Espace DEV
OBSERVATION SPATIALE, MODÈLES & SCIENCE IMPLIQUÉE

Espace-Dev

Leveraging Knowledge Graphs for Earth System Dataset Discovery

Vincent Armant

Vincent Armant, Felipe Vargas-Rojas, Victoria Agazzi, Jean-Christophe Desconnets, Isabelle Mougenot, Valentina Beretta, Stephane Debard, Danai Symeonidou, Amira Mouakher, Joris Guérin, Thibault Catry, Emmanuel Roux

Séminaire Sésame 8 Septembre 2025

RÉPUBLIQUE FRANÇAISE Liberté Égalité Fraternité
IRD Institut de Recherche pour le Développement FRANCE
UNIVERSITÉ DE MONTPELLIER
UR UNIVERSITÉ DE LA RÉUNION
Université de Guyane
RÉPUBLIQUE FRANÇAISE Liberté Égalité Fraternité
UNC UNIVERSITÉ NOUVELLE-CALÉDONIE
Université des Antilles
MUSE MONTPELLIER UNIVERSITY OF EXCELLENCE

Context : Earth System Data Discovery

Data Terra Research Infrastructure

Data Hubs:

Ocean

Atmosphere

Continental Surface

Solide Earth

Geographic Data Standard: ISO 19115

Contexte : Découverte des données du Système Terre

Data Terra Research Infrastructure

Data Hubs:

Ocean

Atmosphere

Continental Surface

Solide Earth

Geographic Data Standard: ISO 19115

Guide of Good Practice, but not really followed

Not suitable to represent various dimensions of observation

Object or Features of Interest,

observable properties, sampling protocol

Contexte : Découverte des données du Système Terre

Data Terra Research Infrastructure

Data Hubs:

Ocean

Atmosphere

Continental Surface

Solide Earth

Geographic Data Standard: ISO 19115

Guide of Good Practice not followed

Not suitable to represent various dimensions of observation

Object or Features of Interest,
observable properties, sampling protocol

Semantic and structural heterogeneities

Obstacles for conducting pluridisciplinary studies involving data from different hubs

UCMM: PluriDisciplinary MetaData Integration Model

Data Terra Research Infrastructure

Data Hubs:

Ocean

Atmosphere

Continental Surface

Solide Earth

User Centric Metadata Model (UCMM)
MetaData integration model (application ontology)
focusing on observation paradigm in pluridisciplinary context

- Eases dataset discovery in multi-source setting
- Relies on SOSA (to represent various dimension of observation)
- Bridges SOSA and DCAT (to represent data catalog)
- Reuses other well known standard : CPM, SWEET, REPR, SKOS, TIME

What is an Ontology, a Data model, or Knowledge Graph ?

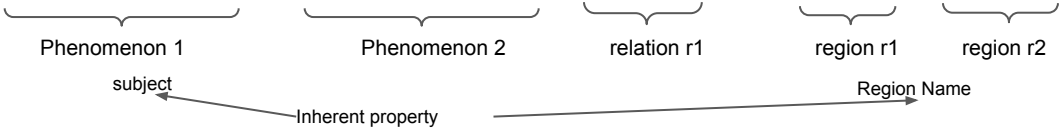
Ontology (computer science):

A set of concepts organised in a graph whose relationships can be semantic, compositional or inheritance.

ex: Let's build one

Informal description to be formalised

The phenomena of 'climate change' and 'vector-borne diseases' are studied in the 'Sahel' and 'Amazon' respectively.



Important when building an ontology

Clearly identify entity/objects/concepts, their types (i.e. Classes), the relationships between objects and classes.

Example: Knowledge Formalisation

Ontology: (Data Model + Knowledge Graph)

ex: Informal Description

The phenomena of ‘climate change’ and ‘vector-borne diseases’ are studied in the ‘Sahel’ and ‘Amazon’ respectively.

Phenomena p1 | Phenomana p2 | relation r1 | region r1 | region r2

ex: Formal Description

Data model, Schema, TBox

- **PHENOMENON** : The class representing phenomena,
- **REGION**: The class representing regions,
- p **isStudiedIn** r : the relation stating that p ∈ PHENOMENON is studied in r ∈ REGION,
- o **isA** c : the relation stating that an object o has the type c,
- p **hasSubject** l : the property of a PHENOMENON stating that p ∈ PHENOMENON o has subject l ∈ STRING.
- r **regionName** n **:** the property of a REGION stating that r ∈ PHENOMENON has region name n ∈ STRING

Knowledge Graph, Fact, ABox

- p1 **isA** PHENOMENON, p1 **hasSubject** ”climate change”
- p2 **isA** PHENOMENON, p2 **hasSubject** “vector-borne diseases”
- r1 **isA** REGION, r1 **regionName** “Sahel”,
- r2 **isA** REGION, r1 **regionName** “Amazon”,
- p1 **isStudiedIn** r1, p2 **isStudiedIn** r2,

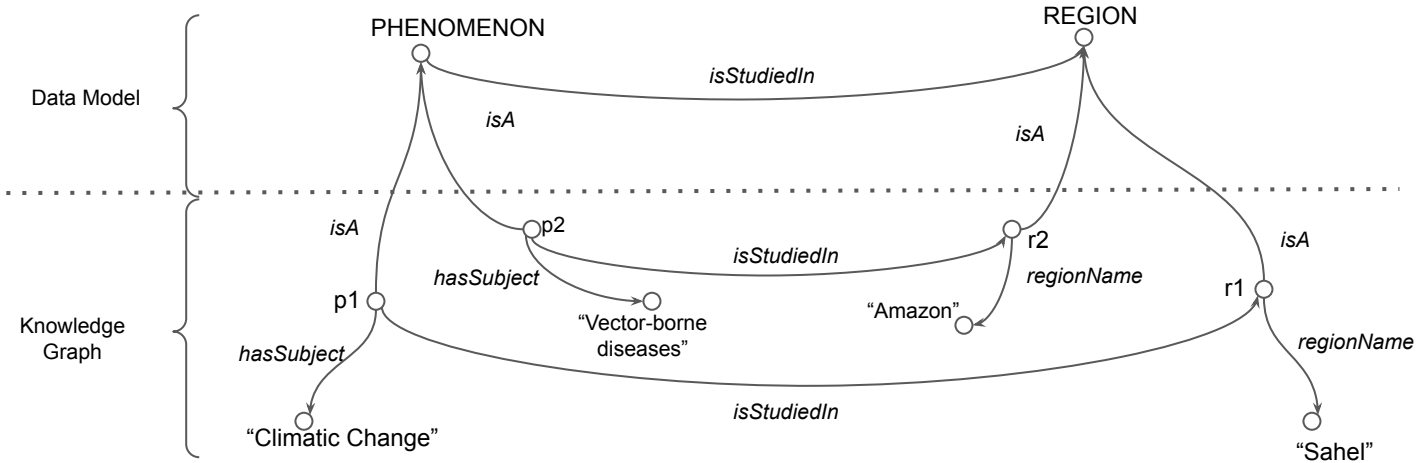
Example : Knowledge Graph

Ontology: (Data Model + Knowledge Graph)

The phenomena of ‘climate change’ and ‘vector-borne diseases’ are studied in the ‘Sahel’ and ‘Amazon’ respectively.

Phenomena p1 | Phenomana p2 | relation r1 | region r1 | region r2

ex: Knowledge graph



Application Ontology : Opening the Knowledge (FAIR Principles)

Reusing semantic relations defined in standardized ontologies to improve data interoperability, data reusability

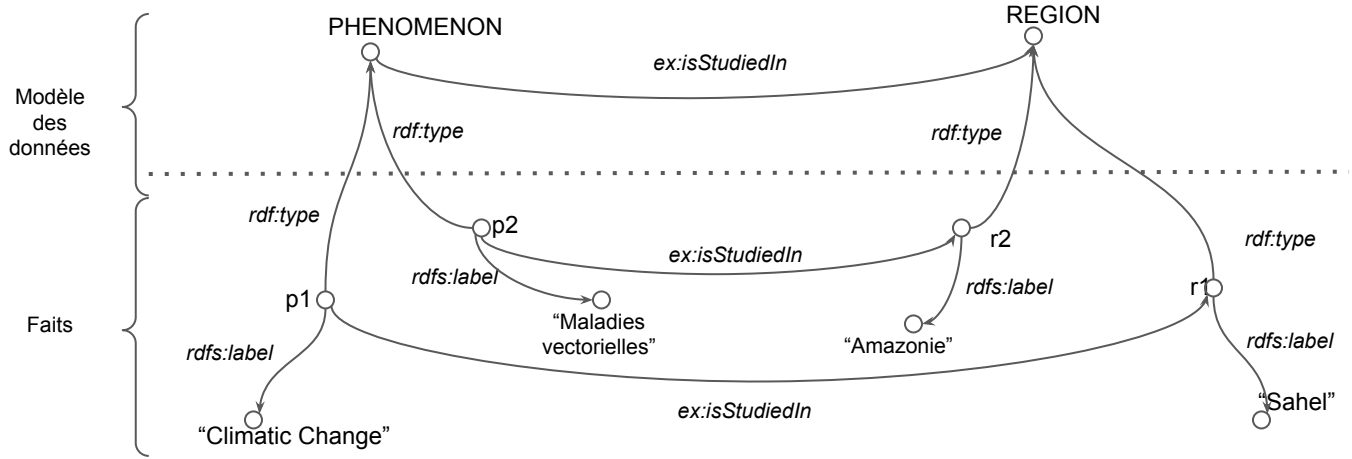
Some CORE ontologies

rdf: Resource Description Framework

rdfs: Resource Description Framework

owl: Ontology Web Language

ex: local KnowledgeGraph



Example of UCMM instance: ISAS-SSS

MetaDAta description ISAS-SSS dataset (portail ODATIS)

ISAS-SSS (In situ Sea Surface Salinity gridded fields)

-Observations from free-drifting profiling floats

-measures up to 2000 m depth

Declaration of namespaces

Prefixes :

i1: http://example.org/IASS-SSS#
dcat : http://www.w3.org/ns/dcat#
dct : http://purl.org/dc/terms/
geo : http://www.opengis.net/ont/geosparql#
repr : http://sweetontology.net/repr/
dtesv : https://terra-vocabulary.org/ncl/FAIR-Incubator/earthsciencevariables/
dtfoi : https://terra-vocabulary.org/ncl/FAIR-Incubator/earthfeaturetype/
skos : http://www.w3.org/2004/02/skos/core#
sosa : http://www.w3.org/ns/sosa/
time : http://www.w3.org/2006/time#
ucmm : http://purl.org/ucmm#

UCMM is an application profile (mainly reuse existing standard)

UCMM: PluriDisciplinary MetaData Integration Model

Data Terra Research Infrastructure

Data Hubs:

Ocean

Atmosphere

Continental Surface

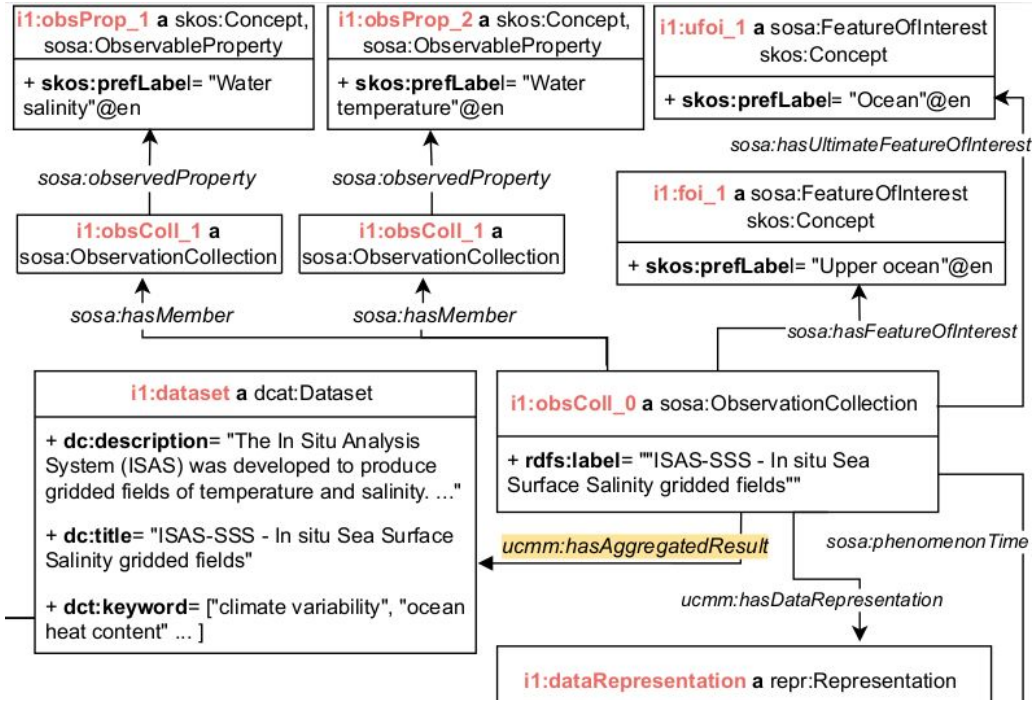
Solide Earth

Geographic Data Standard: ISO 19115

User Centric Metadata Model (UCMM)
MetaData integration model (application ontology)
focusing on observation paradigm in pluridisciplinary context

- Eases dataset discovery in multi-source setting
- Relies on SOSA (to represent various dimension of observation)
- Bridges SOSA and DCAT (to represent data catalog)
- Reuses other well known standard : CPM, SWEET, REPR, SKOS, TIME

Example of UCMM instance: ISAS-SSS

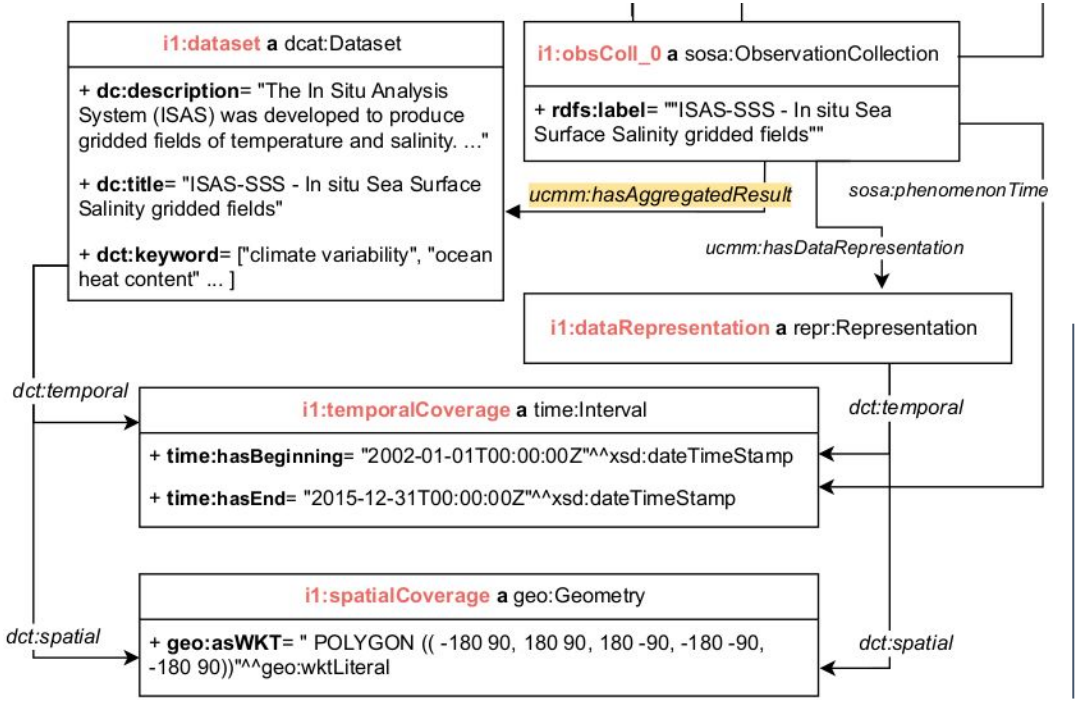


UCMM is based on SOSA Observation Paradigm.

SOSA : Sensor,Observation,Sampler and Actuator

UCMM considers metadata as important as data

Example of UCMM instance: ISAS-SSS

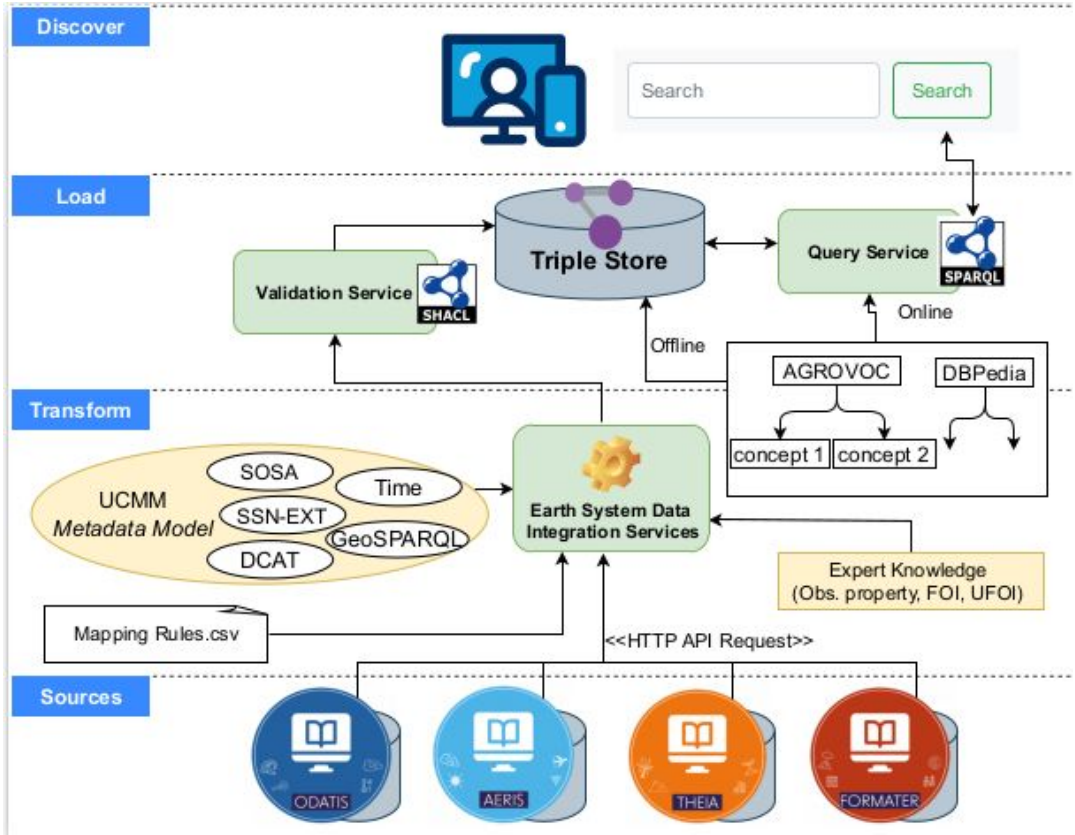


UCMM also relies on DCAT for describing catalogue data and other standard

Data Catalog Vocabulary

UCMM connects SOSA and DCAT

Architecture of the Earth System Data Open Discovery

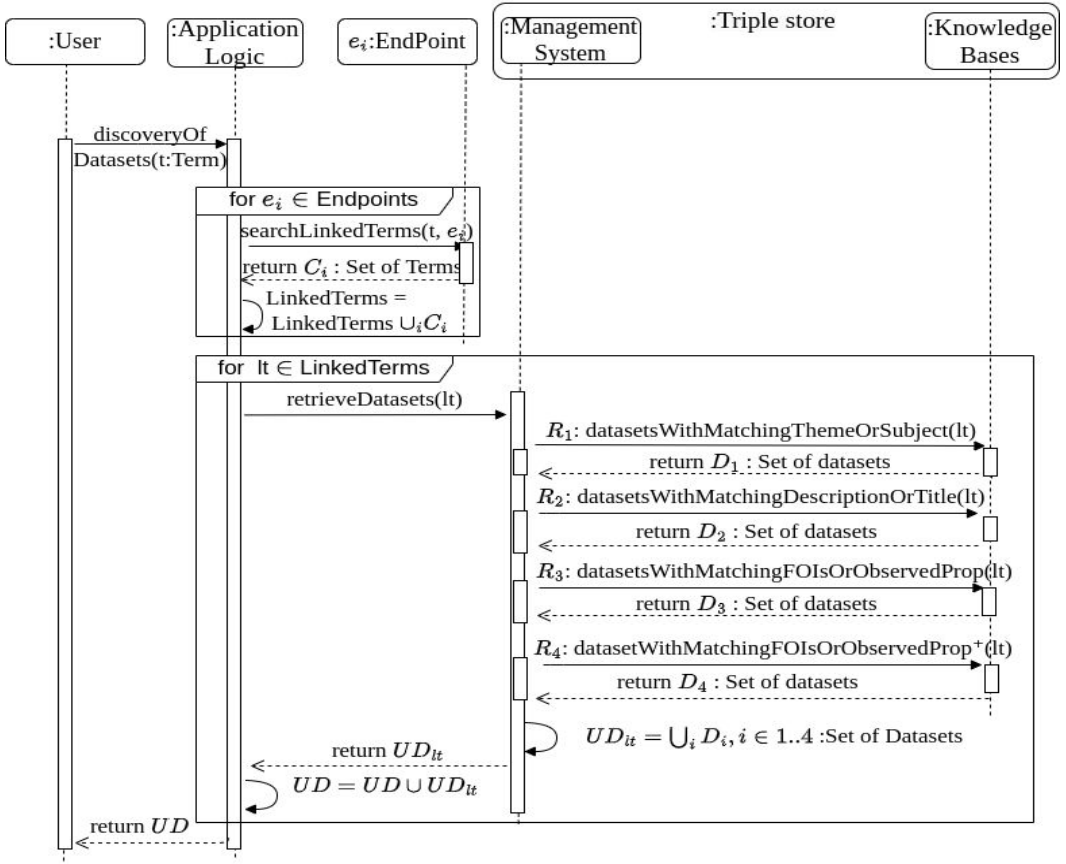


Search engine demo

Earth System Dataset Open Discovery

https://purl.org/earthsystemdatasetdiscovery/

Open Discovery of Datasets using external resources



Impact: Improving the Retrieval of Pluridisciplinary Datasets

	DATA HUB Knowledge Graphs					
	ODATIS	THEIA-LAND	THEIA-HYDRO	FORMATER	AERIS	merged KG
# datasets	10930	337	24594	255	2786	38902
# triples	669857	18071	1032948	13263	53493	1720876

Impact: Improving the Retrieval of Pluridisciplinary Datasets

	DATA HUB Knowledge Graphs					
	ODATIS	THEIA-LAND	THEIA-HYDRO	FORMATER	AERIS	merged KG
# datasets	10930	337	24594	255	2786	38902
# triples	669857	18071	1032948	13263	53493	1720876

	Number of retrieved datasets					
Search term	ODATIS	THEIA-LAND	THEIA-HYDRO	FORMATER	AERIS	merged KG
temperature	1149	80	0	33	378	1640
air	1788	70	25	25	491	2399
water	2427	189	24594	28	200	27438
carbon	268	40	0	2	99	409
conductivity	54	70	0	0	9	133

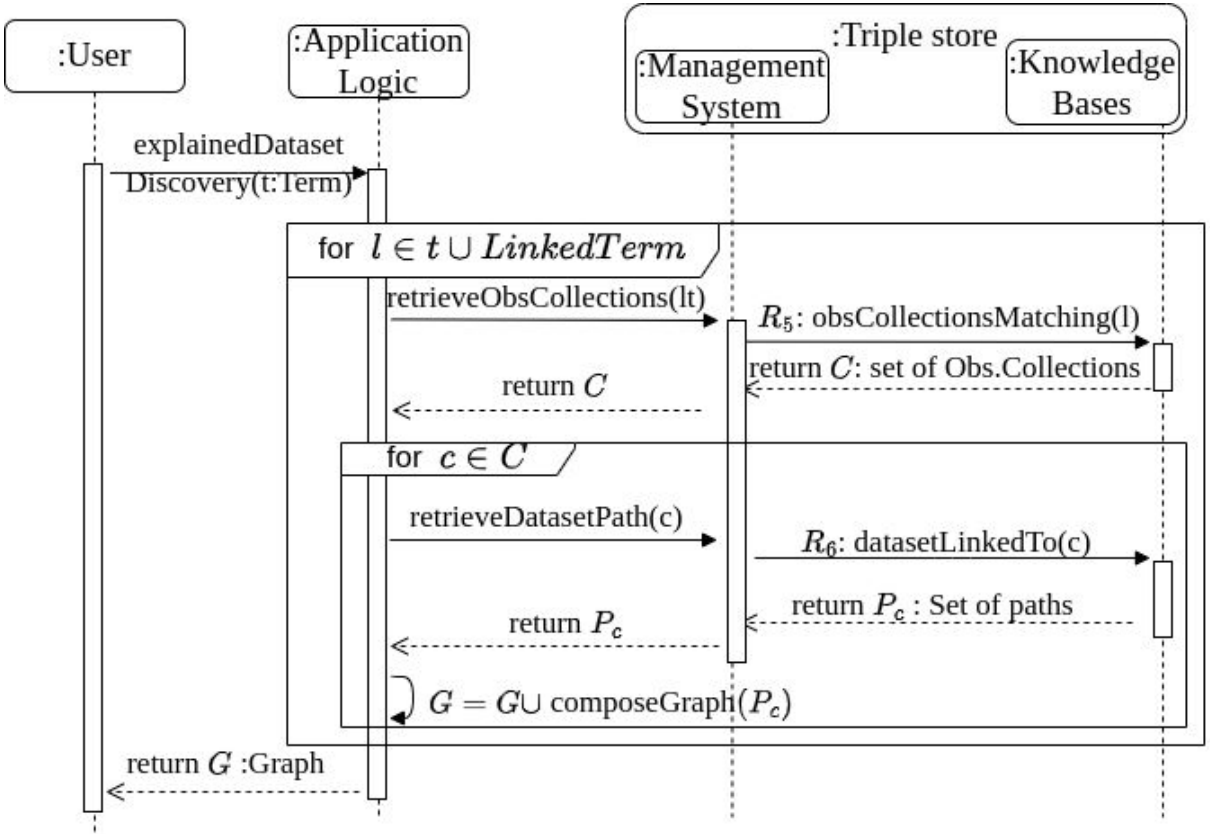
Impact: Improving the Retrieval of Pluridisciplinary Datasets

	DATA HUB Knowledge Graphs					
	ODATIS	THEIA-LAND	THEIA-HYDRO	FORMATER	AERIS	merged KG
# datasets	10930	337	24594	255	2786	38902
# triples	669857	18071	1032948	13263	53493	1720876

	Number of retrieved datasets					
Search term	ODATIS	THEIA-LAND	THEIA-HYDRO	FORMATER	AERIS	merged KG
temperature	1149	80	0	33	378	1640
air	1788	70	25	25	491	2399
water	2427	189	24594	28	200	27438
carbon	268	40	0	2	99	409
conductivity	54	70	0	0	9	133

	Dataset gain ratio					
Search term	ODATIS	THEIA-LAND	THEIA-HYDRO	FORMATER	AERIS	merged KG
temperature	0.43	19.50	-	48.70	3.34	-
air	0.34	33.27	94.96	94.96	3.89	-
water	10.31	144.17	0.12	978.93	136.19	-
carbon	0.53	9.23	-	203.50	3.13	-
conductivity	1.46	0.90	-	-	13.78	-

Explaining discovery (dealing with Observations Collections)



Uptake: User Experience

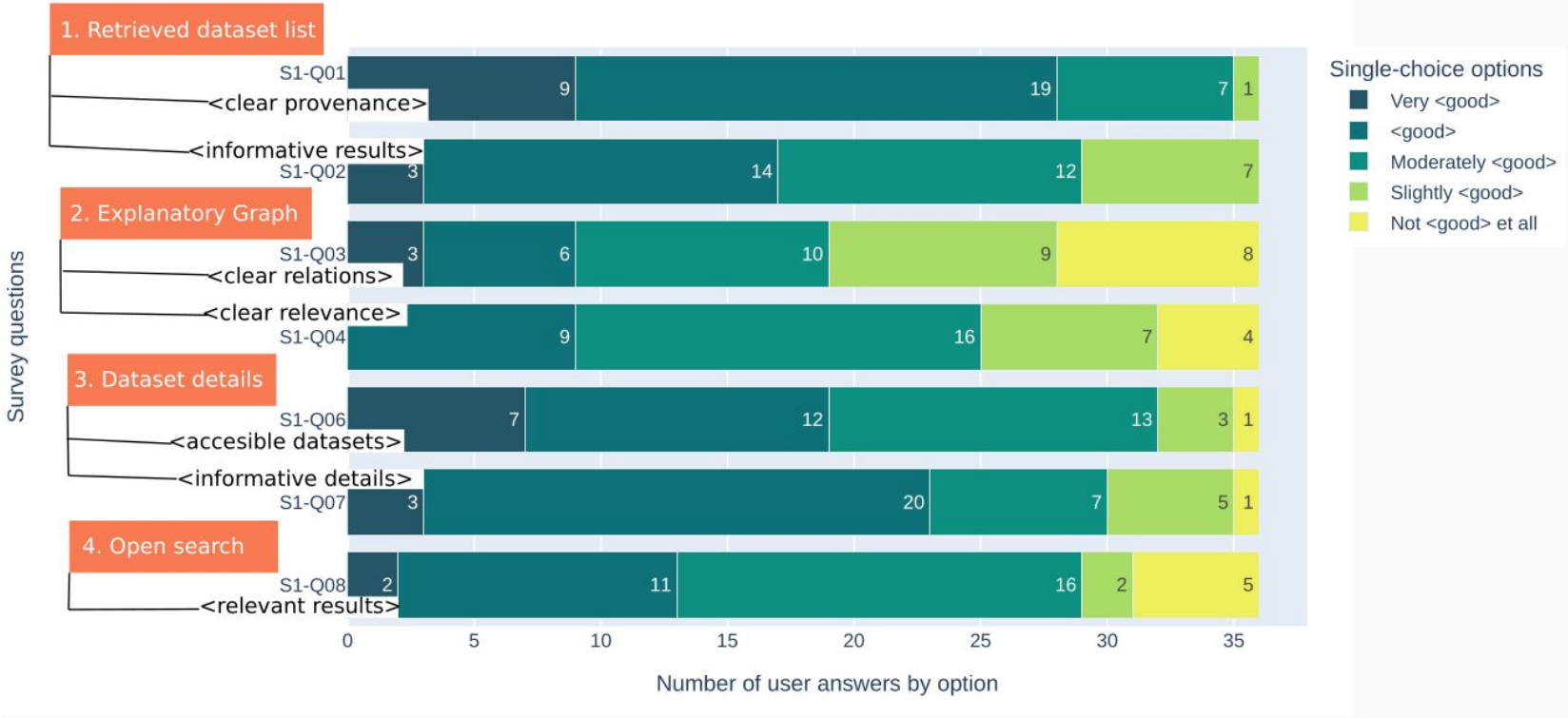


Figure 1: Predefined search: term temperature

Uptake: User Experience

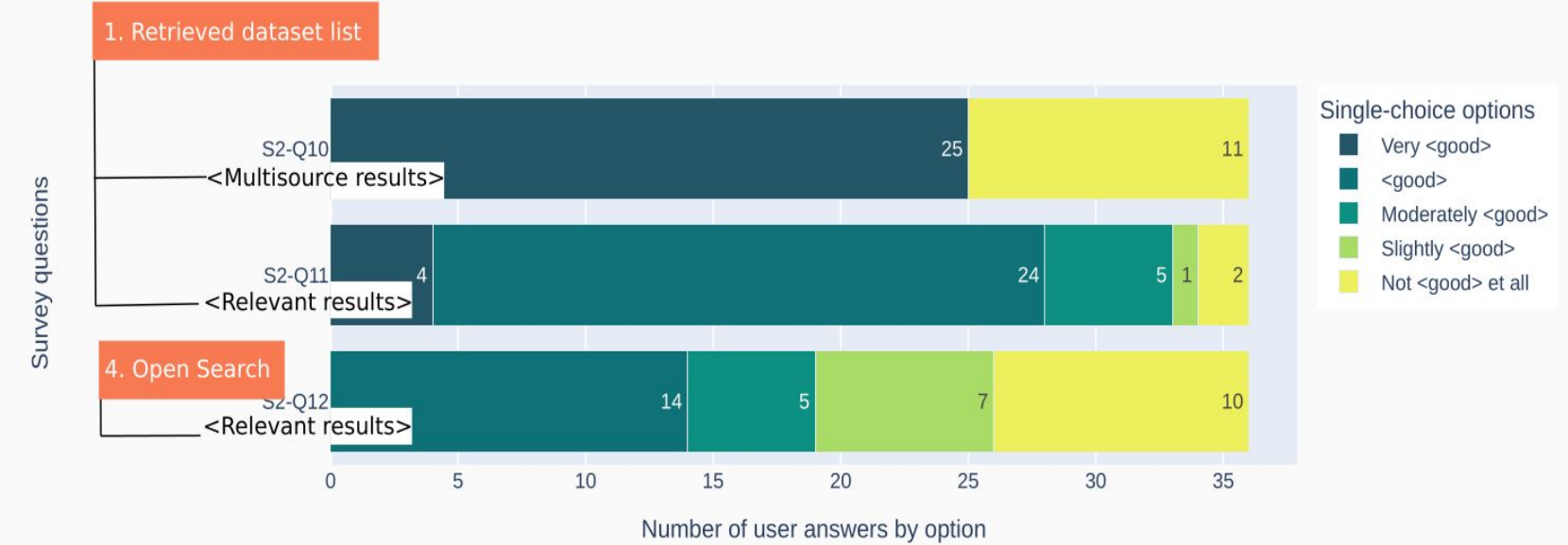


Figure 2: User defined search. E.g., Biodiversity, Ground deformation, Marine litter, Precipitation, Water level, currents, fishing vessels, health, metagenomics, rain, tropical rainforest, coral reef

Conclusion

- UCMM offered more precise annotations for pluridisplinary datasets in the Earth System domain and surpasses ISO 19115
- Multisource and pluridisplinary datasets were integrated in the ESDD system and we quantified the gain ratio
- The results of the user survey showed positive acceptance by the end users and room for improving concerning the explanatory graph

What is next?

- Verbalising the explanatory graph (LLMs)
- Automate the integration of datasets in the observation level
- Expanding the search scope beyond datasets (algorithms, code, reports, ...)

Thanks for your Attention

Demo:

Earth System Dataset Open Discovery

https://purl.org/earthsystemdatasetdiscovery/

ISWC 2024 article:

Leveraging Knowledge Graphs for Earth System Dataset Discovery

Vincent Armant, Felipe Vargas-Rojas, Victoria Agazzi, Jean-Christophe Desconnets, Isabelle Mougenot, Valentina Beretta, Stephane Debard, Danai Symeonidou, Amira Mouakher, Joris Guérin, Thibault Catry, Emmanuel Roux