Thesaurus: a terminology service to facilitate the discovery, interoperability and reuse of data from continental surfaces and critical zone science in interdisciplinary research

Isabelle Braud (1), Charly Coussot (2,3), Véronique Chaffard (3), Sylvie Galle (3), Rémi Cailletaud (2)

1) INRAE Riverly , Lyon, France
2) OSUG - Observatoire des Sciences de l'Univers, Grenoble, France
3) IGE - Institut des Géosciences de l'environnement, Grenoble, France

Session GI2.3: Data and Information Services for Interdisciplinary Research and Applications in Earth Science

# Data sharing in the context of critical zone and interdisciplinary science

Data sharing context

- Data sources : 21 long term French critical zone observatories with a long history, with 21 pre-existing Information systems

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

- Context of interdisciplinary research with the building of a French Earth Science data Research Infrastructure in atmospheric, oceanic, solid earth, continental surfaces and biodiversity science

- Contribution to a European Research Infrastructure eLTER RI

Objectives:

- Expose data from heterogeneous sources on a single data portal

- The search on variables names is a common need of the critical zone science community and more generally of the Earth Science community

$\Rightarrow$ Creation of a controlled vocabulary of environmental variables : the Theia/OZCAR thesaurus

$\Rightarrow$ To facilitate the data discovery by providing research criteria using harmonized variables names

$\Rightarrow$ To enhance data reusability and interoperability by providing rich information on variables

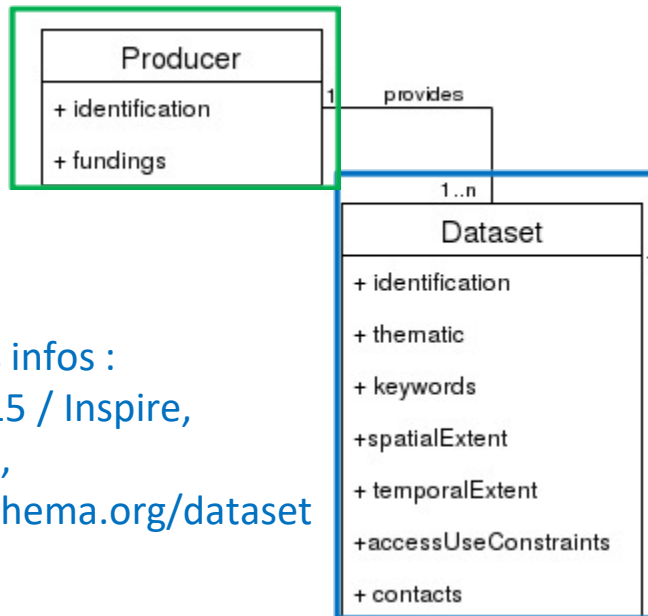# Data fluxes and services between data producers and users
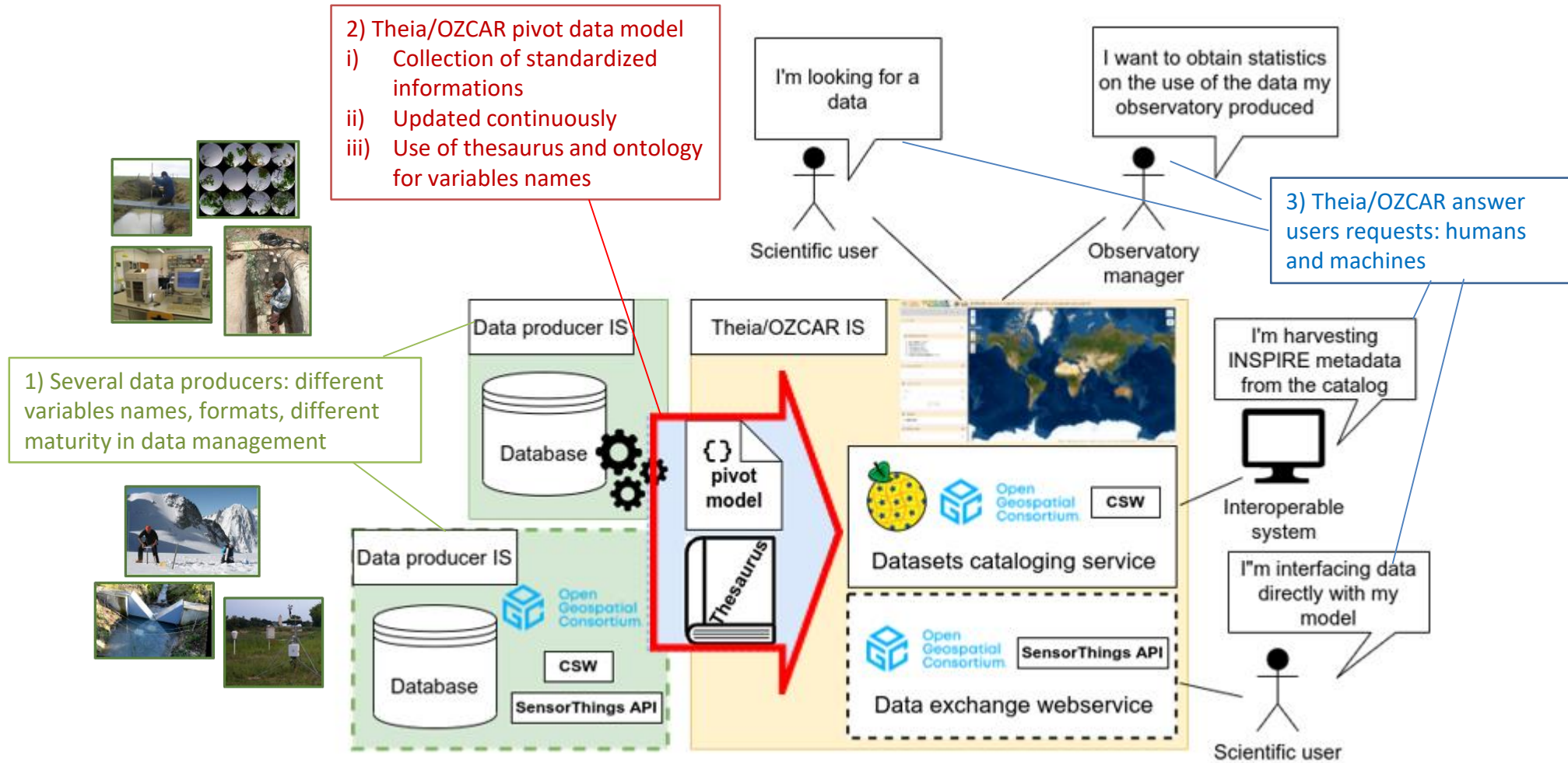
Pivot data model

- **Pivot data model** to harmonize data description, get the required information for the facetted search and set up data exchange web services

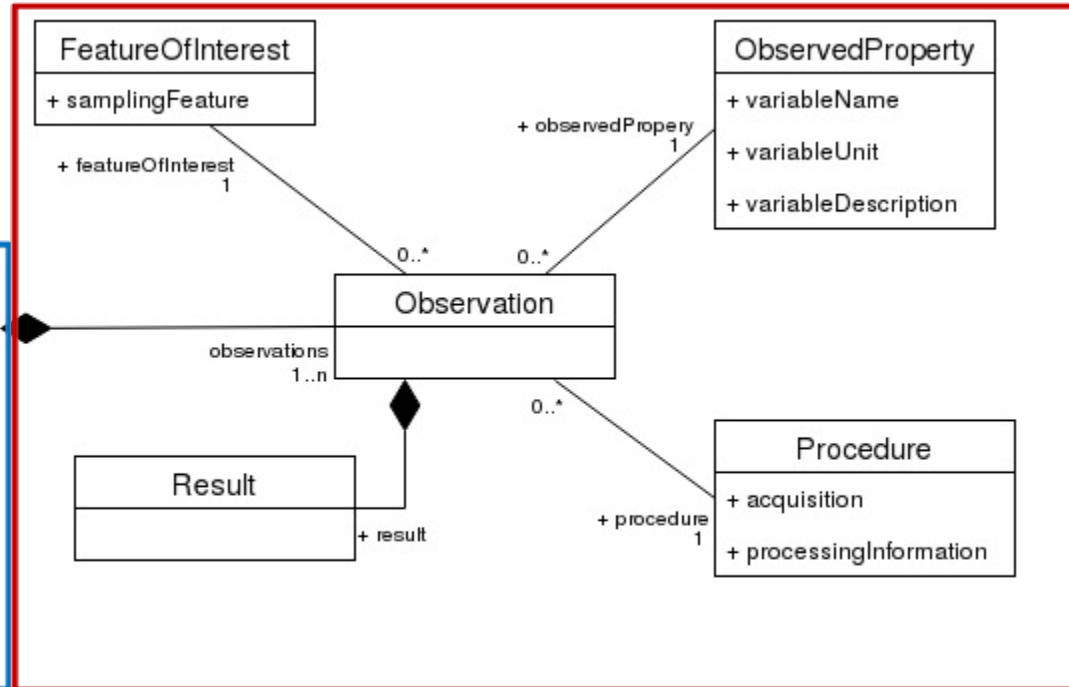- Based on the mapping of different standards

Observations infos: O&M
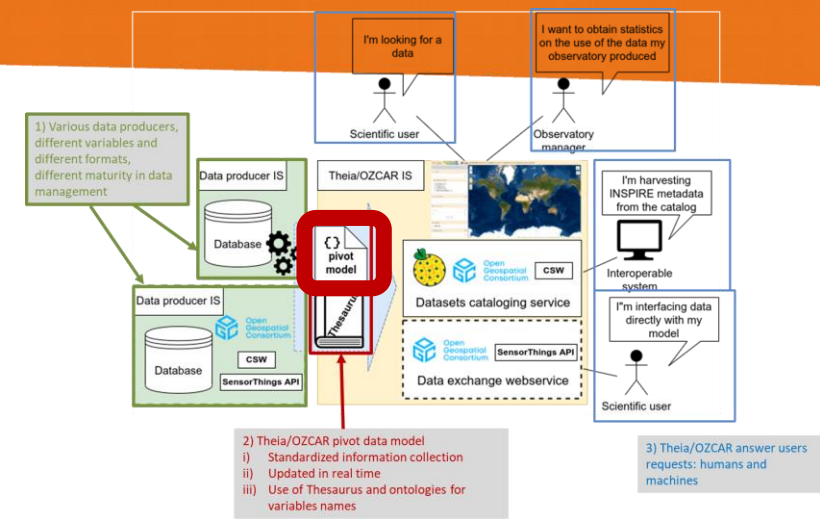
Producers info (DataCite, scanR API)

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

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

Why did we choose to use I-ADOPT ontology in Theia/OZCAR thesaurus?

Needs expressed by the critical zone science community

- Simplified variable names for data discovery (e.g. "precipitation amount")

- Description of observed variables names with rich semantics for unambiguous interpretation of data by a wider user community than the one that produced it and also by machines

- 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 water or groundwater?

Ontologies used to formalize the thesaurus

- SKOS ontology = light ontology. Hierarchical or similarity relations between semantically non-typed terms (skos:Concept). Not precise enough when alignment with other concepts is required

- Comparison of various ontologies for variables names: Scientific Variable Ontology (SVO), Complex Property Model (CPM), Extensible Observation Ontology (OBOE), Interoperable Descriptions of Observable Property Terminology (I-ADOPT)

- **Choice of I-ADOPT** (RDA recommendation I-ADOPT)

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

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

Variables

Objects of interest

Theia/OZCAR thesaurus

English ▾ | Search

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

OZCAR Theia in-situ thesaurus

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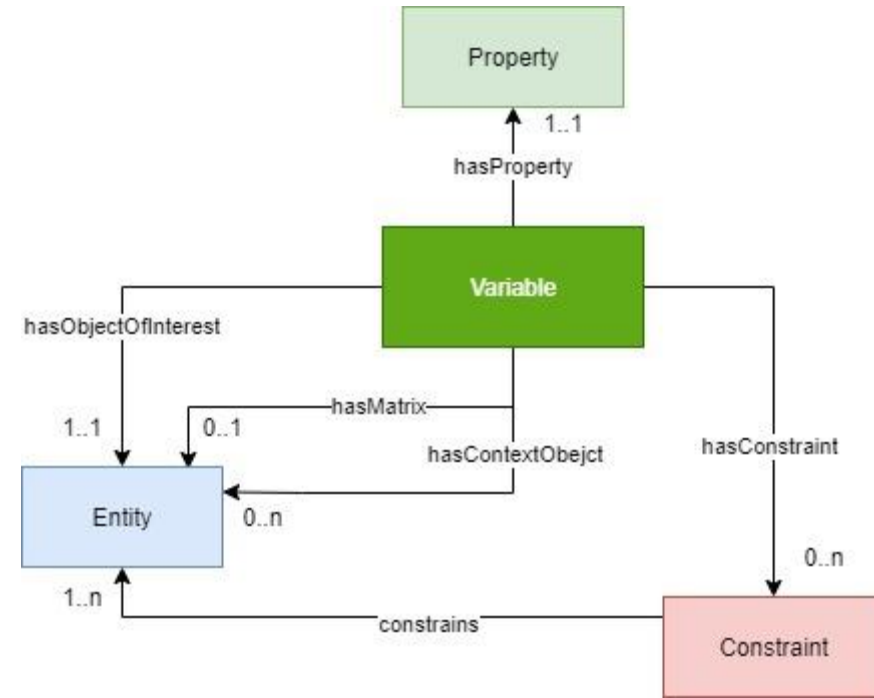
- **Variable**
  - Atmosphere variable
  - Biosphere variable
  - Cryosphere variable
  - Land surface variable
  - Terrestrial hydrosphere variable
    - Groundwater hydrology
    - Karst hydrology
    - Surface water hydrology
      - Surface water chemistry
      - Surface water microbiology
    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

OZCAR Theia in-situ thesaurus

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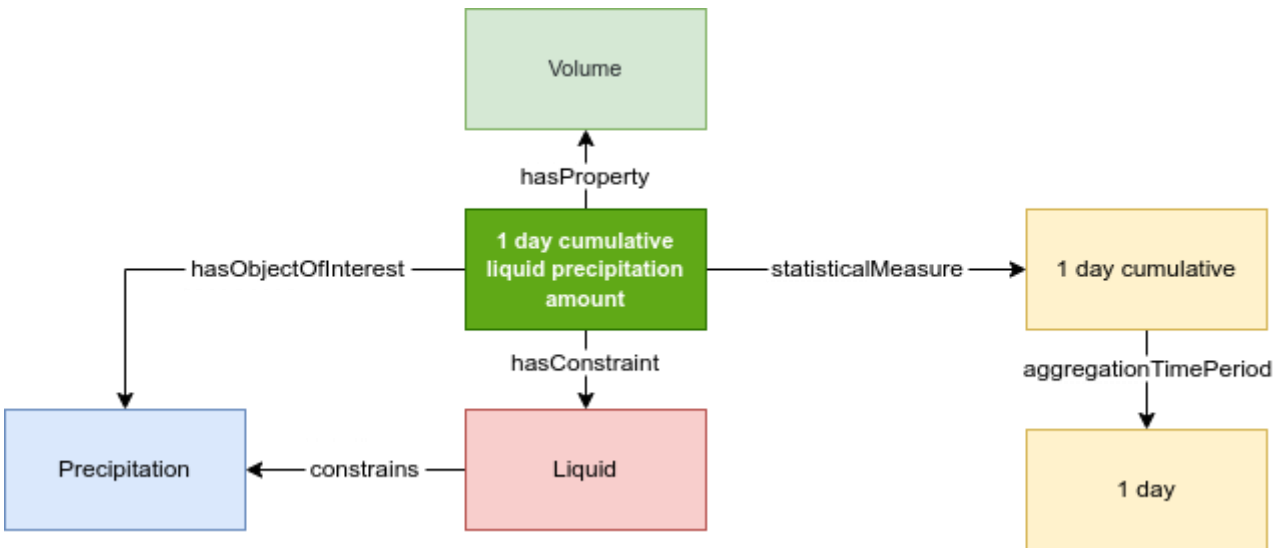
- **Physical entity**
  - Chemical entity
  - Environmental entity
    - Atmosphere
    - Biosphere
    - Cryosphere
    - Hydrosphere
      - Groundwater
        - Karst water
        - Water table
      - Surface water
        - Lake
        - Pond
        - River
        - Spring
      - Water
  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

Implementation of I-ADOPT

- Decomposition of variables names into atomic concepts :

  - Property

  - Entity: roles [ObjectOfInterest, ContextObject, Matrix]

  - Constraint (depth, temperature, wavelength, …)

- 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 dimensions 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

- Variables labels often too complicated to be used directly for data discovery. We implemented our own "Simplified Label" for use on the web portal
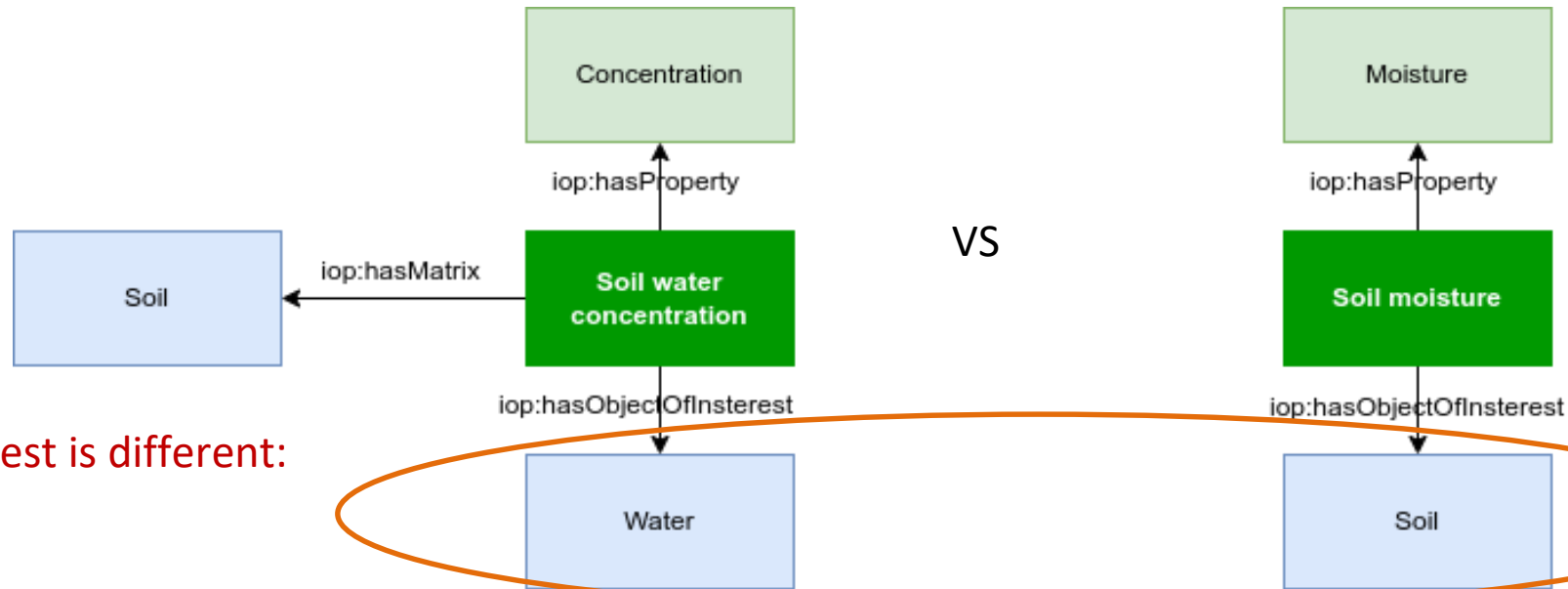
PREFERRED TERM   **1 day cumulative liquid precipitation amount**

SIMPLIFIED LABEL   Precipitation amount

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

The object of interest is different: water versus soil

Limitations of I-ADOPT

- Difficulty for modelling some variables such as a flux between two compartments (ex: carbone dioxide flux between the land surface and the atmosphere)
  - ⇒ Choice to add a Constraint to specify the direction of the flux
  - ⇒ Choice to document the two compartments using the Context concept

Specification of the two compartments

Constraint to specify the direction of the flux
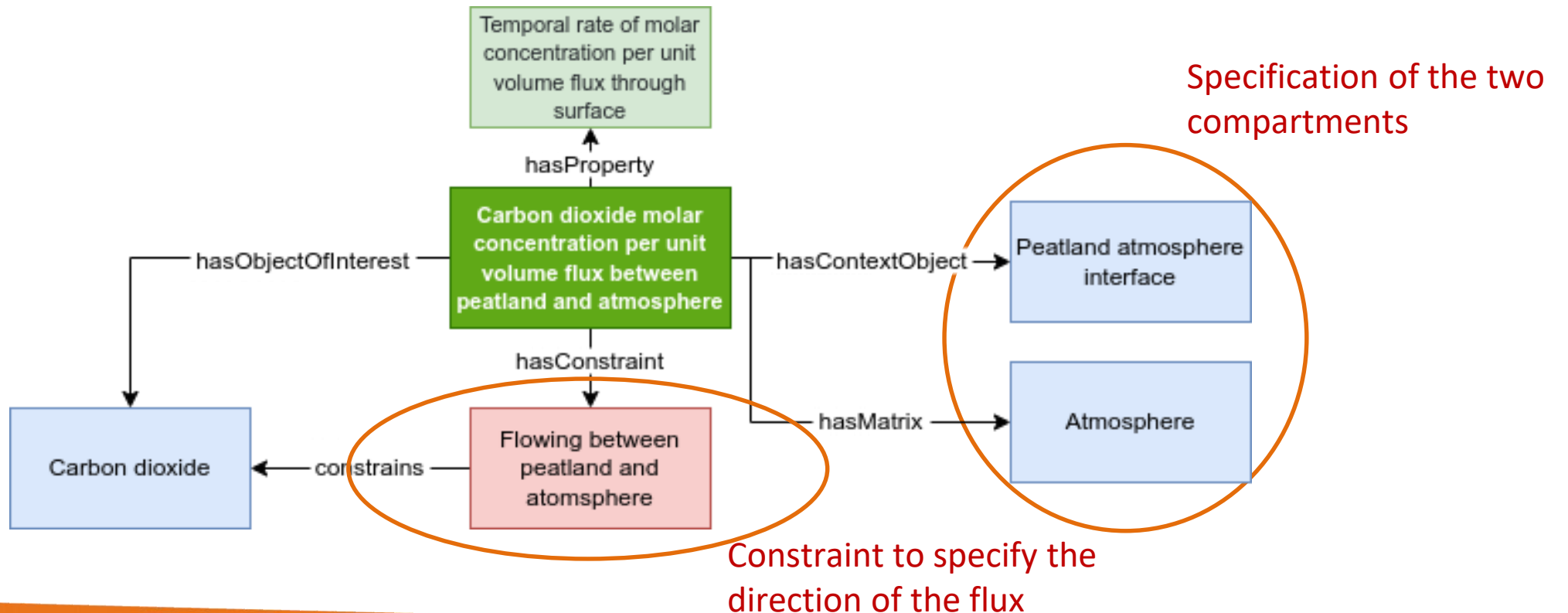
Limitations of I-ADOPT

- Difficulty for modelling some variables such as a ratio between two Properties (ex: $^{11}B/^{10}B$ isotopic ratio)
  $\Rightarrow$ Choice to use SKOS skos:related relationship to specify the two elements of the ratio

Specification of the elements of the ratio using specific SKOS relationhips
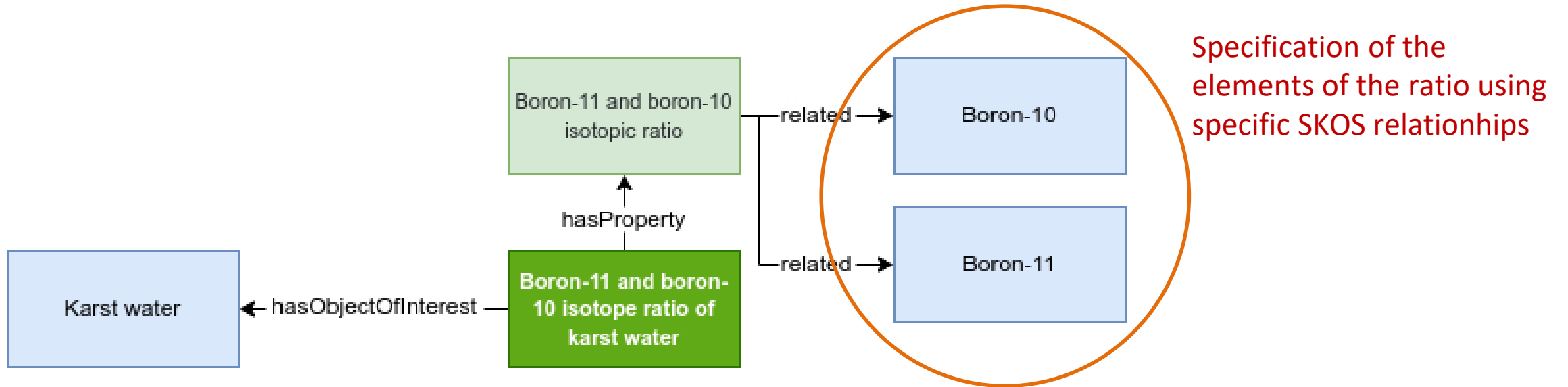
Benefits

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)
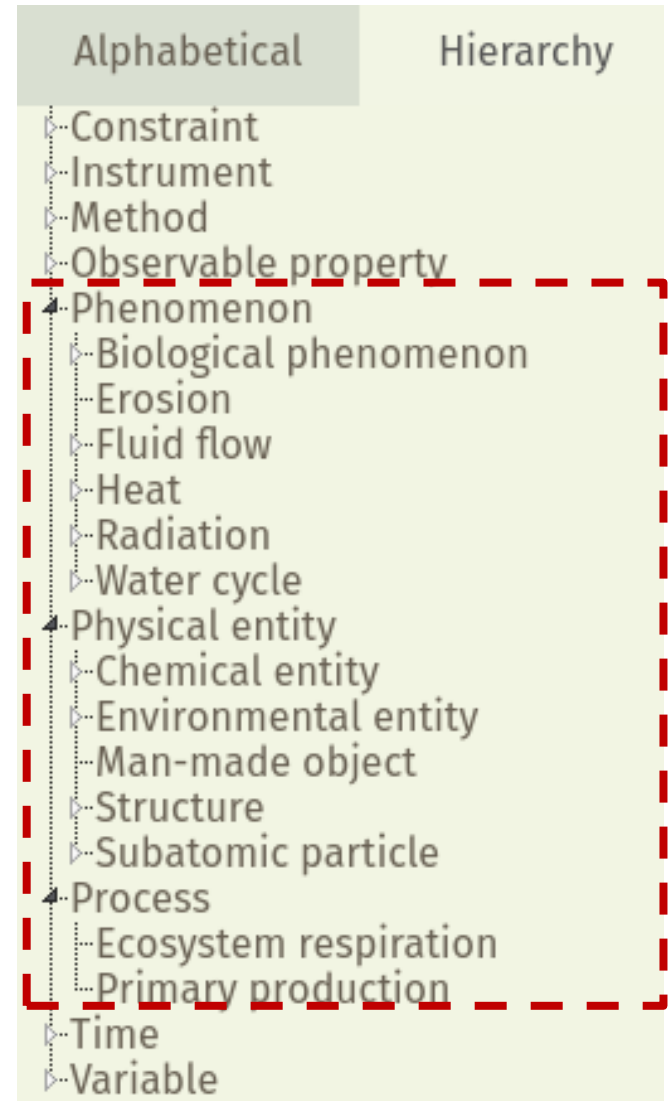
Benefits

✓ 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

Benefits

✓ Allows an unambigous 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
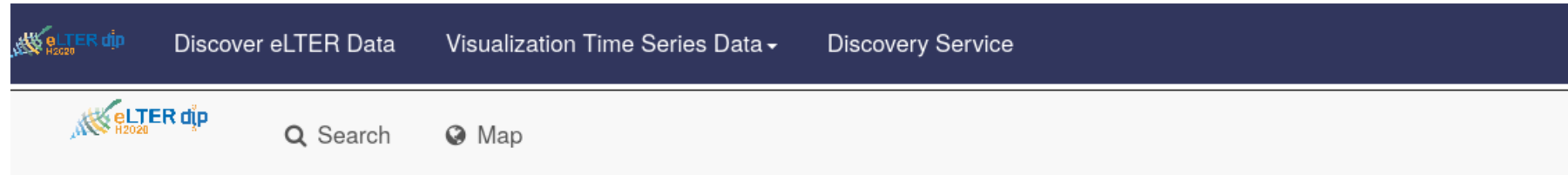
To learn more about the project:

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

Thank you for your attention: Questions ?