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, et al.

Leveraging Knowledge Graphs for Earth System Dataset Discovery

Vincent Armant[1 [0000−0001−5797−7734]], Felipe Vargas-Rojas[1 [0000−0002−3530−7964]], Victoria Agazzi[1 [0009−0005−5249−9127]], Jean-Christophe Desconnets[2 [0000−0002−4142−5289]], Isabelle Mougenot[1 [0000−0002−1287−3269]], Valentina Beretta[1 [0000−0002−1060−3935]], Stephane Debard[1 [0000−0002−4671−109X]], Danai Symeonidou[3 [0000−0003−1152−5200]], Amira Mouakher[1 [0000−0002−1346−3851]], Joris Guérin[1 [0000−0002−8048−8960]], Thibault Catry[1 [0000−0001−9514−1751]], and Emmanuel Roux[1 [0000−0003−2266−8207at]]

[1] ESPACE-DEV, IRD, Université de Montpellier, Université de Perpignan, Université Antilles, Université de Guyane, Université de la Réunion, Montpellier, France
firstname.lastname@{ird.fr, umontpellier.fr, univ-perp.fr}
[2] Mission Science Ouverte, IRD, Montpellier, France
firstname.lastname@{ird.fr}
[3] UMR MISTEA, INRAE, Université de Montpellier, Montpellier, France
firstname.lastname@{inrae.fr}

Abstract. Thanks to open science initiatives, thousands of standardised datasets on Earth System compartments are now available on the web. In particular, we have widely used ISO 19115 to encode spatiotemporal aspects of Earth System observations. However, this standard does not specify the multiple dimensions of observations, including the features of interest, the observable properties, and the provenance. As a result, researchers interested in Earth System multi-disciplinary studies may miss meaningful datasets when querying independently domain-specific data portals. We propose a new Dataset Discovery System based on SOSA and DCAT ontologies, as well as the User-Centric Metadata Model (UCMM), to integrate dataset metadata from multiple data portals, each representing an Earth System compartment. The descriptive UCMM metadata model is exploited simultaneously to address semantic and structural heterogeneities and to build a descriptive Knowledge Graph explaining how retrieved datasets are semantically related to the user's search. We introduce the implementation of two Earth System Dataset Discovery use cases. The experiments and user uptake demonstrate the benefits of the Dataset Discovery System in multi-disciplinary Earth System studies.

Keywords: Open Discovery · Knowledge Graph · Earth System

1 Introduction

Monitoring the Earth System's compartments (atmosphere, solid earth, continental surface, and ocean) and their interfaces is crucial for understanding and

predicting its evolution. This necessitates the acquisition of a continually growing volume of diverse Earth System observations, encompassing satellite imagery, in-situ measurements, and airborne data. These observations are collected by instruments, processed, and subsequently disseminated through domain-specific data catalogs. Within the Data-Terra research infrastructure [13], four data hubs, each aligned with a specific Earth System compartment, provide various data services with machine-readable access. For example, a researcher investigating the effects of deforestation and climate change on the Amazon rainforest would need data from two hubs: 1) THEIA, which offers information on vegetation water stress and land cover changes associated with wildfires; and 2) AERIS, which provides data on the atmospheric impacts of wildfires, particularly aerosol content. THEIA and AERIS focus on observations from the continental surface and atmosphere, respectively. However, despite efforts to standardise data access APIs and output formats, inconsistencies persist in the structure and content of dataset metadata (i.e., descriptions). This hinders full interoperability across data hubs [12].

Due to heterogeneity in dataset metadata, researchers conducting cross-disciplinary studies encounter difficulties in efficiently retrieving datasets focusing on features of interest spanning multiple Earth System compartments. To retrieve datasets mentioning a search feature of interest, a user has to navigate between different data catalogues and explore dataset descriptions that may differ from one Earth System compartment to another. This manual search is surely time-consuming and error-prone. The first challenge this work addresses is improving the integration of dataset metadata and, in this way, the discoverability of multi-source datasets representing Earth System compartments.

From a user viewpoint, all retrieved datasets may not carry the same value. Datasets mentioning a search term in the abstract may not be as captivating as a dataset mentioning a search term as a subject or an observed property. Retrieving datasets in a multi-disciplinary setting is an achievement; understanding how a requested term relates to the retrieved datasets is another. Explaining how a retrieved dataset may relate to a search term is the second challenge addressed by our dataset discovery framework.

In this paper, we present the following contributions: (i) We use the User Centric Metadata Model (UCMM) [6], a pivotal model for integrating rich dataset metadata centred on the observation paradigm, to automatically construct a Knowledge Graph (KG) representing dataset descriptions of the different Earth System compartments of Data-Terra. (ii) We propose an open-linked dataset discovery process that uses external vocabularies to enrich user requests with connected concepts, as well as UCMM as an explanatory model to describe relationships between a requested term and retrieved datasets. (iii) We present a prototype with a user interface, improving both the dataset retrieval and the explainability of the discovery.

This paper is structured as follows: Section 2 reviews existing approaches and methodologies for achieving metadata interoperability in a broader context before narrowing the focus to Earth Science specifically. By concluding this review,

we justify the selection of the UCMM model for integrating dataset descriptions across Earth System compartments. Additionally, we highlight the novel contributions of our approach compared to the existing literature. Section 3 introduces the dataset discovery system's infrastructure, followed by a detailed presentation of the user interface in Section 4. Sections 4 and 5 show two implemented discovery use cases. Finally, Sections 6 and 7 discuss the system's impact and potential for adoption, while Section 8 summarises the key findings and concludes the paper.

2 Related work

The Data Catalogue Vocabulary (DCAT) defines datasets as collections of data managed by a single agent and available in various formats [4]. The W3C Data Cube Working Group describes datasets as collections of observations organised by a common structure [9]. Effective dataset retrieval, crucial for meeting user search criteria, depends on metadata, which ISO 11179 defines as descriptive data about an object [19]. Search engines utilise this metadata and contextual annotations to deliver relevant results, typically through keyword matching or Contextual Query Languages (CQLs) [7].

The standardisation of dataset retrieval through DCAT vocabularies and the recent update by Albertoni et al. [2] emphasise the importance of describing datasets within a catalogue structure using standardised models. This approach enhances interoperability among different catalogues, enabling the effective representation of datasets from various Earth System data talogs. However, the lack of interoperable metadata remains a significant challenge in efficiently discovering relevant datasets across multi-disciplinary and multi-source domains. To address this, two main approaches have been proposed: meta-model agreements and reconciliation [18]. Meta-model agreements involve using a central, standardised metadata model for deeper metadata integration, exemplified by the Earth science data model described by Crystal et al. [11], which uses consistent reporting formats to harmonise data descriptions across disciplines. In contrast, Beretta et al. [6] propose a user-centric approach to metadata interoperability in data lakes, establishing common ground between data providers and consumers to enhance dataset discovery and search. Similarly, Chen et al. [8] focus on exploiting the similarity of values/labels in metadata schemas and query content to improve search relevance. These approaches contrast with meta-model reconciliation methods, which aim to align heterogeneous metadata models. For instance, Nurcan et al. [12] present an ontology alignment method to establish semantic relationships between different ontologies. Khalid et al. [20] propose a metadata reconciliation strategy specifically designed to enhance data integration and classification. Their framework identifies potential matches between user datasets and a standard vocabulary by evaluating individual and inferred column similarities based on relationships between datasets, aiming to semi-automate the interlinking and validation process, thus facilitating the publication of linked data as a unified resource. Building on these efforts, Paneque et

al. [22] introduce e-LION, a novel semantic model specifically for the e-learning domain.

In the context of Earth System compartment observation, a consensus has been established on adopting the core ISO 19115 standard [19] for geographical data description. However, this standard lacks the expressiveness necessary to represent the full scope of Earth System observations. To address this limitation, Aldana-Martin et al. [3] propose a domain-specific ontology for remote sensing to represent and manipulate spatial information from remote sensing datasets. While both ISO 19115 and this domain-specific ontology capture spatiotemporal information effectively, they fall short in representing other crucial dimensions such as provenance, features of interest, variables, and observed properties. To integrate multi-source Earth System observation datasets, we use UCMM, a metadata model centred on the observation paradigm, combining SOSA and DCAT ontologies to represent a rich metadata model describing the various dimensions of observation datasets.

Current approaches to dataset retrieval primarily rely on keyword matching or query languages, often overlooking crucial semantic relationships between user requests and retrievable datasets. This limitation hinders the ability to identify datasets that not only match keywords but also align with the user's specific information needs. Our approach tackles this challenge by introducing a novel descriptive integration model that goes beyond traditional keyword-based methods by explicitly representing the semantic connections between user queries and datasets. This innovative approach enriches retrieved results with semantic context, facilitating a more nuanced understanding of user intent and demystifying the retrieval process. By highlighting these semantic connections, users gain valuable insights into why a particular dataset is relevant, fostering trust and transparency.

3 Earth System dataset discovery pipeline

It is important to integrate dataset metadata from the different Earth System compartments to ease the discovery of datasets for multi-disciplinary studies. To cope with multi-source data, we based our dataset discovery system on the Extract Transform Load process implemented by the pipeline shown in Figure 1.

Metadata descriptions are computed periodically. Consequently, the discovery service does not reflect the instantaneous picture of the up-to-date version of all datahubs' metadata. To achieve a more current version, we can shorten the update period. Due to time constraints during dataset discovery and the complexity of metadata harmonisation (a key aspect of our centralized approach for addressing semantic and structural heterogeneities), an offline implementation is more suitable for our use cases. This process is not suitable for a federated approach that requires responsiveness and online processing. While our approach does not offer up-to-date access to resources, it does provide fast and stable access to periodically updated knowledge bases.

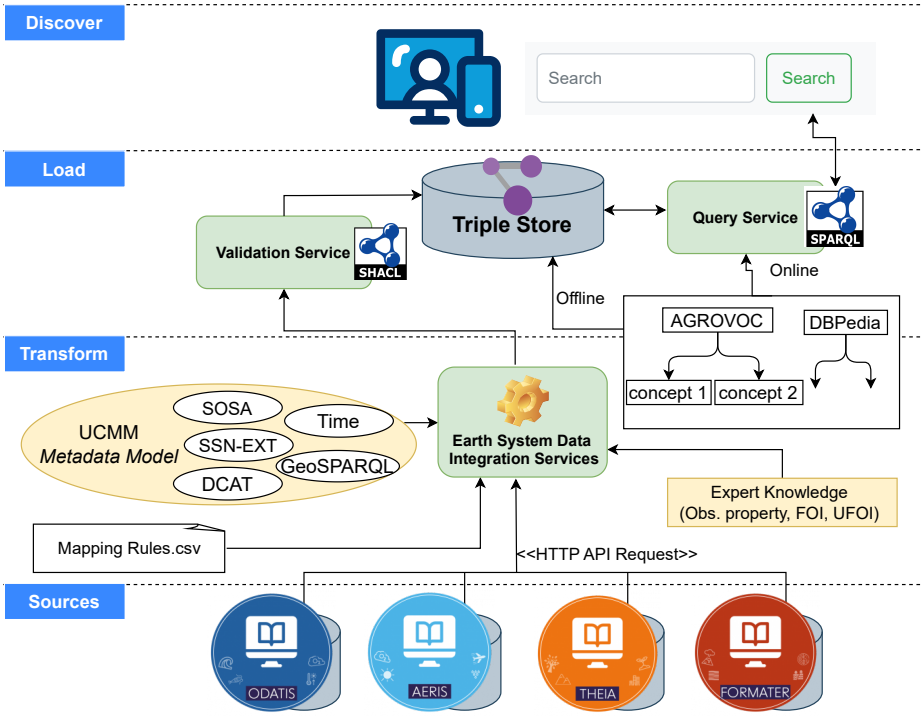


Fig. 1. Multi-source dataset discovery pipeline.

Dataset description harvesting (Extract). To enable multi-source dataset discovery, dataset descriptions are harvested from four data hub services: ODATIS [1], THEIA [23], FORM@TER [17], and AERIS [1], representing each compartment of the Earth System, enabling access to data products and services to support the observation of the ocean, the continental surfaces, the solid Earth and the atmosphere, respectively.

Dataset metadata Integration (Transform). We have made an initial effort to facilitate access to the multi-source dataset metadata. We have adopted the ISO 19115 standard as the dataset model and implemented the GeoNetwork API across the data hubs. Despite this initial effort, some semantic and structural heterogeneities remain in the harvested dataset metadata.

The primary purpose of the ISO 19115 standard is to describe spatiotemporal observations; it does not represent other aspects of observations, such as the subject, features of interest, or observable properties. These aspects of observation have been insufficiently specified and grouped under the keyword entity of ISO 19115. As a result, the harvested metadata contains some semantic heterogeneities. In addition to semantic heterogeneities, structural heterogeneities persist in the harvested metadata. The data hubs may encode the same information at different locations. For instance, the path specifies keywords as an xlink

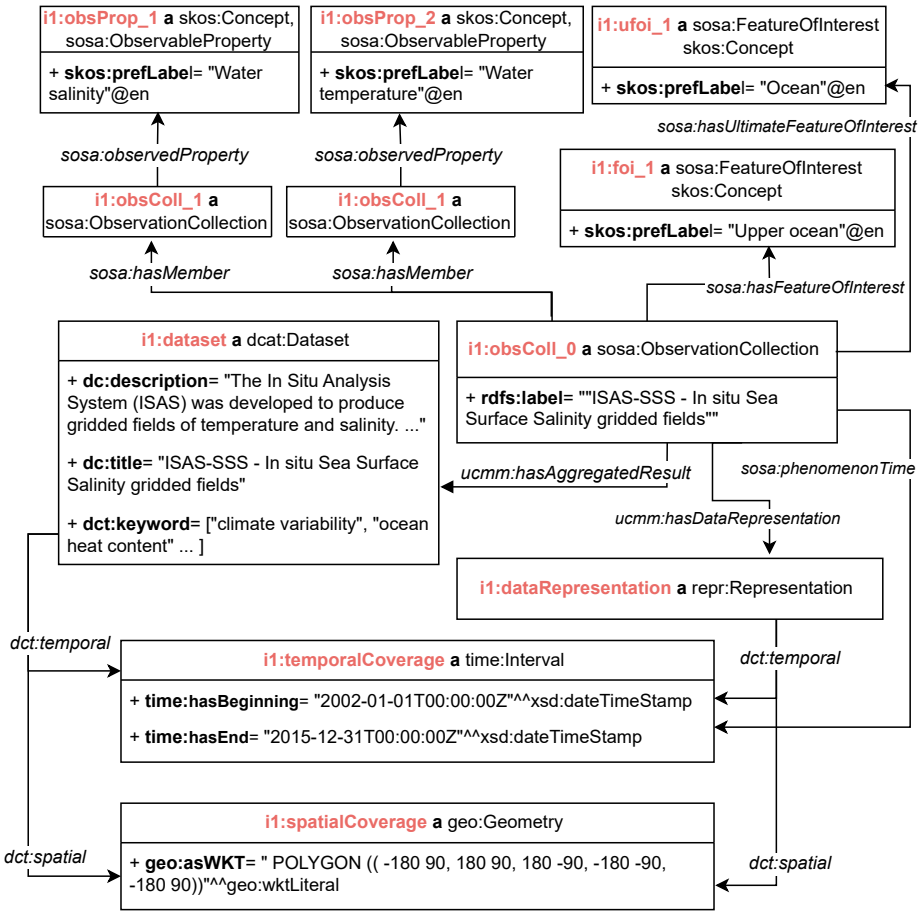


Fig. 2. Instance of UCMM Model: 'ISAS-SSS - In situ Sea Surface Salinity gridded fields'.

attribute[4], or text value at the path[5]. We use UCMM as a central metadata model and turn each harvested ISO19115 metadata into a first UCMM metadata instance in order to harmonise metadata and deal with these semantic and structural differences. One can understand UCMM as a model that aligns DCAT

[4] /MD_Metadata/identificationInfo/MD_DataIdentification/descriptiveKeywords/MD_Keywords/keyword/Anchor

[5] /MD_Metadata/identificationInfo/MD_DataIdentification/descriptiveKeywords/MD_Keywords/keyword/PT_FreeText/textGroup/LocalisedCharacterString

and SOSA concepts, thereby accessing the advantages of both models. Figure 2 exhibits the UCMM general structure along with an instantiation of the model representing the dataset 'ISAS-SSS - In situ Sea Surface Salinity Gridded Fields' extracted from ODATIS[6].

This dataset shows observations made in real time by ARGO[7] global free-drifting profiling floats that are measuring the top 2,000 metres of the ocean. The measurements were collected globally from 2002-01-01 to 2015-12-31, with the primary properties of interest for this work being 'Water salinity' and 'Water temperature'. We can identify the "Upper Ocean" as the feature of interest in the observations. This knowledge is primarily interpretable by domain experts with a deep understanding of the ocean observation system. Non-experts best understand the ultimate feature of interest in the observations, which is the 'Ocean' itself. These multiple observational aspects are crucial in the faceted search for filtering and retrieving relevant datasets. In the current implementation, we declare mapping rules (see Supplemental Material) that help to parse input GCMD entities into in-memory UCMM objects. Because we use in-memory objects that require complex and customised processing, we did not yet use RML [16] at the first step.

Triple Store Ingestion (Load). At the end of the transformation phase, each UCMM metadata instance is validated leveraging SHACL [21], a W3C recommendation. SHACL enables the validation of the GeoSPARQL geometries, the consistency of the time intervals, the tree-based structure of UCMM observation collections, and other constraints. SHACL constraints ensure the proper functioning of the user interface, providing higher-quality metadata. Further information is shared in our repository (see Supplemental Material). A SPARQL endpoint adds the validated instances to a queryable RDF Triple Store. We chose RDF4J-Server to store and access dataset metadata using SPARQL queries. This solution offers a reasonable trade-off between performance and ease of implementation without precluding using a more efficient triple-store management system such as GraphDB or Virtuoso. Note that we manipulate dataset metadata rather than the datasets themselves; as a result, we do not face high volumetric issues.

Dataset retrieval (Discover). After loading dataset metadata into a triple store, various applications can begin to exploit it. The dataset discovery application interacts simultaneously with the triple store and external online vocabularies to retrieve dataset metadata corresponding to the user search criteria. The application interprets a user search and its options into a SPARQL query, which it then sends to the triple store management system, as detailed in the next section.

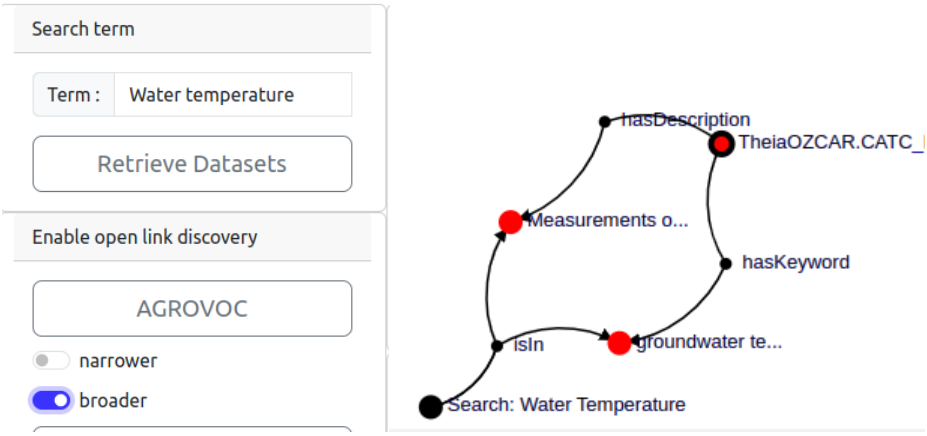
[6] XML available in: <`https://sextant.ifremer.fr/geonetwork/api/collections/OCEANO_PHYSIQUE_SPATIALE/items/97b4842b-94b3-4205-8781-476813d8177b?f=xml`> acceded 25-03-2024

[7] ARGO Operational Oceanography: <`https://www.coriolis.eu.org/Observing-the-Ocean/ARGO`> acceded 16-04-2024

4 Main functionalities of the User Interface (UI)

The main objective of the dataset discovery user interface is to facilitate the retrieval of relevant datasets in a multi-disciplinary context. To this end, the dataset discovery application proposes to users (domain experts and non-experts) a faceted search, allowing them to extend a search term with linked concepts. Not only are the retrieved datasets listed in a table, but a graph also displays the relationships explaining the semantic links between a search term and the retrieved datasets. By clicking on the dataset ID in the table, a user can access the detailed description of a retrieved dataset. In the sequel, we describe the main UI components implementing the dataset discovery use cases.

The **dataset search form** shown in Figure 3 (a) top, invites users to fill up a search term that is expected to match one of the dataset metadata elements loaded in the triple store. To help users formulate their dataset search, we added auto-completion functionality that initially harvests a subset of terms from the triple store. A user can fill up any free text.



(a) Dataset Search (b) explanatory graph

Fig. 3. Selected User Interface components.

The **open vocabulary add-on** shown in Figure 3 (a) bottom, allows users to extend the dataset search with related concepts from external vocabularies. This feature addresses both domain-expert and non-expert user profiles involved in multidisciplinary studies. Experts and general users may want to extend dataset search with other domain-specific vocabularies and more general terms beyond their knowledge to increase the chance of dataset retrieval.

The **Retrieved dataset list** and the **dataset info view** show the result of the Open-linked Discovery use case and display, respectively, the retrieved datasets in a table list and a detailed description of harvested dataset metadata. When a result table displays datasets retrieved from a linked concept, the head of the

table shows how the search term relates to the linked concept. By clicking on the dataset ID, a user can access the details of the dataset.
The **explanatory graph view** shows the result of the explained discovery use case. Each path between a search term and a dataset represents the relationships that connect the search term to the retrieved dataset. In the example shown in Figure 3 (b), the search term is part of the description of the dataset 'TheiaOZCAR.CATC_DAT_CE.Gwat_Odc' and appears as a keyword of the same dataset.

5 Implementation of dataset discovery use cases

We have implemented two complementary dataset discovery use cases that facilitate the retrieval of the Earth System datasets. The first use case, the open-linked discovery of Earth System datasets, consists of extending the dataset search term with linked concepts from open-linked and domain-specific vocabularies. This scenario specifically caters to non-expert or expert users who need to gather datasets from a different area of expertise. The second use case, which explains the discovery of Earth System datasets, uses a knowledge graph to illustrate the semantic connections between the retrieved datasets and the user's search. The entire Dataset Discovery System is based on a Model View Controller (MVC) design pattern where the User Interface implements the view, both the application logic and the Triple Store Management System implement the controller, and the Triple Store Knowledge Base implements the model.

5.1 Use Case 1: Open Discovery of Earth System Datasets

Figure 4 describes the processing workflow that computes the retrieved datasets matching a search term and the linked concepts from external vocabularies. When receiving a dataset search term, the application logic uses external terminologies and open-linked vocabularies to associate additional linked concepts with the initial user request. For each term within the dataset search term and linked concepts, the application logic creates four sub-queries matching the term to specific metadata model entities and merges them into a single query. The Triple Store Management System then receives the merged query for evaluation. After the query evaluation, the Triple Store Management System returns the retrieved datasets to the application logic. For each processed term, the application logic gathers the returned datasets before sending them back to the user.

The SPARQL Listing 1.1 describes the four sub-queries processed during the dataset retrieval. The first sub-query R_1:$datasetsWithMatchingThemeOrSubject(t)$ returns the set of datasets D_1 for which the theme or the subject matches the search term t. The second sub-query R_2:$datasetsWithMatchingDescTitleOrKw(t)$ returns the set of datasets D_2 for which the description, the title or a keyword match t. The third subquery R_3:$datasetsWithMatchingFOIsOrObservedProp(t)$, gives back the set of datasets D_3, which is the result of an

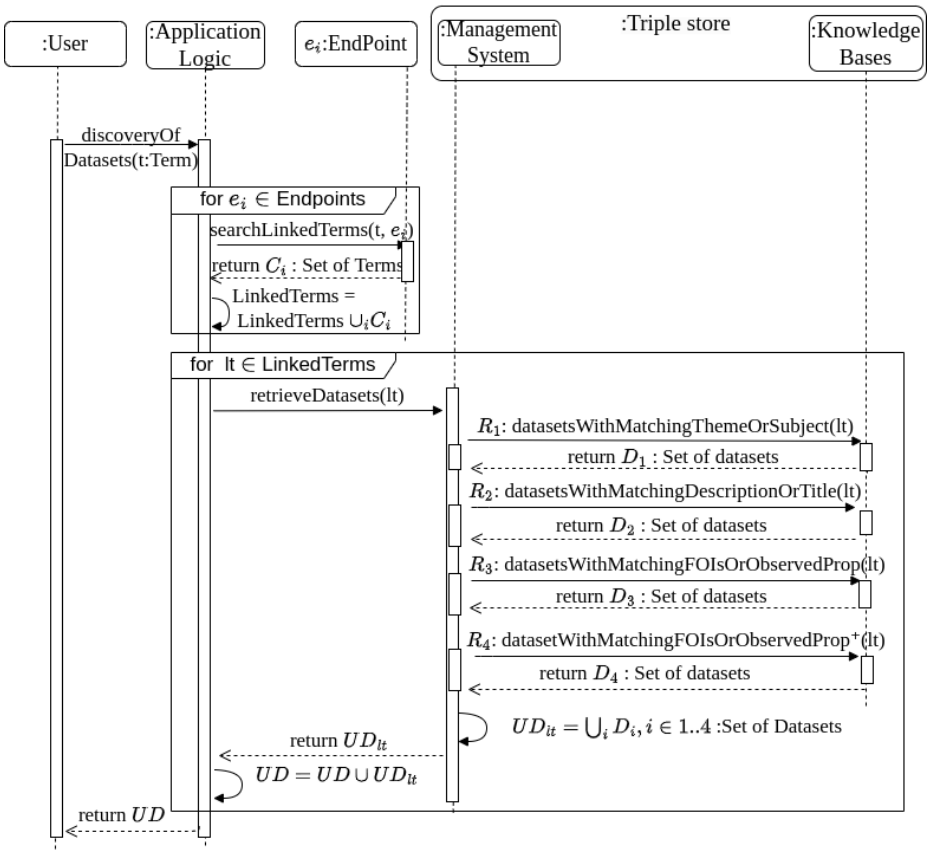


Fig. 4. Open Discovery of Earth System datasets.

observation collection that has at least one observable property and a feature of interest matching t. The sub-query R_4:$datasetsWithMatchingFOIsOrObservedProp^+(t)$ returns the set of datasets D_4 resulting from nested observation collections that contain an observable property or a (ultimate) feature of interest matching the search term. Empty dataset bindings associated with nested observation collections are removed from D_4.

Listing 1.1. retrieveDatasets(t) SPARQL query

```
# full list of name spaces for the paper
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX ucmm: <http://purl.org/ucmm#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dcterms: <http://purl.org/dc/terms/>
```

```
SELECT DISTINCT ?dataset
WHERE{
  {
    # R1: datasetsWithMatchingThemeOrSubject(t)
    ?dataset rdf:type dcat:Dataset .
      dcterms:subject|dcterms:theme ?concept .
    ?concept skos:prefLabel ?t .
  }UNION{
    # R2: datasetsWithMatchingDescTitleOrKw(t)
    ?dataset rdf:type dcat:Dataset .
      dcterms:description|dcterms:title|dcat:keyword ?t.
  }UNION{
    # R3: datasetsWithMatchingFOIsOrObservedProp(t)
    ?dataset rdf:type dcat:Dataset .
      ucmm:hasAggregatedResult ?dataset .
    ?obsColl rdf:type sosa:ObservationCollection .
      ?obsCollProp ?concept.
    ?concept skos:prefLabel ?t .
  }UNION{
    # R4: datasetsWithMatchingFOIsOrObservedProp+(t)
    ?obsColl (^sosa:hasMember)+ ?obsCollParent.
      sosa:FeatureOfInterest|sosa:hasUltimateFeatureOfInterest|sosa:observedProperty ?
          concept.
    ?concept skos:prefLabel ?t .
    OPTIONAL{
        ?dataset rdf:type dcat:Dataset.
        ?obsCollParent ucmm:hasAggregatedResult ?dataset.
    }
    # remove empty dataset bindings
    FILTER (strlen(str(?dataset)) > 0)
  }
  FILTER (lang(?t)="en" && regex(?t,{t}","I") )
}
```

5.2 Use Case 2: Explained Discovery of Earth System Datasets

From a user or domain-expert viewpoint, a search term appearing in a dataset abstract may not have the same value as a search term present as a feature of interest or observed properties of an observation collection. Although the open-linked discovery of datasets increases the possibility of retrieving meaningful datasets, it does not explain how retrieved datasets semantically relate to dataset search terms. In this regard, most dataset discovery approaches, including open-linked discovery, may be considered black boxes.

Figure 5 describes the processing workflow for computing a knowledge graph, illustrating how the dataset searches and the retrieved datasets semantically relate to each other. To achieve this, the application logic initially behaves as the Open Linked Discovery use case when receiving a dataset search term. The system extends the search term by including related concepts. Next, we request the triple store to provide the paths connecting the retrieved datasets to the metadata model's various dimensions, such as theme, subject, and description, rather than just the datasets themselves. Since SOSA allows for the embedding of observation collections, we compute the paths linking datasets to features of interest and observed properties in two steps. We first collect observation collections that match features of interest or observed properties. Next, we draw the paths from the gathered observation collections to the datasets. Upon completion

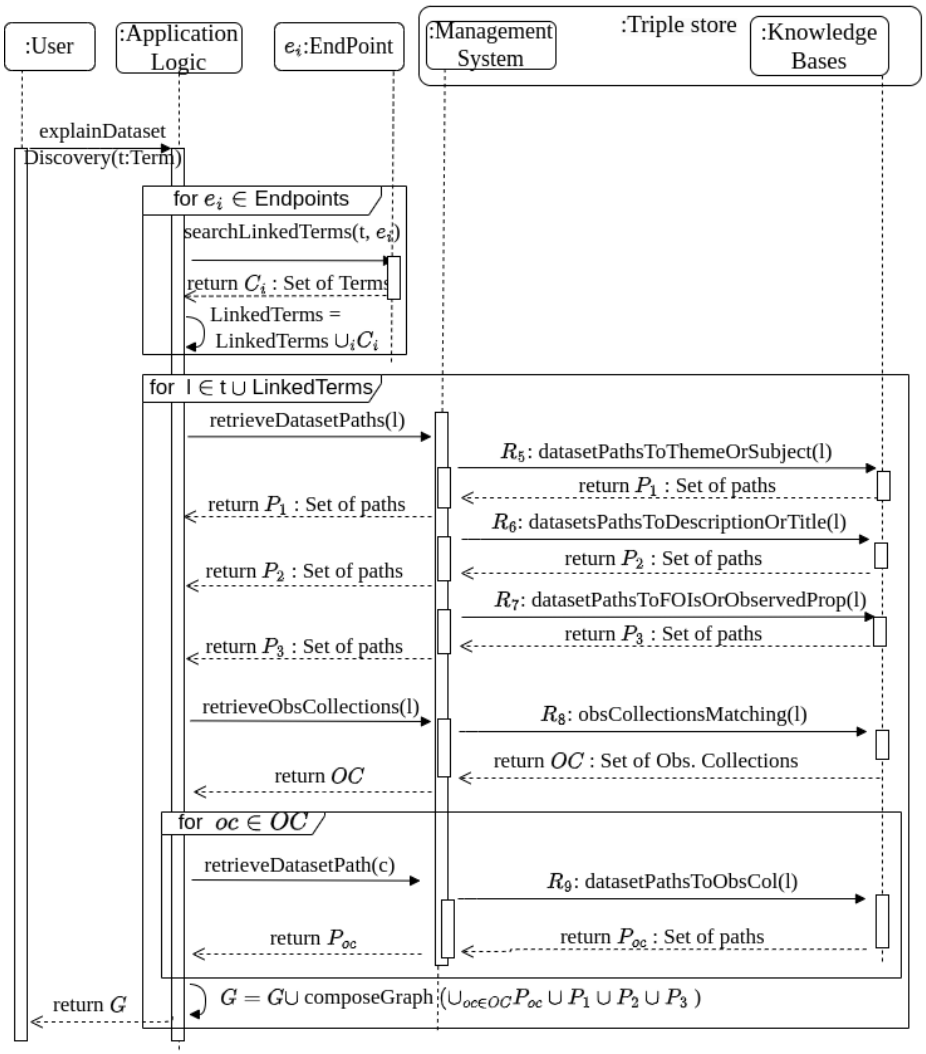


Fig. 5. Explained Discovery of Earth System datasets.

of the workflow, we combine all the computed paths between retrieved datasets and matching metadata elements to construct an explanatory knowledge graph. The graph expresses the computed explanations that emphasize the relationship between dataset search and dataset metadata, not the retrieval process itself.

6 Impact

Without multi-disciplinary Earth System Dataset Discovery (ESDD), a user interested in multi-disciplinary studies would have to individually query different

domain-specific Earth System data portals and collect the returned datasets. Afterwards, the user should manually fix the structural and semantic heterogeneity of these datasets.

Table 1. Evaluation of the discovery system with five search terms (temperature, air, water, carbon, conductivity). We report the number of (#) retrieved datasets for each data hub and the merged version. We calculate a dataset gain ratio as (# results (merged KG) - # results (data hub))/ # results (data hub).

	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
Search term	*Number of retrieved datasets*					
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
Search term	*Dataset gain ratio*					
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	-

To evaluate the impact of Knowledge Graph (KG) technologies for users interested in multi-disciplinary dataset searches, we compared the benefits of using a multi-disciplinary Earth System Data Discovery (ESDD) approach against individual Earth Compartment data hubs. This evaluation, summarised in Table 1, tested dataset discovery using five search terms ('temperature', 'air', 'water', 'carbon', 'conductivity') that represent Earth System concepts intersecting at least two of the five tested data hubs. We assessed dataset retrieval from each individual data hub KG and compared it to the retrieval from the union of all the data hub KGs.

The KGs representing ODATIS, THEIA-LAND, THEIA-HYDRO, FORMATER, and AERIS dataset metadata encode 10930, 337, 24594, 255, and 2786 dataset descriptions, respectively, with 669857, 18071, 1032948, 13263, and 53493 triples, respectively. Within the ODATIS KG, the search term 'temperature' matched 1149 datasets, while the merged KG matched 1640 datasets, indicating a 43% increase in retrieved datasets. For THEIA-LAND KG, we observed a significant gain ratio of 19.5 when comparing locally retrieved datasets (80 datasets) to those retrieved from the merged graph (1640 datasets).

The dataset gain ratio varied between 0.43 and over 978, depending on the search term. These quantitative results demonstrate the substantial improvement

in discovery capacity achieved by integrating metadata from different Earth System data hubs compared to conducting dataset discovery within a single data hub KG.

7 Uptake

Recent efforts in Data-Terra's data hubs aim to harmonise dataset descriptions in alignment with the FAIR principles endorsed by the Research Data Alliance. This initiative has led to two significant outcomes. First, a best practice guide[8] for creating onto-terminologies has been developed, which adheres to FAIR principles and is detailed in [10]. Second, the Data Terra on-terminologies incubator is now accessible via the Terra Vocabulary Linked Data Registry service[9]. The Earth System Dataset Discovery framework builds upon this vocabulary harmonisation to enhance dataset retrieval across multiple data portals. These standardised onto-terminologies are used by the application profile for federated Data Terra metadata[10], which is based on DCAT, to describe dataset themes using the *dcat:theme* property.

In addition to adopt Data-Terra guidelines and best practices, we carried out a user survey to assess the uptake and usability of the Earth System Dataset Discovery tool.

Survey structure: the survey aimed to evaluate the relevance and clarity of the ESDD functionalities through two scenarios: S1 (predefined search) and S2 (user-defined search). In S1, users begin with a predefined search term ('temperature') and extend the search using narrower concepts from the AGROVOC vocabulary. In S2, users input their own keywords and expand their search with concepts from external vocabularies.

User profiling: the survey was conducted with 36 participants, including researchers (57%), students (14%), and engineers (28%). Over 80% of respondents claimed expertise in at least one scientific field (hydrology, energy modelling, remote sensing, machine learning, biostatistics, marine ecology, etc) and familiarity with FAIR principles or Linked Open Data. Approximately 60% had prior experience with Data-Terra data hubs.

Survey analysis: Figure 6 presents the results of 10 single-choice questions evaluating user experiences regarding (i) the retrieved dataset list, (ii) the explanatory graph, (iii) dataset details, and (iv) the open search functionality. Dark colours indicate positive user feedback. Notably, S2-Q10 is a yes-or-no question.

[8] Best practices for Data Terra thesaurus creation: <https://gitlab.in2p3.fr/gaia-data/wp3-services/vocabulaires/linked_data_registry/guidelines_thesaurus> acceded 16-04-2024, language French

[9] Data Terra terminology service: <https://terra-vocabulary.org/ncl/_FAIR-Incubator> acceded 16-04-2024

[10] Guidelines Data Terra profile application: <https://gitlab.in2p3.fr/gaia-data/wp3-services/vocabulaires/dataterra_ap/guidelines_profil_application> acceded 16-04-2024, language French

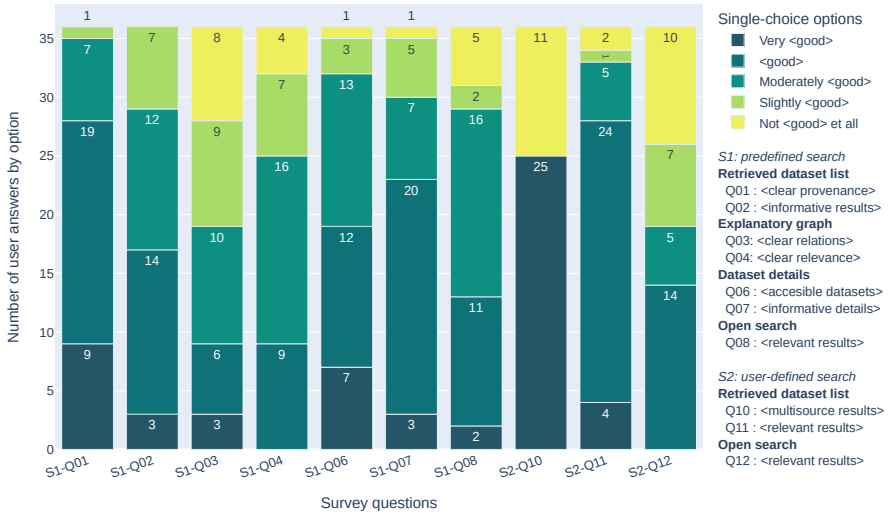


Fig. 6. Distribution of 36 users' answers to scenarios S1 and S2 questions.

About 70% of users reported retrieving results from multiple data hubs for their search terms (S2-Q10), with 72% indicating clear provenance (S1-Q01) and 50% finding the information informative or very informative (S1-Q02). Approximately two-thirds of users (S1-Q07) found the dataset details component, which provides additional information about a selected dataset, to be informative or very informative, and 90% reported the dataset details to be very easy to access (S1-Q06). One of the study's objectives was to evaluate the relevance of the open search functionality, which offers additional search terms from external vocabularies to enhance the retrieved dataset list. The outcomes for this functionality differed between the two scenarios. For predefined keyword searches (S1-Q08), 36% of users found the open search results relevant, and 44% assessed them as moderately relevant. For user-defined searches (S2-Q12), 39% of users found the results relevant, and 14% rated them as moderately relevant. This suggests that while the open search functionality is generally useful, its effectiveness is slightly higher when users define their own search terms, possibly because these terms are more tailored to their specific information needs. Another objective was to help users better understand the semantic relationships between a search term and the retrieved datasets using an explanatory graph. About half of the users found the relationships displayed by the graph understandable (28% moderately understandable, 17% easily understandable, and 8% very easily understandable, S1-Q03). Approximately 70% of users found the graph helpful in assessing datasets' relevance (S1-Q04). However, 53% of users suggested adding more textual representation, such as verbalising the explanatory graph. The full survey and results are available as Supplemental Material.

8 Conclusion and perspectives

To conduct multi-disciplinary studies, experts gather datasets from several data portals, each providing datasets from observations of a different compartment of the Earth System. The widely used ISO 19115 standard, which focuses on the spatiotemporal aspects of Earth System observations, calls for a harmonisation effort. However, Earth System datasets still have some semantic and structural differences. These heterogeneities, which are due to the ISO 19115 standard's lack of modelling of the multiple dimensions of observations, represent obstacles to the collection and use of data sets. We created a new Earth System Dataset Discovery (ESDD) system based on SOSA and DCAT ontologies to make it easier to search for datasets and deal with issues related to heterogeneity. We use the UCMM metadata model as the primary metadata model to align dataset metadata from various data portals. The ESDD framework uses the UCMM metadata model to not only integrate dataset metadata and facilitate dataset discovery, but also to clarify to users the semantic relationship between the dataset search and the returned datasets. The results of the user survey, which assesses the user experience, show a positive acceptance by the end users. When you combine metadata from various Earth System data hub knowledge graphs, you can find a lot more information than when you just look for data in one data hub. There are several ways to improve and facilitate data search and discovery for multi-disciplinary research activities while still allowing effective open science codes, algorithms, and models to ensure they are available and understood by different communities. The following perspectives can aid in achieving this objective:

Expanding the discovery scope. Extending the framework beyond datasets to include the discovery of other relevant research resources [14], such as services, code repositories, and computational models, would provide users with a more holistic research environment.

Enriching explanations with LLMs. Leveraging Large Language Models (LLMs) throughout the query and explanation process has the potential to provide more comprehensive and nuanced justifications for retrieved datasets [15]. LLMs could look at the semantic connections between user queries, dataset descriptions, and maybe even the content of the datasets themselves. This would help give more complete and useful explanations [24]. This integration would require careful consideration of potential LLM biases and limitations, but it holds promise for significantly improving the user experience [5].

Supplemental Material Statement. Application prototype is available from: `https://purl.org/earthsystemdatasetdiscovery`. Code, User Survey, SHACL validation constraints, and mappings are available in this github repository: `https://purl.org/earthsystemdatasetdiscovery/supplementalMaterial`.

Acknowledgments. This study was made possible thanks to the metadata and data provided by ODATIS, AERIS, THEIA and FormaTer clusters of the Data Terra infrastructure and was partly funded by the MOSAIC Horizon Europe project (grant number: 101137398).

References

1. AERIS. https://www.aeris-data.fr/en/catalogue-en/, [Online; accessed - March-2024]
2. Albertoni, R., Browning, D., Cox, S., Gonzalez-Beltran, A.N., Perego, A., Winstanley, P.: The w3c data catalog vocabulary, version 2: Rationale, design principles, and uptake. Data Intelligence pp. 1–37 (2023), publisher: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ...
3. Aldana-Martín, J.F., García-Nieto, J., Roldán-García, M., Aldana-Montes, J.F.: Semantic modelling of Earth Observation remote sensing. Expert Systems with Applications **187**, 115838 (Jan 2022). https://doi.org/10.1016/j.eswa.2021.115838, https://linkinghub.elsevier.com/retrieve/pii/S0957417421012008
4. Archer, P.: Data catalog vocabulary (DCAT) (w3c recommendation) (2014-01), https://www.w3.org/TR/vocab-dcat/
5. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 610–623 (2021)
6. Beretta, V., Desconnets, J.C., Mougenot, I., Arslan, M., Barde, J., Chaffard, V.: A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences. Computers & Geosciences **154**, 104807 (2021). https://doi.org/10.1016/j.cageo.2021.104807, https://linkinghub.elsevier.com/retrieve/pii/S0098300421001060
7. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset search: a survey. The VLDB Journal **29**(1), 251–272 (2020-01). https://doi.org/10.1007/s00778-019-00564-x, http://link.springer.com/10.1007/s00778-019-00564-x
8. Chen, Z., Jia, H., Heflin, J., Davison, B.D.: Leveraging schema labels to enhance dataset search. CoRR **abs/2001.10112** (2020), https://arxiv.org/abs/2001.10112
9. Consortium, W.W.W., others: The RDF data cube vocabulary (2014), publisher: World Wide Web Consortium
10. Cox, S.J.D., Gonzalez-Beltran, A.N., Magagna, B., Marinescu, M.C.: Ten simple rules for making a vocabulary FAIR. PLOS Computational Biology **17**(6), e1009041 (Jun 2021). https://doi.org/10.1371/journal.pcbi.1009041, https://dx.plos.org/10.1371/journal.pcbi.1009041
11. Crystal-Ornelas, R., Varadharajan, C., O'Ryan, D., Beilsmith, K., Bond-Lamberty, B., Boye, K., Burrus, M., Cholia, S., Christianson, D.S., Crow, M., et al.: Enabling fair data in earth and environmental science with community-centric (meta) data reporting formats. Scientific data **9**(1), 700 (2022)
12. Dang, V., Aussenac-Gilles, N., Megdiche, I., Ravat, F.: Interoperability of Open Science Metadata: What About the Reality? In: Nurcan, S., Opdahl, A.L., Mouratidis, H., Tsohou, A. (eds.) Research Challenges in Information Science: Information Science and the Connected World, vol. 476, pp. 467–482. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-33080-3_28, https://link.springer.com/10.1007/978-3-031-33080-3_28
13. DATA TERRA. https://www.data-terra.org/en/, [Online; accessed -March-2024]
14. Desconnets, J.C., Giuliani, G., Guigoz, Y., Lacroix, P., Mlisa, A., Noort, M., Ray, N., Searby, N.D.: GEOCAB Portal: A gateway for discovering and accessing capacity building resources in Earth Observation. International Journal of Applied Earth Observation and Geoinformation **54**,

95–104 (Feb 2017). https://doi.org/10.1016/j.jag.2016.09.010, https://linkinghub.elsevier.com/retrieve/pii/S0303243416301672
15. Dhole, K.D., Chandradevan, R., Agichtein, E.: An interactive query generation assistant using llm-based prompt modification and user feedback. arXiv preprint arXiv:2311.11226 (2023)
16. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: Rml: A generic language for integrated rdf mappings of heterogeneous data. Ldow **1184** (2014)
17. ForM@Ter. https://en.poleterresolide.fr/data-access/catalog/#/, [Online; accessed -March-2024]
18. Haslhofer, B., Klas, W.: A survey of techniques for achieving metadata interoperability. ACM Computing Surveys **42**(2), 1–37 (Feb 2010). https://doi.org/10.1145/1667062.1667064, https://dl.acm.org/doi/10.1145/1667062.1667064
19. ISO: Geographic information-metadata. ISO 19115: 2003 (2003)
20. Khalid, H., Zimanyi, E., Wrembel, R.: Metadata reconciliation for improved data binding and integration. In: Beyond Databases, Architectures and Structures. Facing the Challenges of Data Proliferation and Growing Variety: 14th International Conference, BDAS 2018, Held at the 24th IFIP World Computer Congress, WCC 2018, Poznan, Poland, September 18-20, 2018, Proceedings 14. pp. 271–282. Springer (2018)
21. Knublauch, H., Kontokostas, D.: Shapes constraint language (shacl). Tech. rep., W3C (07 2017), https://www.w3.org/TR/shacl/
22. Paneque, M., Del Mar Roldán-García, M., García-Nieto, J.: e-lion: Data integration semantic model to enhance predictive analytics in e-learning. Expert Systems with Applications **213**, 118892 (2023)
23. THEIA. https://catalogue.theia-land.fr/, [Online; accessed -March-2024]
24. Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., Wen, J.R.: Large language models for information retrieval: A survey. arXiv preprint arXiv:2308.07107 (2023)