eosc | FAIR-IMPACT

Expanding FAIR solutions across EOSC

Stories of practical implementation of the **FAIR principles**

Real-life use cases on social sciences and humanities, photon and neutron sciences, life sciences and agri-food and environmental sciences brought to you by the FAIR-IMPACT partners.

FAIR-IMPACT identifies practices, policies, tools and technical specifications to guide researchers, repository managers, research performing organisations, policy makers and citizen scientists towards a FAIR data management cycle. The focus is on persistent identifiers (PIDs), metadata, ontologies, metrics, certification and interoperability, starting with real-life use cases on social sciences and humanities, the photon and neutron sciences, life sciences and agri-food and environmental sciences.

Disclaimer

January 2025

Table of Contents

EMBL-EBI - Providing harmonized information about organisations in identifiers.org using ROR registry

- **Key topic** Interoperability | Metadata & Ontologies PIDs
- **Scientific domain** Life science | Physical and Technical Sciences
- **Leading organisation** EMBL-EBI
- **Contributors** Henning Hermjakob, Renato Juacaba Neto

Overview

European Bioinformatics Institute is Europe's largest provider of public biomolecular data resources. The institute is co-located with Elixir Hub and partnered with many relevant EU projects, including EOSC-Life, FREYA, and BY-COVID. In addition to supporting life sciences, EMBL-EBI is increasingly collaborating with other domains, e.g., social sciences in COVID-19 research. **EMBL-EBI provides consistent access to life science data by leveraging compact identifiers through the Identifiers.org resolution service**. Anyone can use identifiers.org for consistent and globally unique references to objects in registered data collections. For example, researchers can use our link to their objects using our URIs and these will remain resolvable even if the object's online location changes. This use case focuses on the addition of metadata artefacts to the identifiers.org registry. **This mainly consists of the use of the ROR registry as a source of truth for institution data**. Allowing the identifiers.org registry to correctly link its resources to the institution in charge of managing it and verify its own metadata on them based on RORs metadata.

Introduction

The ambition of EMBL-EBI in this use case **is to curate and update components of Identifiers.org**. The updates will be aligned with community standards and needs while following FAIR practices. Our tasks for this goal revolve around verifying the set of annotations available at the identifiers.org registry meets the community guidelines for PIDs. Thus including:

- ☐ Adding or removing attributes as necessary for our system to provide minimal and sufficient context to PIDs
- ☐ Curating the registry to verify that the correct metadata is associated with entries

For this means, we added ROR identification to the institutions in our registry. The ROR registry of institutions acts here as a metadata catalogue and it is a source of truth on information such as home pages, type of organization and other identifiers for organizations.

In the identifiers.org registry, we link institutions to their ROR ID when possible, the associated ROR ID of the institution in charge of a resource can be seen in our registry pages. **This association is performed manually by our curators**. This association provides us with the means to identify institutions through a persistent identifier but our registry also stores other metadata on them locally that overlaps with other metadata managed by ROR. For example, both ROR and identifiers.org contain institution name and home page. **This provides use with the opportunity of updating our registry by using the ROR registry as a source of truth**. However, considerable curation effort is needed to add and maintain ROR IDs in current and future entries. Additionally, ensuring that the metadata from ROR is equal to the one in our registry takes considerable time.

The addition of ROR IDs to our registry is a small change that leverages both the identity of institutions and allows for cross validation of other metadata values. **It is an example of a minimum change that makes our PIDs more useful, but even this small example considerably increases our curation efforts**. This displays the importance of carefully considering what additions make sense in our type of service.

In the future, we wish to consider to expand the registry with associations to other metadata registries such as FAIRsharing and Wikidata and allow users to resolve compact identifiers to their metadata at these providers and use metadata from these services in our search services.

Challenges that need to be addressed

There is a need to find solutions to overcome some challenges related to PID practices and associated metadata, especially in cases where billions of data objects are included in a large number of resources, where:

- Repositories produce a large amount of entries,
- frequent data updates occur, and
- there is a high barrier to the adoption of global PID systems.

For these, **the identifiers.org registry tries to keep the minimum amount of metadata to minimize the work to keep these up to date with repositories**. Also, adding these associations will require our team to dedicate further efforts to keeping these up to date with providers. Such effort could make securing enough financing difficult. Thus, we aim for a solution where data providers are the ones in control of the associated metadata with zero or minimal maintenance from our team.

Our ROR ID association works illustrates this as the small addition of ROR IDs to institutions in our registry adds considerable curation effort to maintain overlapping metadata in sync and to ensure our registry entries have a valid ROR ID when available.

Expected impact of the Use Case

More efficient use of PID metadata will greatly improve interoperability and discoverability, two of the main FAIR principles, by using well-defined and harmonized data. For example, users can use ROR IDs from our registry to link entries with other resources in knowledge graphs that also use ROR IDs. **Co-designing the PID practices through EOSC alignment provides a research environment which is responsive to the needs of the various research communities**. EMBL-EBI is experienced in working with metadata standards, integration, discoverability, and display through its work with the Identifier.org service. **Furthermore, EMBL-EBI can bring its valuable expertise from its partnership with a large international consortium, the European Bioinformatics Institute, which is a public biomolecular data resource provider**. EMBL-EBI also has vast expertise gained from its partnership in many relevant EU projects, among others EOSC-Life, FREYA, and BY-COVID.

Expected outputs

The use case will bring evolving Identifiers.org practices into a broader EOSC context and provide solutions to overcome some challenges related to PID practices. Making:

- Our registry more accessible through the use of semantic artifact catalogs
- Working towards determining minimum metadata for PID providers

AnaEE use case: Semantic interoperability in ecosystem studies

- **Key topic** Interoperability | Metadata & Ontologies
- **Scientific domain** Earth and environmental sciences
- **Leading organisation** INRAE
- **Contributors** Christian Pichot, Philippe Clastre, Brett Choquet

Overview

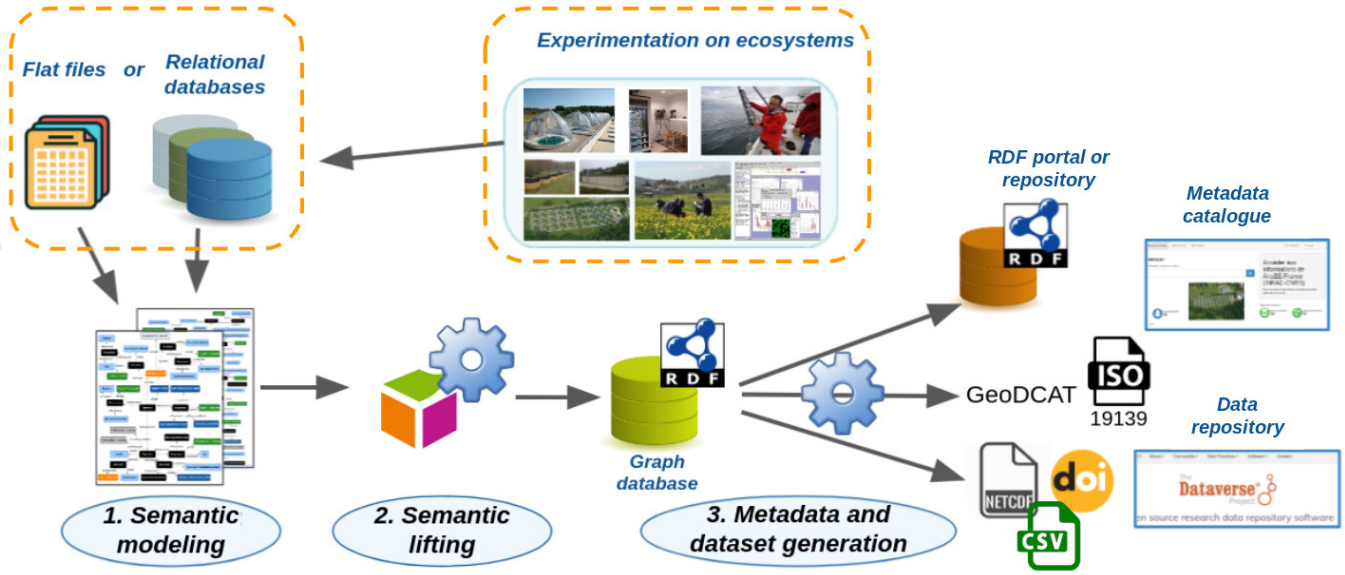
The AnaEE (Analysis and Experimentation on Ecosystems) Research Infrastructure offers experimental facilities for studying ecosystems and biodiversity. A pipeline featuring several applications is developed, based on interoperability of its components and the use of shared semantic artifacts, mainly AnaeeThes thesaurus and OBOE-based AnaeeOnto ontology (to be published). The goal of this pipeline is to generate interoperable datasets and the associated metadata records. The AnaEE semantic workflow consists of 3 steps:

At step 1 the observed /measured variable and the acquisition contexts are modeled as a generic graph based on the ontology.

At step 2 a dedicated pipeline allows the automation of the annotation process and the production of graph-hosted semantized data.

At step 3, a second pipeline is devoted to the exploitation of these semantized data through the generation of standardized metadata records (presently GeoDCAT and ISO) and, data files (presently in NetCDF and csv format) from selected perimeters (acquisition sites, measured variables, experimental factors, years...) (https://hal.inrae.fr/hal-03234155 and figure 1).

The semantized data and metadata produced are intended to feed discovery and access portals, data repositories and (Virtual Research Environment) VRE-type platforms.



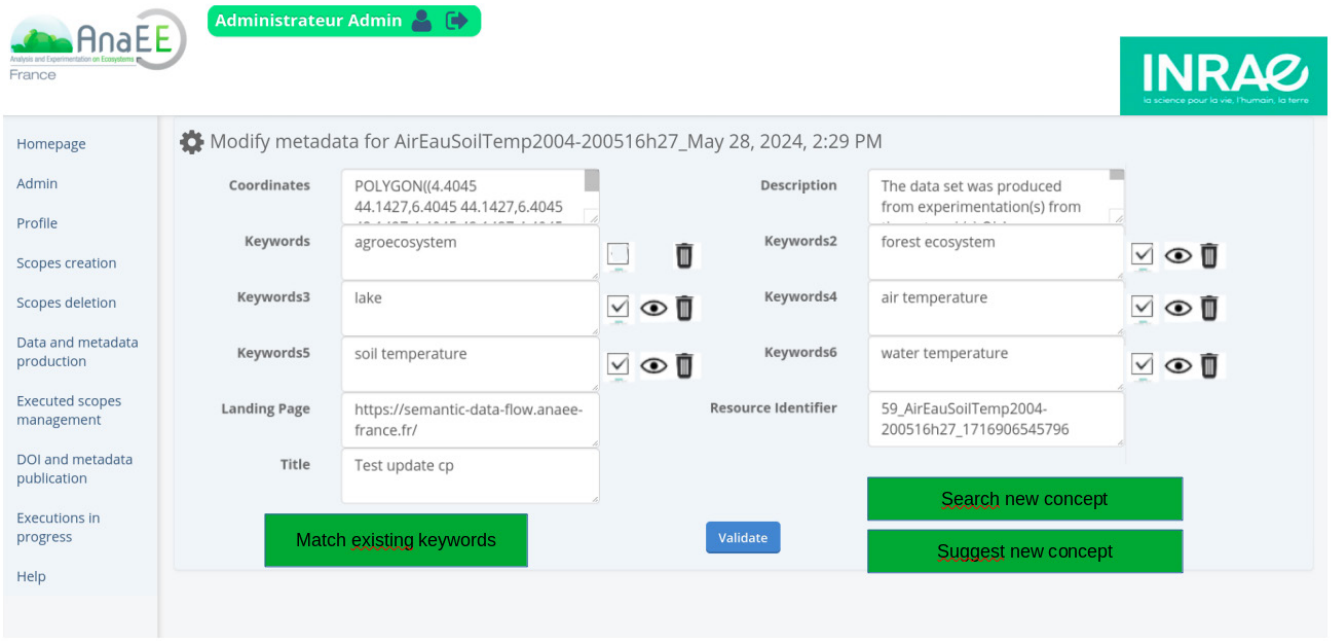
As shown above, the generated datasets and metadata are pushed to the Recherche Data Gouv platform (based on Dataverse). For that reason, AnaEE metadata enriched with controlled keywords must remain fully compatible with the Dataverse metadata standard.

Semantic artefacts (SA) in use within the AnaEE semantic tools and data repositories

In the context of FAIR IMPACT, we focus on the use of SA at the 3rd step of the workflow, i.e for metadata generation. However, it should be noted that SA are already involved in the modeling phase (step 1). Referred to as SemData, the application aims to enrich the description of the dataset by the addition of 'keywords' supplied by controlled SA in order to complete the description of the data asset provided by the modeling step. Three types of enrichment processes are proposed to the data providers :

1. Keyword automatic checking & enrichment

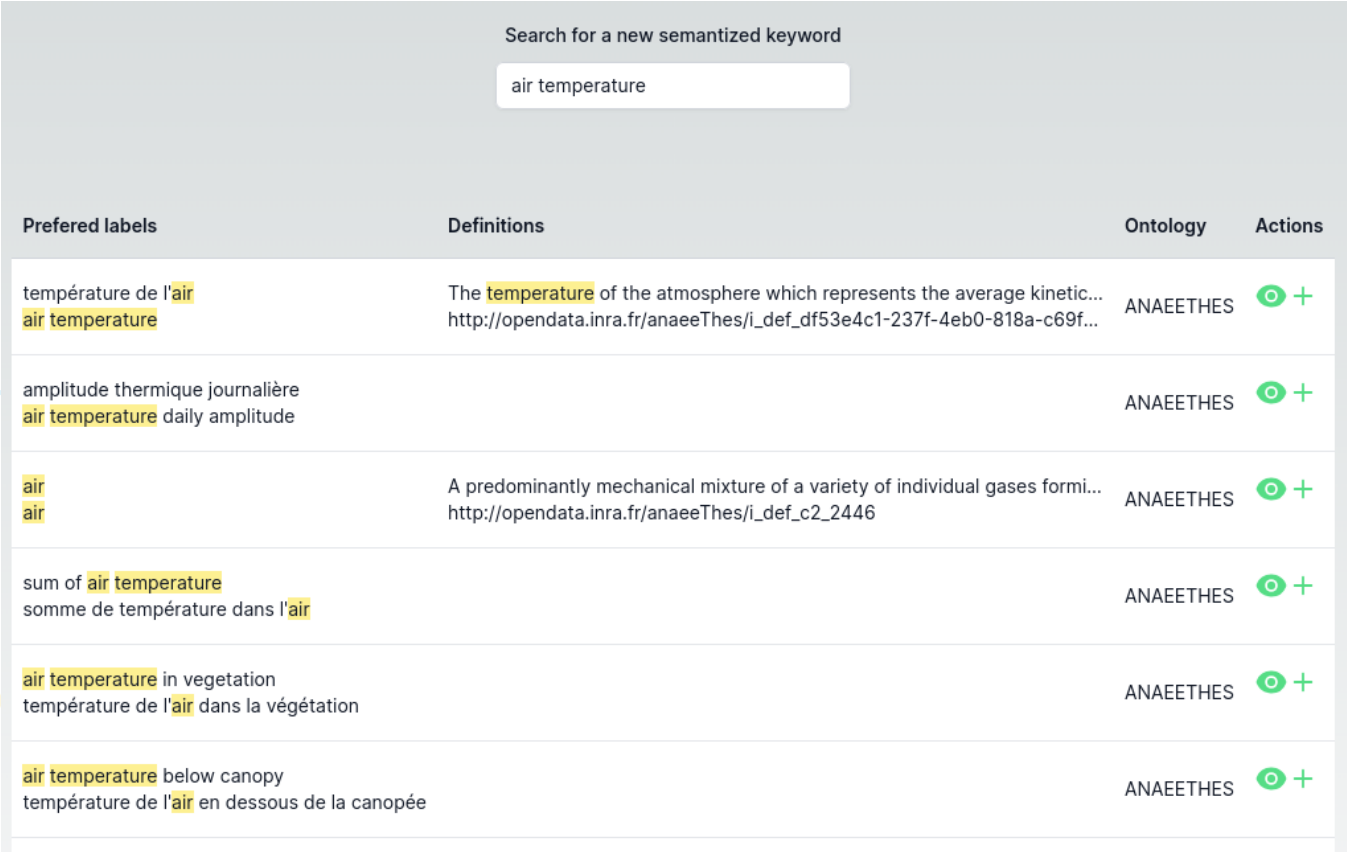
In this situation, some keywords resulting from the pipeline are simple strings, without semantic reference. The application uses the Agroportal search API to search for the matching concept(s) in the AnaEE Thesaurus. Then, as shown in figure 2, the user is able to access keyword full description, and for keywords that do not match (unchecked box), decide to get rid of them (trash icon). Meaningful icons next to each keyword text box allows access to the full keyword description as provided by Agroportal (definition, synonyms, URI…) and possibly delete the keyword.



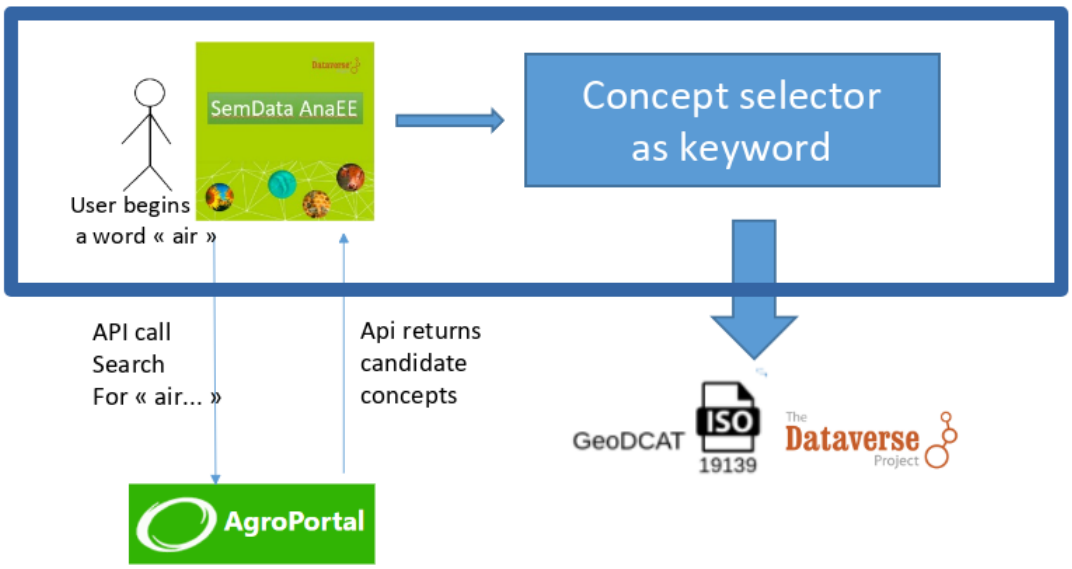
2. Keyword manual selection

In this second situation, the data producer wants to enrich the metadata record with additional keywords. The interface of the AnaEE dataset and metadata generation tool must allow to pick up concepts from thematic vocabularies, especially from AnaeeThes, to fill in the "keywords" metadata elements.

The user starts writing a keyword and the agroportal API call proposes several candidate keywords. One or more of the proposed keywords is/are then selected. These keywords can be added to the metadata records or ignored (Figure 3).



Here, the user entered “air temperature” and the search function of Agroportal returns some matching keywords. The UI displays them and allows the user to add one or more keywords to the metadata records.



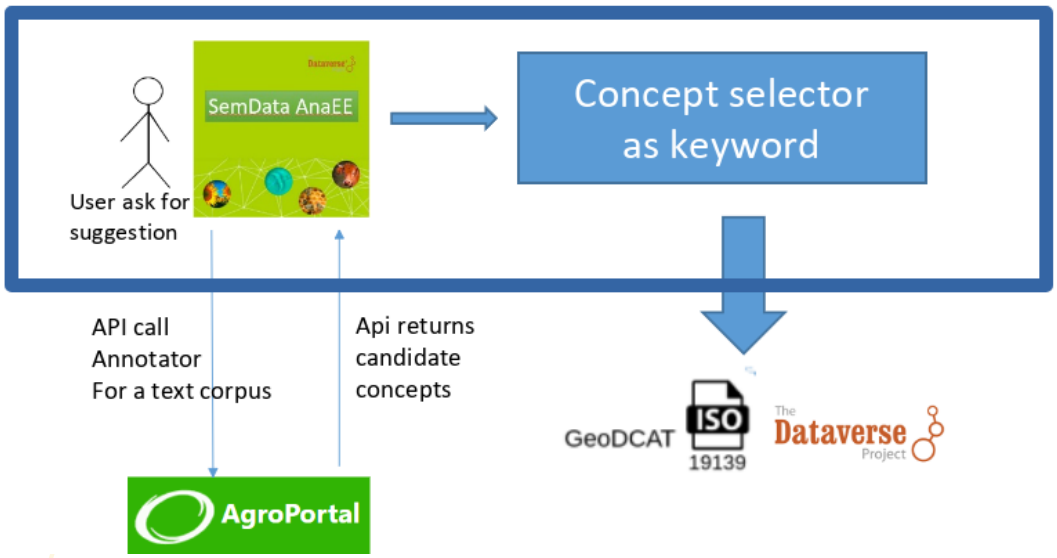
AnaeeThes is used as the default SA, but the UI allows the use of other controlled vocabularies.

3. Keywords suggestion from textual description

Here, the objective is to offer the possibility to enrich the metadata records with keywords by matching a SA on a narrative description of dataset content. This description is automatically generated by SemData from information collected by the pipeline. As an abstract-like description of the dataset, it provides information about sites, variables and time periods selected by the user as criteria generating the dataset (i.e : ‘The data

set was produced from observations collected on the site(s) {sites} in the ecosystem(s) {ecosystems}. Measurements are about the following variables: {variables}... '.

The user can press the "suggest new concept" button (see Figure 2). The abstract has to be pasted in a combo box. The user may also complete (or replace) this default text with some other meaningful narrative. Then the user launches an "annotator" call to the Agroportal API on the description text, which returns candidate concepts.



For sub use cases 2 & 3, the "Concept Selector" is designed to facilitate the choice for the user. We focus on displaying fully qualified information. The preferred term and concept definition are needed. The user can also open a new web browser tab to consult the concept landing page.

In the application, each keyword is stored with its preferred term, its URI, the name and URI of the SA it comes from..., elements that cover the information expected by Dataverse instances like Recherche Data Gouv, for qualifying keywords.

Challenges that need to be addressed

In order to implement all the functionalities described within the AnaEE use case, we inserted Agroportal API calls in the AnaEE application. We also changed our data model inside the application, to store all keyword related information. We also adapted API calls to the dataverse repository, to include keywords defined by several parameters: concept, URI, SA URL, SA name.

Expected impact of the Use Case

After the integration of the API calls to Agroportal, the keywords associated by default in each dataset are fully described and above all compatible with the description of the keywords in the dataverse repository of recherche.data.gouv. The user also benefits from a mechanism for suggesting new keywords, based on the description of the dataset. For datasets already published, there will be no impact of this development, unless they are regenerated through the adhoc functionality offered by the AnaEE pipeline.

Expected outputs

Production of semantically enriched, vocabulary based, description of the datasets. Keywords need to be fully compatible with the dataverse keyword data model. As a consequence, our thesaurus updating process needs to be more accurate with the changes involved by adding new variables, when new information systems benefit from the semantic annotation and data publication services offered by the pipeline.

FAIR vocabularies in DANS Data Stations

- **Key topic** Interoperability | Metadata & Ontologies | Metrics, Certification and Guidelines | PIDs
- **Scientific domain** Life science | Social Sciences and Humanities | Physical and Technical Sciences | Archaeology
- **Contributors** Slava - Vyacheslav Tykhonov, Pascal Flohr, Andrea Scharnhorst

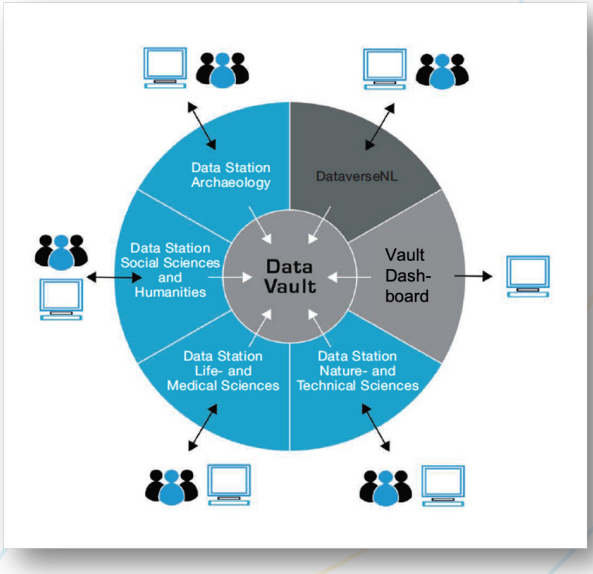
Short Use Case overview

DANS (Data Archiving and Networked Services) improved the FAIRness (Findable, Accessible, Interoperable, and Reusable) of its repository service by transitioning from a generic repository system, EASY, to four discipline-specific repositories called "Data Stations." Each Data Station is curated with relevant communities, enables the addition of custom metadata fields and discipline specific controlled vocabularies, improving metadata quality and interoperability. Data is mapped to and can be exported in multiple formats like DublinCore, DataCite, and Schema.org. The new Data Stations use Dataverse software, as opposed to EASY, which was based on the FEDORA system and became outdated.

Use case description

Use case: DANS Data Stations

DANS (Data Archiving and Networked Services) is the Dutch national expertise centre and repository for research data. Its main objective is to facilitate and improve the long-term storing and sharing of research datasets, thus aiding its mission of enhancing the reusability of research data and the quality of scientific research. DANS is running various Dataverse instances for different disciplines paired with a long-term preservation repository, and calls them Data Stations. DANS' repository services currently include four discipline-specific data stations; a research data repository service for institutions, DataverseNL; and a preservation repository, the Data Vault. DANS also offers expertise and training on topics like Research Data Management, FAIR data, and Open Science through its Research Data Management Expert and Training teams.



Context and objectives

To improve the FAIRness of its repository service, DANS moved from one generic repository, Electronic Archiving SYstem or simply EASY, to four discipline-specific 'Data Stations': Archaeology (https://archaeology.datastations.nl, launched in 2022), Social Sciences and Humanities (https://ssh.datastations.nl, SSH, launched in the spring of 2023), Life Sciences (https://lifesciences.datastations.nl, soft launch in December 2023), and Physical and Technical Sciences (https://phys-techsciences.datastations.nl, soft launch in December 2023). For each of the Data Stations there are also special links to the communities. For instance the Physical and Technical Sciences Data Station is curated together with the Technical University of Delft. The Archeological Data Station serves as a reference archive for all archeological reports. This role is commissioned by the RCE ("Rijksdienst voor Cultureel Erfgoed", the Dutch Cultural Heritage Agency).

The implementation of discipline-specific archives has facilitated the addition of discipline-specific metadata with their own specific controlled vocabularies, in addition to the already present rich generic metadata. Metadata has been mapped to, and can be exported as, DublinCore, DataCite, Schema.org, OpenAIRE, DDI, and more. EASY was developed by DANS team internally based on the FEDORA system, however the Data Stations rely on the Dataverse software (https://dataverse.org/).

For several generic metadata fields and for many of the discipline-specific metadata fields, the depositor chooses values from a dropdown menu or automatic completion field. The lists, or semantic artefacts, underlying these values are mostly hard-encoded in the DANS Dataverse instance(s), and using Dublin Core and DataCite 3.0 and 4.0 properties (Table 1). For language and spatial coverage ISO standard lists are used, ISO 639-2:1998 and ISO 3166-1:2013, respectively. Also for the (more recently added) discipline-specific metadata blocks, the values are derived from internationally used vocabularies, greatly improving interoperability. For the Data Station SSH this concerns terms from the CESSDA ELSST Thesaurus and the CESSDA Topic Classification; addition of DDI terms is planned for the near future. For the Data Station Archaeology this concerns terms from the 'Archeologisch Basis Register' (ABR) of the Netherlands (https://data.cultureelerfgoed.nl/term/id/abr.html).

The process of adding domain-specific metadata and vocabularies is still ongoing for the Data Station Life Sciences and the Data Station Physical and Technical Sciences in collaboration with relevant partner organisations to assess community needs.

Table 1

Data Station	Metadata element label	Source	Source URL
All	Indentifier Type	List created by DANS based on identifiers known to be used by the DANS community	n/a
All	Subject	DANS, based on Dataverse list	n/a
All	Language	ISO 639-2:1998	https://www.loc.gov/standards/iso639-2
All	Spatial Coverage	ISO 3166-1:2013	https://www.iso.org/obp/
All	Contributor Type	A mix of terms from DataCite ContributorType version 3.0 and 4.0	https://schema.datacite.org/meta/kernel-3.0/; https://schema.datacite.org/meta/kernel-4.0/
All	Audience	NARCIS	https://vocabs.datastations.nl/NARCIS/en
All	Collection	DANS collections	https://vocabs.datastations.nl/DansCollections/en/
All	Relation, Related Material Type	Elements from DataCite and DublinCore. A mix of elements from Dublin Core properties and DataCite 3.0 and 4.0	https://www.dublincore.org/specifications/dublin-core/dces/; https://schema.datacite.org/meta/kernel-3.0/; https://schema.datacite.org/meta/kernel-4.0/
All	Personal Data in Dataset	List created by DANS	n/a
SSH	Keyword ELSST	ELSST	https://thesauri.cessda.eu/elsst-4/en/

Data Station	Metadata element label	Source	Source URL
SSH	Topic Classification CESSDA	CESSDA Vocabulary Service	https://vocabularies.cessda.eu/vocabulary/TopicClassification
SSH	[Nothing yet, planned use]	DDI Analysis Unit, Mode of Collection; Sampling Procedure; Time Method; Type of Instrument	https://ddialliance.org/controlled-vocabularies (links to CESSDA Vocabulary Service)
ARCH	Archaelogy report	ABR+	https://vocabs.datastations.nl/ABR/en/
ARCH	Relation metadata	NARCIS	https://vocabs.datastations.nl/NARCIS/en
ARCH	Methods of recovery	ABR+	https://vocabs.datastations.nl/ABR/en/

The third-party vocabularies are available through a semantic artefact catalogue relying on the SKOSMOS, web-based tool providing services for accessing controlled vocabularies: https://vocabs.datastations.nl/en/. The vocabularies to be used in the DANS Data Stations are fed into here. The Data Station is then connected to the SKOSMOS instance to fill out the dropdown lists of the metadata elements, providing an easy way for users to select values from the third-party vocabularies. Custom settings required to connect vocabularies available in SKOSMOS with a plugin implemented as an external Javascript application, available through https://github.com/gdcc/dataverse-external-vocab-support.

Technologically, by using the Dataverse archival software, DANS became an active participant in the Open Source community around Dataverse. Led by Harvard University, regular updates of the so-called 'master/main' branch are developed with new features. For example, a Controlled Vocabularies plugin was developed by the DANS team in the SSHOC project and was incorporated to Dataverse by Harvard and GDCC, and since then has become part of the out-of-the-box distribution for the community. The institutions hosting Dataverse instances are free to implement new versions into their local systems, depending on their own requirements. However, being part of the Dataverse community also enables DANS to explore new innovative solutions in projects, based on experiences gathered from the day-to-day use and curation of Data Stations. When it comes to challenges and implemented solutions we will indicate which of them has been already implemented in the production system, which of them have been taken up in the Dataverse main releases (but not yet implemented in the DANS instance), and which of them are explorations even new for the Dataverse main release discussion. Within T4.2 we are working with other use cases (e.g., the OntoPortal/Dataverse connector) to reach out the same maturity by learning from our experience and making the connector part of the default Dataverse distribution.

Challenges and solutions implemented

Challenges (identified, not all to be addressed):

Data Station Archaeology terms such as keywords are available in Dutch taxonomy only and filled during depositing in the DANS Data Station. Available for searching in English in the ARIADNE portal.

- ☐ Objective: Look into augmenting the vocabularies with an English translation. ODISSEI semantic enrichment workflow can be used to get multilingual translations available in Data Stations.

Not all depositors use the available metadata fields.

- ☐ Solutions to link external controlled vocabularies to the DANS Data Stations have been implemented in the DANS production system and in Dataverse main releases, but their uptake by depositors remains a little limited.

- Objective: Increase the use especially of rich metadata, especially also those fields that use international vocabularies.
- Solution: Write a clear guide on why and how to use the metadata fields and their controlled vocabularies.

Currently the keywords are free-text.

- Objective: Connect them to internationally recognised controlled vocabularies / thesauri.
- Solution: The technical solution already exists.

Using a SKOSMOS instance has clearly been a useful solution of connecting published vocabularies to Dataverse instances. However, while it has been documented on GitHub, reusability of the solution would be greatly improved by additional documentation.

- Solution: document the whole approach to make it reusable.

Vocabulary sustainability: how are updates to the original vocabularies dealt with? For example, ELSST Thesaurus in the DANS SKOSMOS instance has version 3 but CESSDA already updated it with version 4.

- Objective: track changes and synchronise the same vocabularies hosted by different parties
- Solution: requires development of some workflow to archive new versions of controlled vocabularies before publishing in SKOSMOS or OntoPortal, and keep provenance.

Vocabularies available in OntoPortal instances, currently especially relevant for the Data Stations Life Sciences and Physical and Technical Sciences. How can these be directly connected to Dataverse instances like the DANS Data Stations?

- Objective & solution: develop a connection between OntoPortal and Dataverse in partnering with the INRAE's use case (RDG/OntoPortal) in the same task (T4.5)

This solution is new even for the Dataverse main release discussion.

Proposed solutions to work on as part of FAIR-IMPACT T4.5:

1. Documenting the approach

- How to transform the free text terms by reference to controlled vocabulary terms?
- How to connect several controlled vocabularies to a single data repo/metadata element? (Could there be a way to have AAT AND ABR?)

2. Documenting the use of vocabularies in the Data Stations. How can we take user requirements about vocabularies integration into account?

- How are the semantic terms used in the repository index, how are they exported in the various export flavours or served by APIs?
- How are the terms used by systems that harvest the data stations and mapping technique?
- How is the harvesting done from the DS Archaeology by ARIADNE (for example) where mapping of ABR to AAT should be implemented by ARIADNE?
- Will DS SSH be harvested by CESSDA -> EOSC in future? Are the semantics then retained?
- Comparison between before and after the vocabularies were connected. Are people increasingly using these vocabularies? Is searching easier?
- Connecting OntoPortal vocabularies to DANS Data Stations?
- Practice report, for the use case / question 'How are the DANS Data Stations using external specific vocabularies? Practice report, with pros/cons, lessons learnt, things to be aware of technically and beyond.
- Documentation SKOSMOS – Dataverse available here: https://zenodo.org/records/8133723
- OntoPortal – Dataverse connection

Expected/Measured Impacts

- ☐ Use / increase in use of SKOSMOS – Dataverse connector
  - ☐ Compare use before FAIR-IMPACT project, with after Slava started in FAIR-IMPACT (after initial presentations, after more presentations, after document published, etc.). Is there an increased use?
  - ☐ use of the information by others, e.g. already use in OntoPortal-DV connection: https://github.com/IQSS/dataverse/pull/10145
- ☐ Increased use of metadata elements with controlled vocabularies by DANS Data Station depositors.
  - ☐ e.g. use before the metadata with controlled vocabs like ELSST and CESSDA were present and after one year; and/or compare now with one year from now (pre-guide and post-guide / or pre and post increased guidance in general).
  - ☐ Connection possibility of OntoPortal instances with Dataverse instances. Ideally also take up and use of this, e.g. connection of an OntoPortal instance with the new DANS Data Stations.

Reference Materials

- ☐ DANS Website
- ☐ DANS Data Stations guideline
- ☐ Dataverse support for external vocabulary services
- ☐ DANS Skosmos

Semantic Artefacts Alignment for Improving Interoperability in Astronomy

- **Key topic** Interoperability | Metadata & Ontologies
- **Scientific domain** Physical and Technical Sciences
- **Leading organisation** INRAE | Observatoire de Paris
- **Contributors** Baptiste Cecconi - Observatoire de Paris, Laura Debisschop - Observatoire de Paris, Sophie Aubine - INRAE

Short Use Case overview

The astronomy community is structured in several sub-communities, with matured but siloed semantic artefact ecosystems. This use case brings all semantic artefacts in the same catalogue. The goal is to improve the semantic interoperability between the astronomy communities, and in turn the semantic artefact FAIRness.

Use case description

The astronomy community is composed of three main semantics sub-communities:

- **celestial astronomy** (objects are referenced to with their sky coordinates, e.g., stars, galaxies, etc);
- **planetary sciences** (the study of the Solar System objects, e.g., planets, comets, asteroids, etc);
- **heliophysics** (the study of the Sun, the plasma environments throughout the Solar System).

Each of these sub-communities have developed interoperability and semantic ecosystems, which are rather siloed up to now.

Context and objectives

The sky astronomy community is organized **around the IVOA** (International Virtual Observatory Alliance, https://ivoa.net), which is maintaining an operational interoperability framework used by data repositories and science application platforms throughout the world. In this community, semantic artefacts (SA) are composed of terms (vocabularies) and schemas (data models). **The Semantics Working Group of the IVOA is managing the vocabularies used in the IVOA standards**. The vocabularies are available from a dedicated web page (https://ivoa.net/rdf) and are accessible using IVOA or RDF tooling.

The planetary science community is less organized than the sky astronomy one. Two main frameworks co-exist with different scopes: the IPDA (International Planetary Data Alliance, https://planetarydata.org), which proposes an advanced data archiving information model for planetary exploration datasets; and the OGC (Open Geospatial Consortium, https://www.ogc.org/), which is used by the teams studying the planetary surfaces.

The heliophysics community is organized around the IHDEA (International Heliophysics Data Environment Alliance, https://ihdea.net), which is proposing a set of tools and standards for finding and accessing datasets in this domain. Semantic artefacts in this community were historically of two kinds:

- **the SPASE** (Space Physics Archive Search and Extract, https://spase-group.org) Ontology[1] (XML schema) includes list of terms, properties and classes for defining various objects (Persons, Observatories, Instruments, Datasets, Repositories, etc);
- **SOLARNET** set of keywords[2] (dedicated to Solar observations).

1 https://spase-group.org/data/schema/index.html
2 https://zenodo.org/records/5719255

The IVOA, IPDA and IHDEA alliances are all worldwide working groups, consensus and bottom-up driven, and based on best effort contributions. Interdisciplinary links between these communities have been developed thanks to the Europlanet/VESPA (http://www.europlanet-vespa.eu/) project, focusing on discoverability and implementation of plugins to extend the capabilities of existing tools. The semantic interoperability across the sub-communities approach started only recently, with the ongoing development of two common semantic artefacts: a vocabulary for "observation facilities" and another on for "reference frames".

The objective of this use case is to enable semantic interoperability between sub-communities of astronomy in a first step, and explore interoperability with neighboring fields such as the Earth and environmental sciences, or particle physics.

Challenges and solutions implemented

Within FAIR-IMPACT Task 4.2, an OntoPortal instance has been set up with the goal of gathering semantic artefacts from the various astronomy sub-communities in the same place. The astronomy ontology portal is now available, and a series of relevant semantic artefacts have been ingested therein (39 SA, at the time of writing), covering sky astronomy and heliophysics.

The main challenge of this use case is to produce semantic artefacts in the form (RDF, OWL or SKOS) that can be ingested into the OntoPortal instance. Most of the current semantic artefacts (except those from the sky astronomy) are in diverse forms, from lists of terms in XML schemas, to unformatted lists of metadata in specification documents.

This lifting shall be done by the semantic working groups or authorities of the relevant communities (e.g., the IVOA Semantics WG or the IHDEA dedicated teams), with the support of the ObsParis FAIR-IMPACT team. This interaction has started as shown in the few examples below.

In the IVOA context, **all the semantic artefacts (list of terms) were available on a web page, and an RDF version of each SA was available on the respective landing pages**. The IVOA vocabularies are managed according to an IVOA recommendation[3] defining rules and conventions, and specifically how should be designed the RDF version of the IVOA SA, with a limited subset of SKOS and OWL properties. This design decision is trying to limit the external dependencies and ensure the sustainability of the IVOA infrastructure.

The semantic artefact management in the IVOA relies on VEP[4] (Vocabulary Enhancement Proposal), which is a process to propose, update and deprecate a term. The VEPs process implies a consensus-based decision, after a community discussion.

In the IHDEA context, a general rehauling of the semantic artefacts has been initiated since the 2023 IHDEA meeting. **The previous state was based on the SPASE information model, serialized solely by an XML schema**. The lists of allowed values are embedded in the SPASE schema, requiring frequent release of new versions (e.g., for each new item in a list of allowed values). It became clear that many lists of terms needed to be updated, and that convergence with the IVOA semantic artefacts was desirable. The first work on a joint semantic artefact concerns a vocabulary for "solar system reference frames", which will be merged with the IVOA RefFrame vocabulary[5].

The goal of the FAIR-IMPACT use case with the IVOA and IHDEA communities is to enhance the semantic artefact quality, especially on the interoperability. Part of the planned outcome is to update the semantic artefact management practices (e.g., new version of the Vocabulary in the VO recommendation).

3 https://ivoa.net/documents/Vocabularies
4 https://wiki.ivoa.net/twiki/bin/view/IVOA/VEPs
5 http://www.ivoa.net/rdf/refframe

Expected/Measured Impacts

From the point of view of the interoperability alliances, the setting up of the OntoPortal instance and the assessment / preparation of the semantic artefacts have been a quantitative and qualitative improvement.

- **IVOA context:** As an example, a revision "Vocabularies in the VO" document (https://ivoa.net/documents/Vocabularies) is being prepared to include a set of terms required for fixing SKOS-based semantic artefact catalogues in the IVOA. This work also started an ongoing fondamental discussion of reusing external semantic artefacts, versus keeping things "simple" (but disconnected) by reducing external semantic dependencies.
- **IHDEA context:** Several teams have started working on producing linked data and metadata using the RDF tooling. The OntoPortal instance for astronomy has been an important incentive for adoption of this framework.

From a user perspective, the enhanced FAIRness of the semantic artefacts will enhance the FAIRness of published datasets. The semantics artefacts can be plugged into smart DMP tooling (URIs for terms, rather than free text) or search interfaces, and in turn allow more refined queries and selections on data discovery interfaces. In this respect, The OSTrails (https://ostrails.eu) project development for the astronomy thematic pilot will be built on the semantic artefact catalogue tooling developed thanks to FAIR-IMPACT.

Enabling interoperability between AgroPortal and PHIS information system data repository for enhanced phenomics data annotation and exchange

- **Key topic** Interoperability | Metadata & Ontologies
- **Scientific domain** Life science
- **Leading organisation** INRAE
- **Contributors** Anne Tireau - INRAE, Arnaud Charleroy - INRAE, Llorenc Cabrera-Bosquet - INRAE, Clement Jonquet - INRAE

Overview

The goal of this use case is to illustrate the benefit of using AgroPortal ontologies **to describe, annotate and structure phenomics data within the PHIS platform**, an open source information system for Plant Phenomics. We like: (i) to ease the reuse of semantic artefact objects (classes, concepts, properties, etc.) within PHIS to describe data and (ii) enable the push back of knowledge objects created by domain scientists within PHIS to application or domain ontologies hosted in AgroPortal.

Context and materials

In recent years, **plant phenomics** –i.e., the discipline of biology related to the measurement and analysis of the observable physical and biochemical characteristics of plants as they interact with their environment– has generated vast amounts of datasets from experiments conducted in both field and controlled conditions, encompassing hundreds of genotypes across various scales of organization. These datasets represent unprecedented resources for identifying and testing novel mechanisms and models [https://doi.org/10.1016/j.cub.2017.05.055].

However, **assembling and organizing such datasets is challenging due to the heterogeneous nature of the data** (e.g., environmental data, phenotypic variables, images, and metadata) and the difficulty in accessing information distributed across multiple sources.

To address these challenges, the phenomics community has proposed **an ontology-driven information system**, **called PHIS (Phenotyping Hybrid Information System)**, inspired from the FAIR principles. PHIS serves as a solution for integrating, organizing, and managing multi-source and multi-scale phenomics data obtained from field and greenhouse conditions [https://doi.org/10.1111/nph.15385].

PHIS is based on the **generic OpenSILEX technology, developed and maintained by INRAE-MISTEA.** It is an ontology-driven open source information system tool designed for life science data. The software suite implements original management methods for the exploitation of semantics, production of FAIR data and adopts an architecture adapted to heterogeneity and increasing volumes of data. PHIS is a specific instance of the OpenSILEX technology for plant phenotyping deployed in various categories of installations (field, glasshouse) developed in part within the H2020 EPPN2020, EMPHASIS ESFRI and national PHENOME infrastructure projects.

One of the major obstacles for data interoperability and reusability is the accurate identification and definition of measured variables (see the work of the I-ADOPT RDA working group). For instance, the commonly measured variable "plant height" can have various definitions depending on the crop, can be measured using different methods (e.g., image analysis or a ruler), and can be expressed in different units (e.g., cm or mm). To address

this challenge, the Entity-Characteristic-Method-Unit model (https://cropontology.org/) has been adopted to facilitate the standardization of measured variables (as illustrated in Fig. 1):

- **Entity**: Refers to the object being targeted (e.g., plant, canopy, air, leaf).
- **Characteristic**: Denotes the type of measurement, encompassing physical quantities as well as observed qualities like irradiance, temperature, area, height, etc.
- **Method**: Describes the approach used for estimating the variable (e.g. manual measurement, image analysis, visual score).
- **Unit**: Describes the units used to quantify the variable (e.g., g, kg, W/m², unitless).

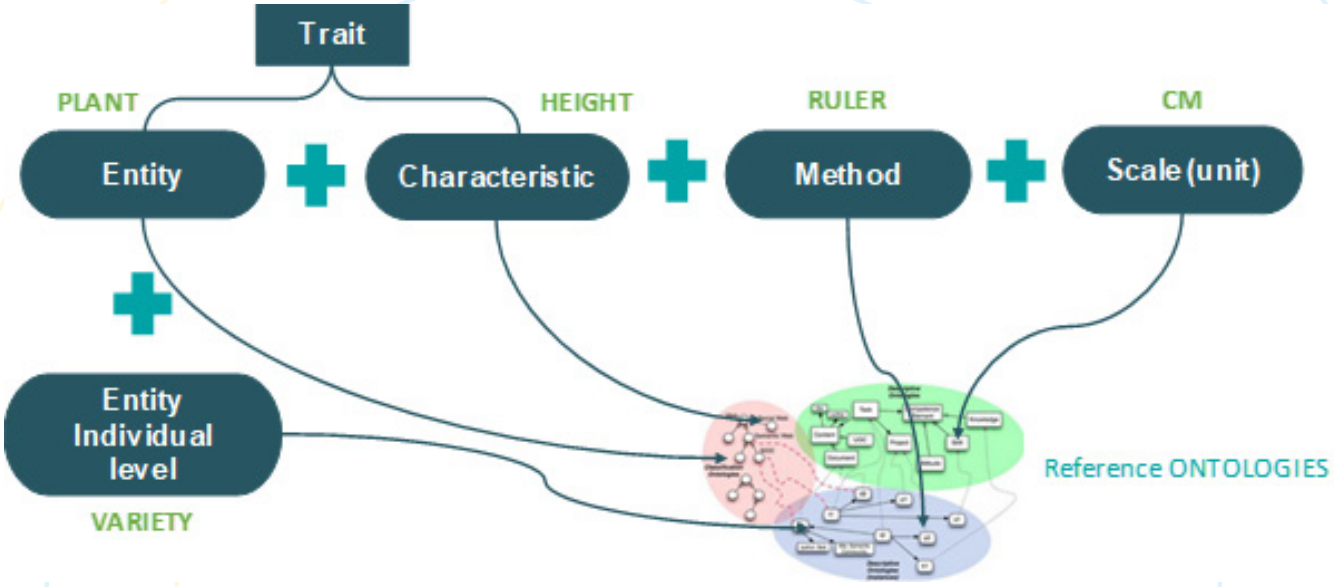


Figure 1. Entity-Characteristic-Method-Unit model used in OpenSILEX/PHIS to represent measured variables

Within PHIS, each of these building blocks of a measured variable are mapped as much as possible to reference ontologies such as the Plant Ontology and the Crop Ontology or the SOSA ontology for sensors.

As an example, the air temperature is modeled according to this scheme as:

Air_Temperature_ShelterInstantMeasurement_DegreeCelsius

Where, the Entity is the Air, the Characteristic is the Temperature, the Method used is an instantaneous measurement using a shelter and the units are in °C.

For this variable in PHIS, the different components are mapped (when possible) to existing reference ontologies (here fetched from AgroPortal) and unique internal URIs generated by PHIS are also associated with the variable and to each of these components. In this given case, no reference term was found for the method 'ShelterInstantMeasurement,' thus the system generates a new term and associates it with an internal URI. Eventually this term will be a candidate for extending an ontology. While the former example demonstrates the flexibility and freedom of PHIS to create new terms when users are unable to find them for any reason (such as a lack of computer science skills, time constraints, or familiarity with existing repositories and resources), this approach does not promote the reuse of existing terms and limits interoperability with other resources.

- Variable URI => phis:id/variable/ev000001
- Entity (Air) => http://aims.fao.org/aos/agrovoc/c_224 (see in AgroPortal)
- Characteristic (temperature) => http://purl.obolibrary.org/obo/PATO_0000146 (see in AgroPortal)
- Method (ShelterInstantMeasurement) => phis:id/variable/method.shelter_instant
- Unit (degree Celsius) => http://purl.obolibrary.org/obo/UO_0000027 (see in AgroPortal)

This structured approach allows for the creation of new variables by combining these building blocks, for instance, by changing the method or the unit.

To get the relevant ontology terms for the variable, PHIS users are encouraged to use AgroPortal. It is a vocabulary and **ontology repository built as a reference catalogue for hosting, sharing and serving semantic artefacts for agri-food** communities, developed and maintained by INRAE-MISTEA and University of Montpellier [https://doi.org/10.1016/j.compag.2017.10.012]. AgroPortal is based on the generic technology OntoPortal developed jointly by the OntoPortal Alliance. AgroPortal allows users to search and browse for terms in a user-friendly interface (see Fig. 2). The semantic artefact catalogue can be called automatically by tools thru its API.

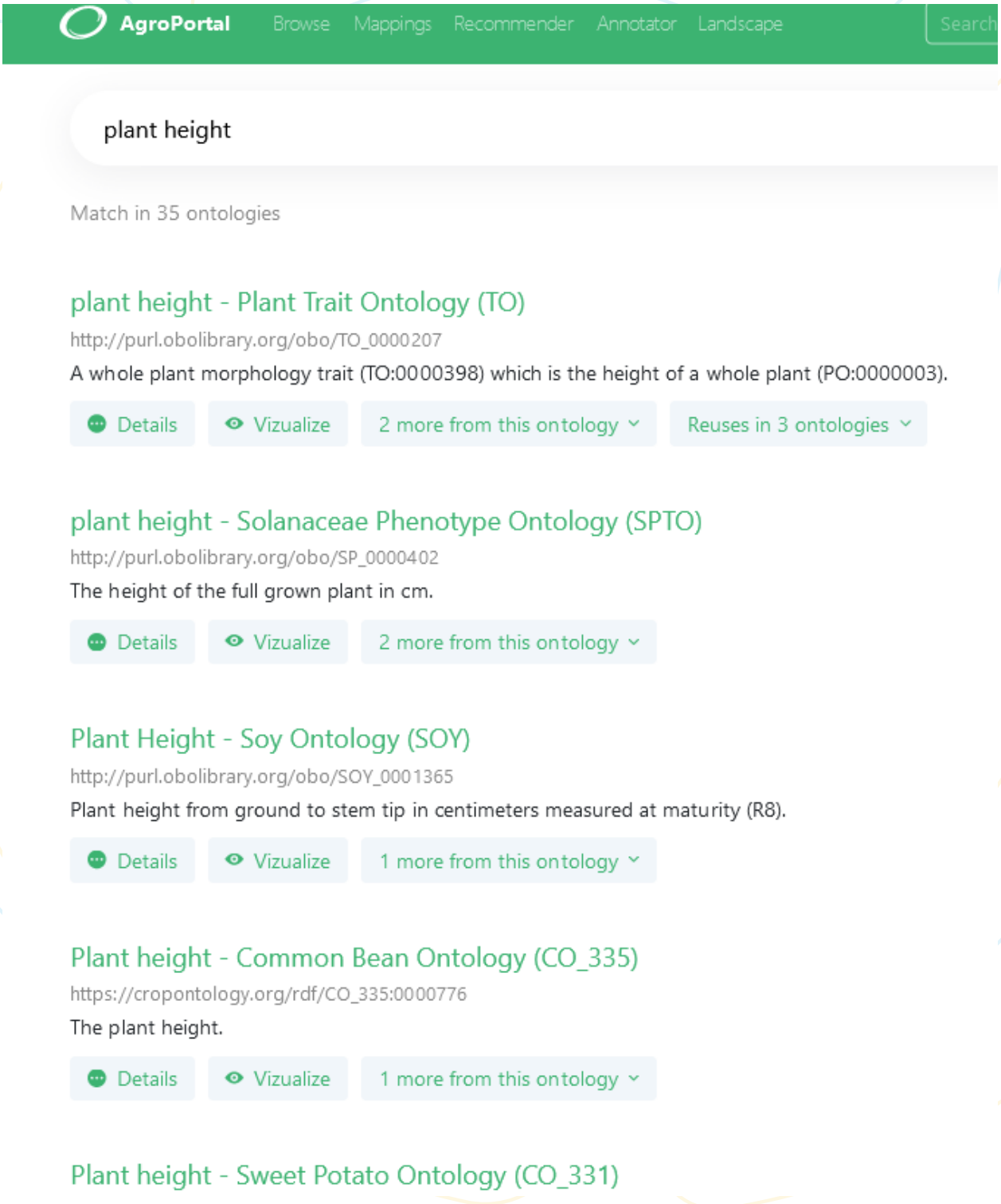


Figure 2. AgroPortal Search interface

Challenges and objective

While the plant phenomics community has embraced ontologies to standardize the description of experimental variables, many users with a background in biology lack computer science skills and are unfamiliar with the use of ontologies or semantic artefacts, leading to **difficulties in retrieving information** (as illustrated in the example above). Additionally, most available resources are not centralized, which further complicates the process of gathering information from multiple sources and mapping concepts.

Currently, for PHIS users, fetching ontology terms from AgroPortal or directly from multiple sources and ad-hoc vocabulary systems is a manual process. This process requires going to another web application or tool, performing a search and manually copy/pasting the information found (if any) related to the selected ontology term. This information is then used to fill in the necessary field for the mappings by specifying the mapping (more specific, more general). This manual process considerably prevents and slows down the reuse of standard ontology terms when describing objects within PHIS.

The goal of our use case in FAIR-IMPACT T4.5 is **to build a connector between PHIS and AgroPortal to ease the re-use of ontology terms when building variables and other scientific objects** within PHIS.

Prototype connector between PHIS and AgroPortal

We (INRAE-MISTEA and INRAE-LEPSE) are working on a prototype (currently developed within the generic OpenSILEX technology and later moved to the PHIS instance) so **PHIS users can easily describe, through Web interfaces**, their measures, observations as scientific variables –using the model presented above– as well as their experimental vocabulary. The connector allows users to **search and grab from AgroPortal either a term URI and its related information** (name, synonyms, definition) or create a new term within PHISand describe it with information and mappings coming from AgroPortal. All the **descriptions are stored in PHIS in RDF**, the pivotal language of semantic knowledge graphs. Mapping between variable description and other domains ontologies are made by users with SKOS mapping relations (e.g., skos:exactMatch, skos:broadMatch).

The prototype connector under development (Fig. 3) aims to address this challenge by providing a semi-automatic ontology term fetching tool embedded within PHIS. The connector is executed within an OpenSILEX instance (here PHIS) and relies on AgroPortal API to get the information. The connector will offer a number of features to make the fetching process easier and more efficient, including:

- An **ergonomic search interface** that allows users to easily find the terms they are looking for.
- Integration with **pre-selected semantic artefacts** made available by AgroPortal, such as the AGROVOC thesaurus, the Crop Ontology or multiple references ontologies for plant sciences.
- A **mapping functionality** that allows users to specify the links between terms from different ontologies especially in the case a new term is created in PHIS and AgroPortal is only used to grab information and mappings to other close terms (Fig. 4).

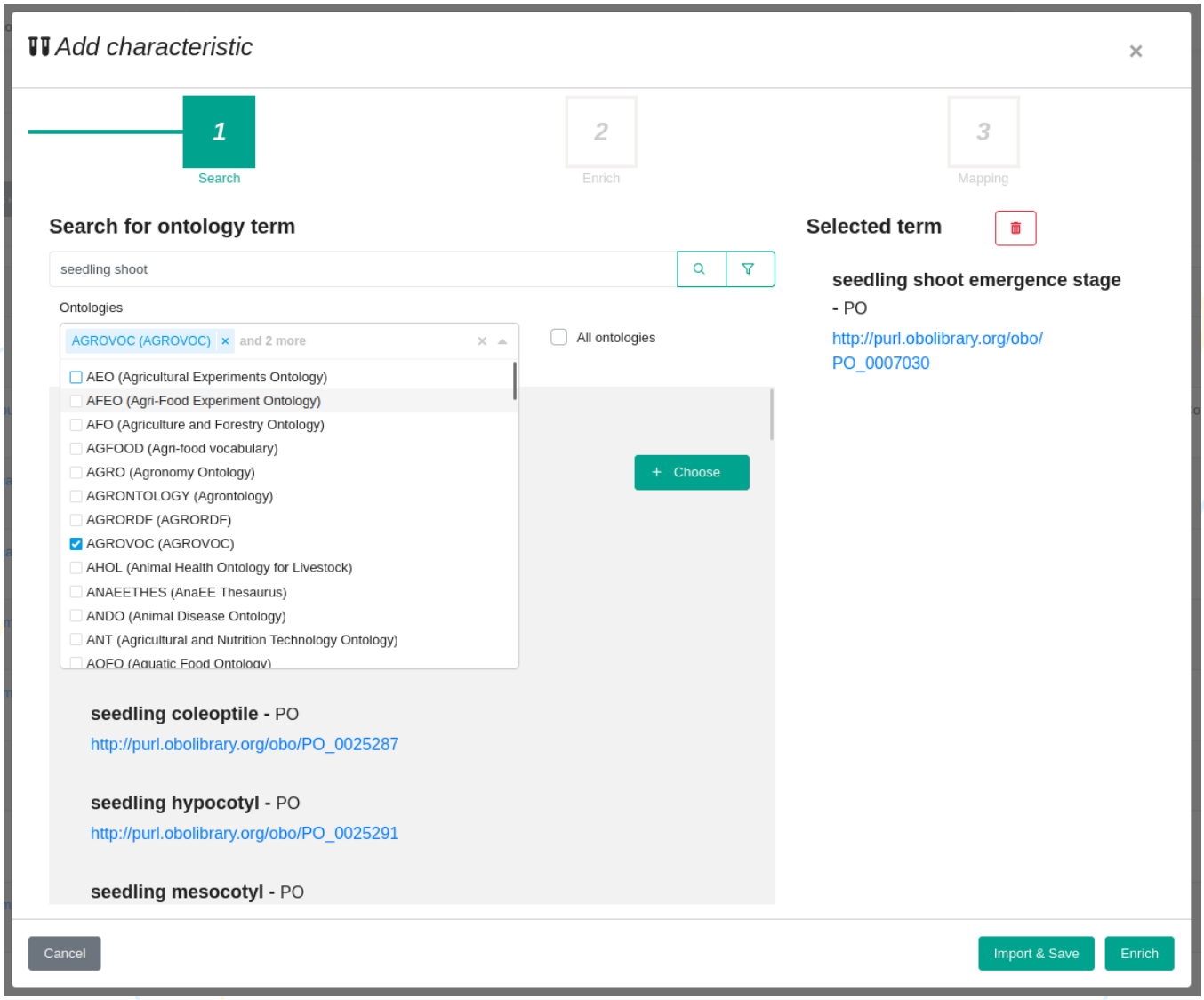


Figure 3. Prototype connector between PHIS and AgroPortal (under development)

This connector will facilitate the search for new terms to PHIS users.

In the future, following the same philosophy and technical behavior (API calls) in addition to consuming the content from AgroPortal, the connector will open up the possibility of proposing content (e.g., terms and mappings) to AgroPortal ontologies and semantic artefacts. **This will valorize the new intrinsic contributions made in creating scientific objects** (variables, terms, properties, etc.) by PHIS data experts which would engage with external ontology experts and users who validate these terms.

The connector will provide a number of benefits to users of PHIS and will ultimately contribute the AgroPortal's content, including:

- **Reduced time and effort required for reusing or mapping the terms** used in PHIS.
- **Improved interoperability** between different dataset and ontologies.

The connector is currently under development as a demonstrator within FAIR-IMPACT T4.5.

Eventually, **the connector would be made completely generic** to work with any instance of OntoPortal (on the semantic artefact catalogue side) and on any instance of PHIS (on the data repository side).

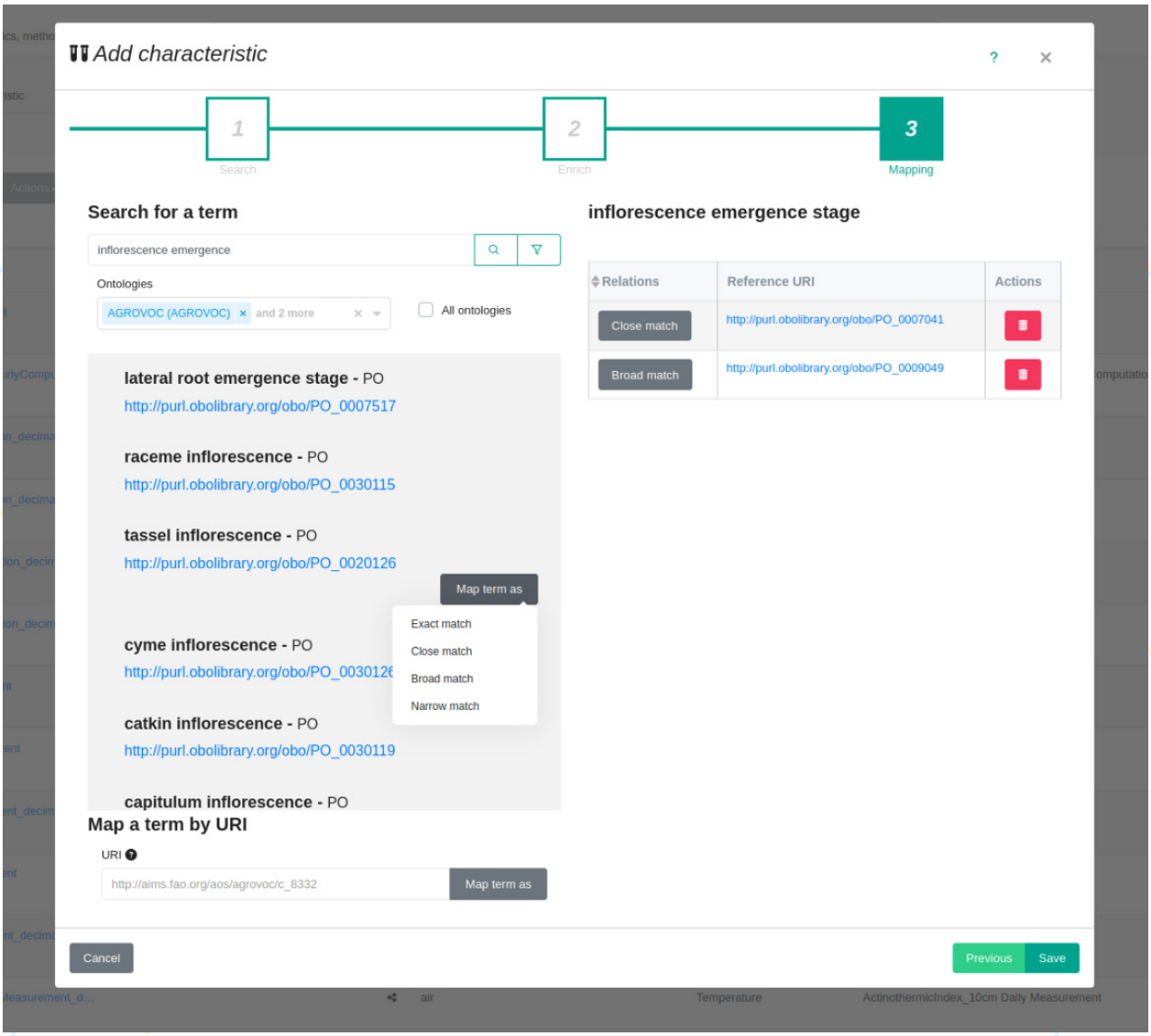


Figure 4. Connectors' mapping functionalityof terms

Enhance the semantic functionally of the national Earth & Environmental Data Repository by integrating it with the EarthPortal

- **Key topic** Interoperability | Metadata & Ontologies
- **Scientific domain** Earth and environmental sciences
- **Leading organisation** Dataterra (CNRS)
- **Contributors** Christelle Pierkot, Guillaume Alviset. Hélène Bressan

Overview

Easy Data is the French repository for long-tail data relating to the Earth and the Environment. EaSy Data uses vocabularies initiated by scientists to fill in specific metadata elements such as **topics** or **keywords**. These vocabularies are incomplete and do not reflect the complex diversity of data deposited in EaSyData. Other vocabularies exist and could be used to complete this vocabulary. Our aim is **to improve the semantic functionality of EaSyData by connecting it to the EarthPortal**. By combining the vocabularies displayed in the EarthPortal and EaSy Data vocabulary with associated services such as the annotator, we can offer researchers a more comprehensive set of tools to specify metadata. This will allow us to propose new ways for the researcher to specify metadata and thus improving the semantic capabilities of EaSyData. With four data clusters related to the Earth system and the environment (atmosphere, solid earth, continental surface and ocean), the French research infrastructure Data Terra wants to improve its data repository service (EaSyData) with a better use of semantics.

Context and objectives

EaSy Data is a French national data repository dedicated to long-tail data related to Earth and Environment, **using the ISO 19115 standard to describe them**. EaSy Data uses Geonetwork in the back office to store metadata, and an ad-hoc application layer has been developed to support the data repository and related functions (deposit, search).

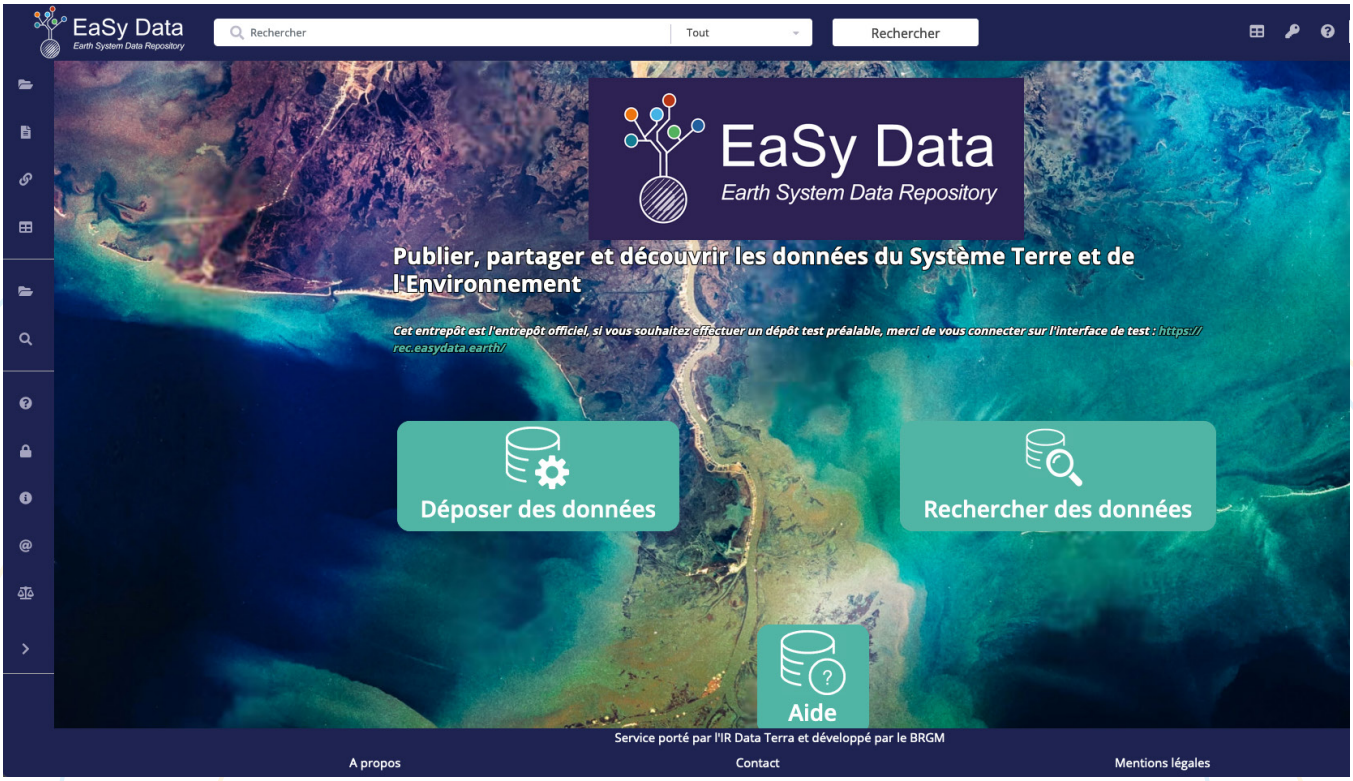


Fig. 1: EaSy Data repository

To fill the metadata elements related to keywords or topics, community controlled vocabularies are used. These specific vocabularies have been defined by scientists and are maintained in an experimental registry based on UKGovLD which has limited services to administrate and suggest new keywords.

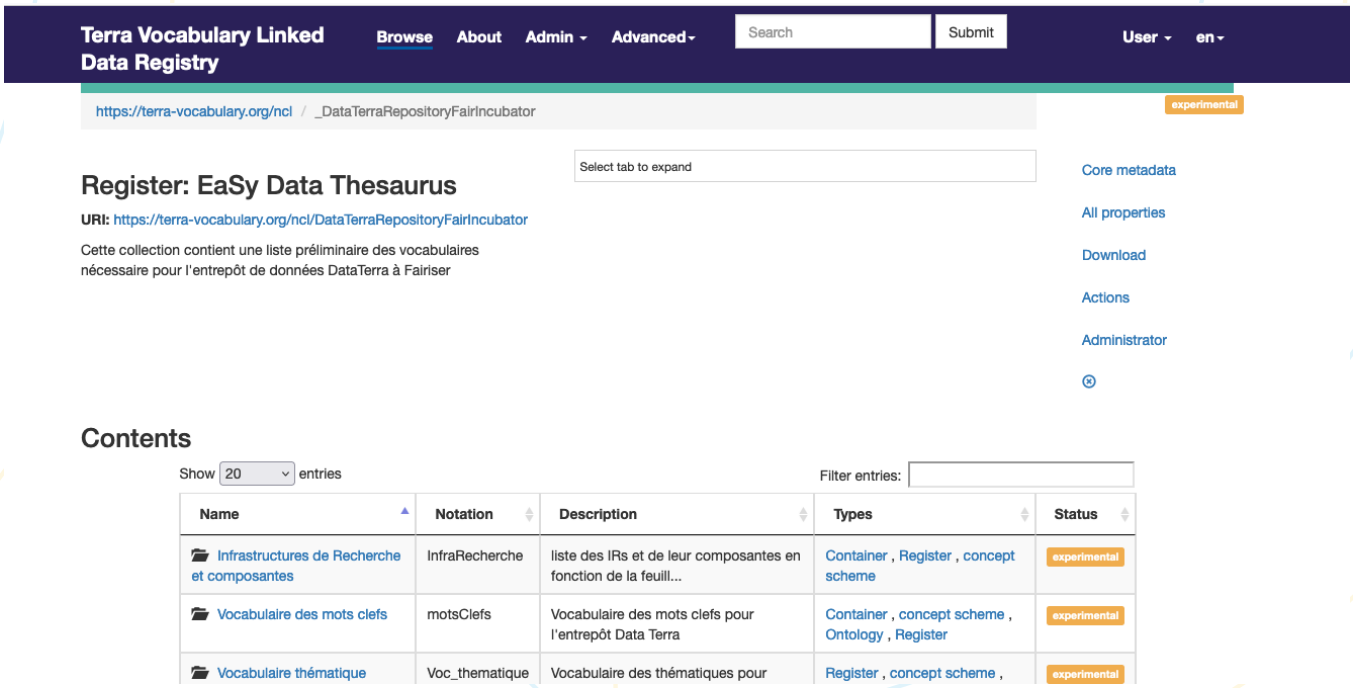


Fig. 2: EaSy Data Thesaurus (UKGovLD)

Furthermore, **these vocabularies are incomplete and do not reflect the complex diversity of data deposited** in EaSyData. Other vocabularies exist and could be used to complete this vocabulary. For example, some of the thesauri produced by French data clusters, such as those supplied by Theia/Ozcar, or by European infrastructures, such as those defined by EPOS or ACTRIS, could be used.

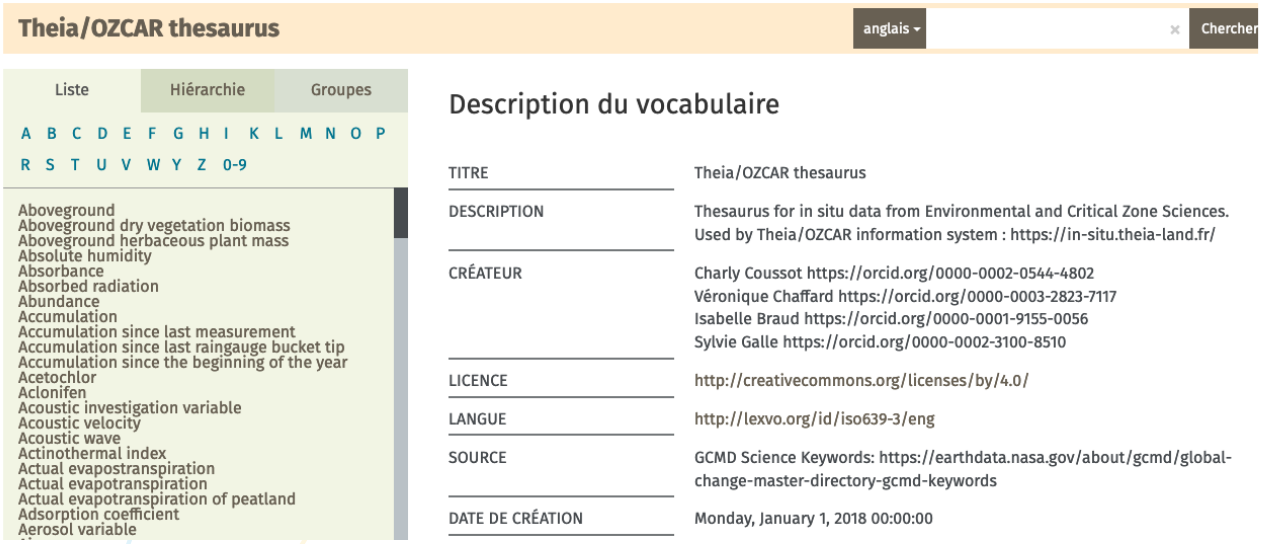


Fig. 3: Example of Theia/ Ozcar vocabulary (Skosmos)

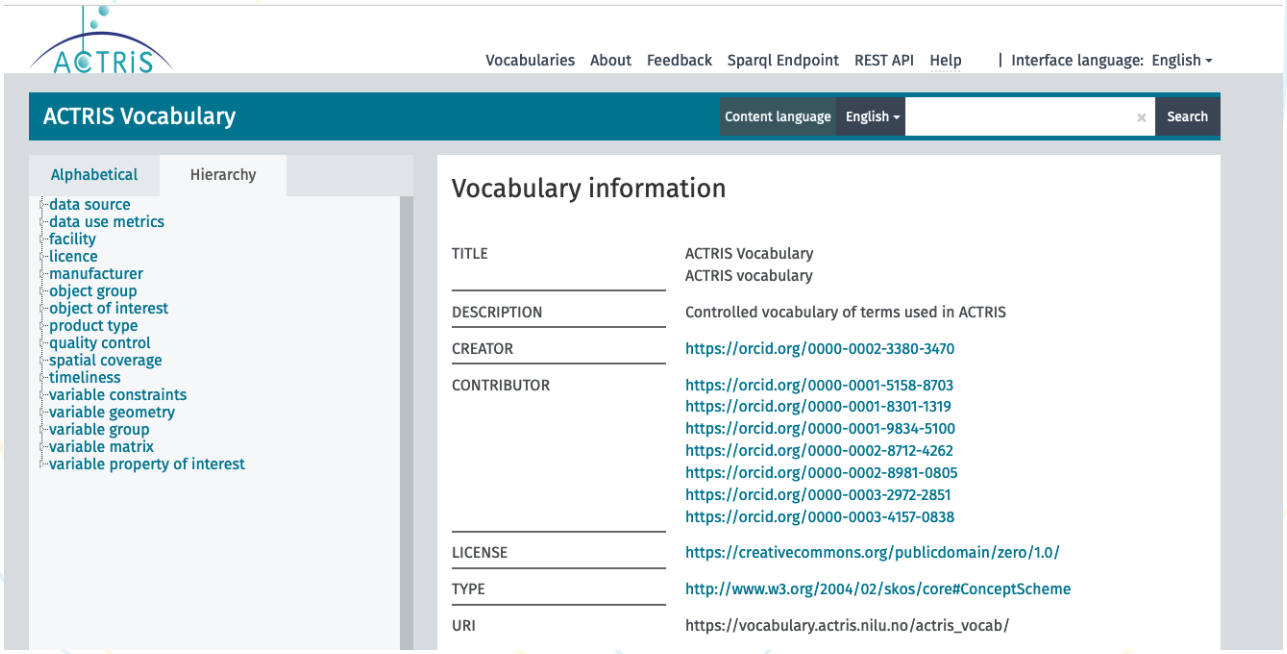


Fig. 4: Example of ACTRIS vocabulary (SKOSMOS)

Thus, one of our aims is to improve the semantic functionality by using these additional SA in EaSy Data.

Challenges and solutions implemented

In EaSy Data, we need to use community vocabularies at several levels: **administration** (to manage users by topic, e.g., moderators linked to specific topics), **depositors** (to better describe datasets and avoid spelling biases, etc.) and **users performing searches** (to improve search guidance). This is in line with the FAIR principles, which explicitly require the use of FAIR community vocabularies. The use of a catalog to reference

and update existing vocabularies, and potentially to host specific vocabularies for EaSyData, adds significant value.

To achieve this goal, we need to change the vocabulary tool used in EaSyData to better manage vocabularies and benefit from enhanced services.

The EarthPortal is a thematic semantic artefact catalog and repository for the Earth sciences using the OntoPortal technology. It has been deployed in the context of the FAIR-IMPACT Task 4.2 to host Earth and Environmental semantic artefacts (SA) that can be used by external applications through its REST API. Moreover, EarthPortal provides tools such as **the Annotator** (makes term suggestions based on text input), **the Recommender** (suggests relevant SA based on text input) and **Mappings** (generates, stores and displays mappings between SA). These tools could be useful to improve the EaSy Data semantic functionality.

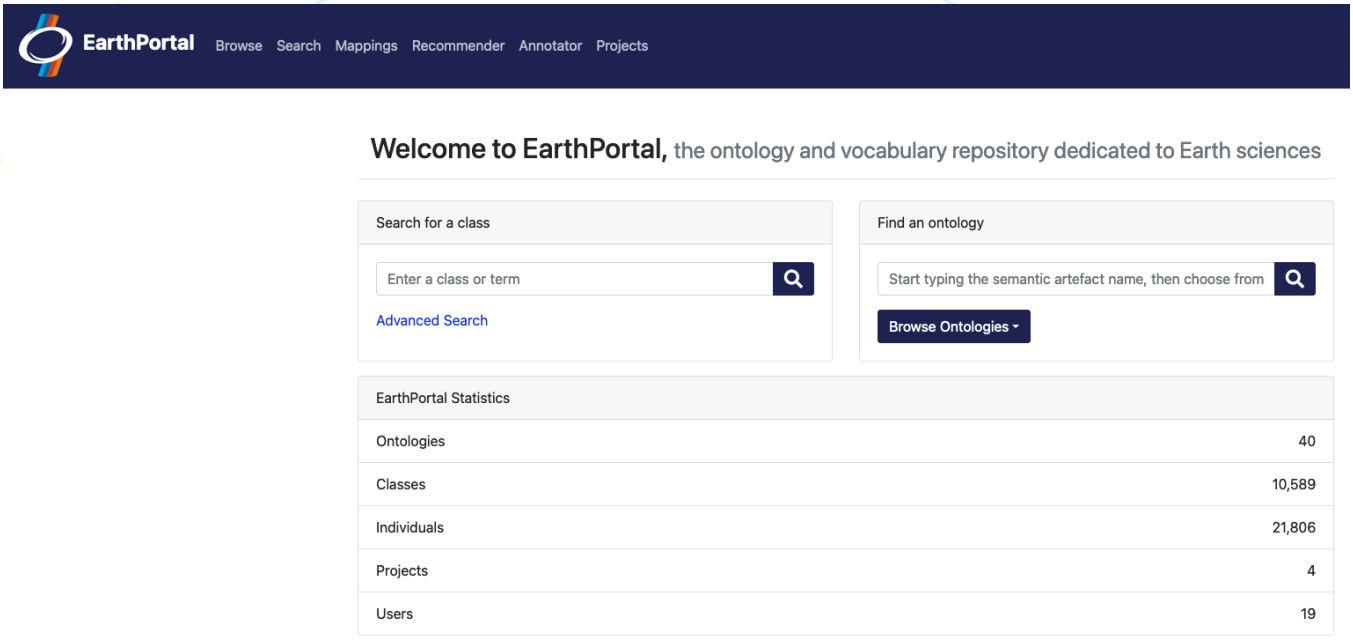


Fig. 5: EarthPortal homepage

Data Terra's thesauri, as well as vocabularies produced by European infrastructures and other more generic but commonly used SA (e.g. SOSA, Sweet, etc.) are already available in EarthPortal.

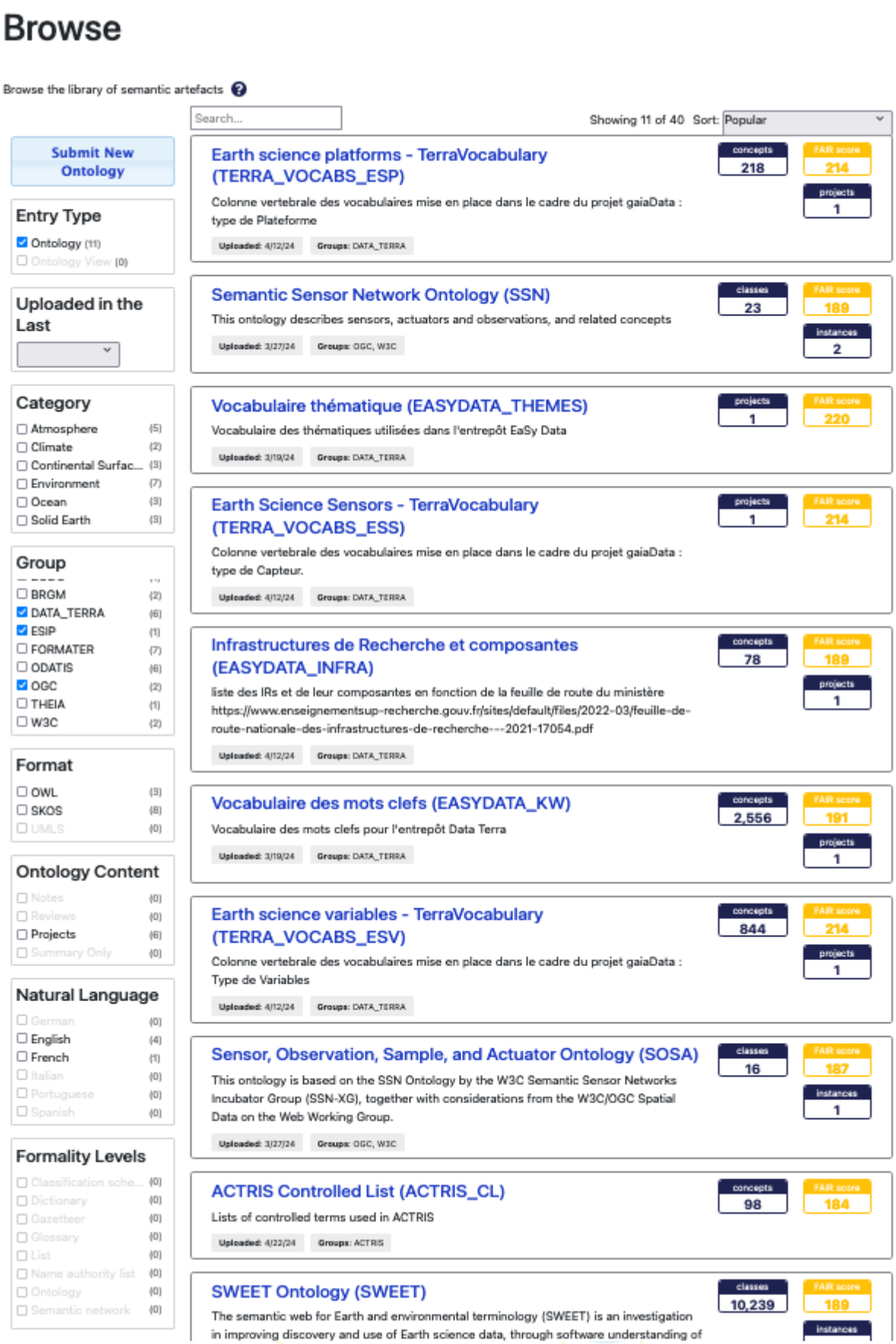


Fig. 6: Example of semantic artefacts in the EarthPortal

EaSy Data will harvest these vocabularies directly from the Earth Portal's existing REST API, in order to offer users more terms than those initially defined. EaSy Data will also use the annotation service offered by EarthPortal.

This will improve the user experience in two ways:

- **To populate keywords and topics metadata**, the EarthPortal will allow the user to check additional vocabularies to those currently used by EaSy Data.
- **Annotator service will be used to suggest new terms** from the user-written abstract to enrich the metadata with new terms.

Expected/Measured Impacts

From a user perspective, we expect several improvements that should enhance the FAIRness of the datasets and related publications in EaSy Data:

- **A better access to the vocabularies** and the possibility to contribute to them;
- **An extended semantic description of the datasets**, allowing better discovery of related resources;
- **Semantically enriched metadata of the datasets** by looking at the related SA to suggest related vocabulary concepts.

From an administrator's point of view, we expect a better management of the EaSy Data vocabularies and a smarter use of other vocabularies.

Leveraging AgroPortal ontologies to ease metadata completion and data discovery in Data INRAE

- **Key topic** Interoperability | Metadata & Ontologies
- **Scientific domain** Life science
- **Leading organisation** INRAE
- **Contributors** Bilel Kihal, Carmen Corre, Clement Jonquet, Dimitri Szabo, Jerome Roucou, Sophie Aubin

Overview

Data **INRAE is a French institutional data repository** (INRAE ; France's National Research Institute for Agriculture, Food and Environment), part of "Recherche Data Gouv" (French Data Repository), and based on Dataverse technology. **Datasets are referenced with key words, selected by dataverse managers**. In the current way, these managers can use any terms or semantic artefacts and few belong to control vocabularies. This use case aims to connect AgroPortal with Data INRAE. **AgroPortal is a semantic artefacts catalog for agri-food and related domains**. The goal of the connection is to facilitate the control vocabularies use for keyword completion. Control vocabularies improve our ability to find and reuse stored data and participate in their interoperability. In fine, a better keyword usage should improve data INRAE FAIRness. In this use case, we will evaluate the practices and data FAIRness evolution.

In recent years, an increasing number of data repositories have been deployed to address the need of research data publication and reuse. In the case of INRAE, France's National Research Institute for Agriculture, Food and Environment, research data is either shared via domain repositories or via an institutional repository: Data INRAE, now a part of the French federated national research data platform Recherche Data Gouv.

This national repository is based on the open source research data repository software Dataverse. Data repositories softwares such as Dataverse allow datasets to be documented by metadata, but these metadata fields often function as sole texts rather than semantic concepts, without enrichment, expanded search on related terms or multilingualism.

Semantic artefacts of interest to INRAE are hosted in AgroPortal, a repository for ontologies and other semantic artefacts in agri-food and related domains. AgroPortal is based on the generic technology OntoPortal developed jointly by INRAE-MISTEA, University of Montpellier and the OntoPortal Alliance. AgroPortal allows users to search and browse for terms in a user-friendly interface and can also be called automatically by tools through APIs.

This use case aims at bridging the gap between these platforms; data repositories and semantic artefacts catalog, by developing a connector in Data INRAE to be able to use semantic artefacts from AgroPortal in an user-friendly way, and make it usable and available to all users of Dataverse or Ontoportal technologies.

Implementation solution: Connector between Data INRAE and AgroPortal

Concretely, with this new feature (cf. Fig2),

- ☐ The dataverse manager selects relevant semantic artefacts from AgroPortal
- ☐ The data depositor picks keywords from ontologies and thesauri that have been connected
- ☐ Dataverse records URIs and any relevant information for these keywords
- ☐ Information from the selected keywords can be exported with other metadata of a dataset

Fig. 1: Current Data INRAE metadata feeding. Four boxes are available for (1) Term, (2) Term URL, (3) Vocabulary, and (4) Vocabulary URI. Users feed boxes with information from any semantic artefact catalogs.

Fig. 2: Use Case goal, Data INRAE' s metadata feeding, after AgroPortal and DataInrae connection. An unique search box allows the user to find AgroPortal's terms. Terms and vocabularies associated informations are

search box allow the user to find AgroPortal's terms. Terms and vocabularies associated informations are imported at the same time.

Challenges that need to be addressed

This integration should make it possible to propose relevant semantic artefacts to the community with terms from agri-food vocabularies being available on AgroPortal (and potentially in several others semantic artefact catalogs).

#user-friendliness - the evolutions proposed must ease the work of describing and searching datasets while remaining intuitive and fast-responding. The users will be involved in the specifications and tests of the new features.

#multilinguism - users of Data INRAE generally work in English or French and AgroPortal semantic artefacts can be multilingual. Evolutions of the Agroportal API are considered in link with T4.2.

#modularity - In the case of pluridisciplinary data repositories like the Recherche Data Gouv repository, depositors from various communities may need to select concepts from specific vocabularies. To address this need, the set of vocabularies to be used must be configurable at dataverse collections level.

#sustainability - At this time, a key challenge in this use case is to connect Data INRAE and AgroPortal, while making this connection generic and reusable for other installations of OntoPortal and Dataverse. For instance, we have to change the export feature to include the URIs of controlled terms used to index. The evolutions brought by the use case are discussed and shared with both developers' communities.

#evaluation - the use case must allow evaluating the impact of the evolutions proposed on the data FAIRness, Findability in particular.

Expected impact of the Use Case

The use case outputs will improve the FAIRness of Data INRAE's datasets (research and technical agri-food data). The connection with the vocabularies repository (AgroPortal) will result in more and richer keywords added by data producers. In addition, the use of semantic artefacts will enable the multilingual and synonymy search for increased findability and accessibility of datasets.

The use of documented, identified, unambiguous and more domain-relevant keywords thanks to external and specific semantic artefacts systems, which will allow increasing interoperability of the datasets. This will also promote and ease the use of standard terms and semantic artefacts for the communities.

These benefits will be achieved, while providing a faster and easier metadata completion, thanks to the simplification of metadata fields and their filling.

In addition, this should result in an increased use of the semantic artefacts in platforms using this connector. For example, AgroPortal will benefit from Agricultural Science users' needs for additional or improved terms from Data INRAE. Other scientific domains willing to be able to benefit from such features should also result in new terms and artefacts in semantic artefacts catalogs.

As these developments will be compatible not only with Data INRAE and Agroportal installations but for the softwares they use (respectively Dataverse and Ontoportal). The feature and its benefits will be reusable on all installations of these softwares in various domains.

Expected outputs

The tangible outcomes of this use case are:

- ☐ Dataverse-OntoPortal connector publically available in the community's repository
- ☐ Global improvement of semantic artefacts use in Dataverse
- ☐ Analysis of current practices and impact of the evolution on users and on datasets findability
- ☐ Guidelines on criteria to select relevant semantic artefacts for a data repository

Improving ecological (meta)data FAIRness through semantic services: integration of EcoPortal in LifeWatch Italy new platforms

- **Key topic** Interoperability | Metadata & Ontologies | Metrics, Certification and Guidelines
- **Scientific domain** Ecology and Biodiversity
- **Leading organisation** LifeWatch ERIC
- **Contributors** Ilaria Rosati, Enrica Nestola, Martina Pulieri, Parham Ramezani

Overview

LifeWatch Italy serves as the Italian Distributed Center for the LifeWatch ERIC Infrastructure, contributing significantly to the ERIC's functionality. Focused on biodiversity and ecosystem research data management, LifeWatch Italy enhances data sharing, integration, and analysis through its Data Portal and Metadata Catalogue. Recent efforts by LifeWatch Italy aimed at improving FAIRness involve integrating EcoPortal, a semantic artefacts catalogue, with the Data Portal and Metadata Catalogue. EcoPortal supports the scientific community in managing semantic artefacts in the ecological domain and employs the Ontology FAIRness Evaluator (O'FAIRe) tool for FAIRness assessment. Challenges we are trying to address include enhancing metadata annotation with FAIR semantic artefacts. The expected impacts of the integration between EcoPortal and the new Data Portal and Metadata Catalogue of LifeWatch Italy encompass easier ecological data discovery, annotation, machine-actionable meta(data), and a push towards Linked Open Data.

Context and objectives

LifeWatch Italy (LW ITA) is the Italian Distributed Center of LifeWatch ERIC e-Science infrastructure for biodiversity and ecosystem research and contributes in-kind to its functioning. LifeWatch Italy activity focuses on the management of biodiversity and ecosystem research data, as well as the development of tools and services for their sharing, integration and analysis. Among the different offered services, LW ITA provides the Data Portal and the Metadata Catalogue.

The Data Portal is a data repository, based on DSpace, that provides FAIR data and metadata. It helps scientists to share their (meta)data and also to reuse data created by others. The data schema is based on the Darwin Core standard and controlled vocabularies. The metadata schema associated with each dataset is the Ecological Metadata Language profile LifeWatch (EML 2.2.0; Vaira et al., 2022).

The LifeWatch Italy Metadata Catalogue is an information management system based on GeoNetwork 4.2.2, designed and implemented to enable access to several resources from a variety of providers through descriptive metadata, enhancing and promoting the information exchange and sharing among organizations and researchers. The LifeWatch Italy Metadata Catalogue gives access to different resources and their metadata that are based on two main standards:

- ISO 19139 (VREs, services, workflows and research sites);
- EML 2.2.0 (datasets).

Each metadata profile is organized in sections, which reflect the main information related to each specific resource. Each section contains optional and mandatory metadata elements (Vaira et al., 2022).

One of the objectives of the last implementation was to provide a new version of the Data Portal and the Metadata Catalogue and also integrate both with EcoPortal. The integration aimed to ensure the annotation and description of (meta)data using semantic artefacts.

EcoPortal uses an advanced version -collaboratively developed and experimented in FAIR-IMPACT– of the OntoPortal technology. It supports the scientific community in the creation and management of semantic artefacts and their use to harmonize (meta)data. The recent updates of EcoPortal have supported the improvement of the metadata schema (MOD 2.0) associated with semantic artefacts and its alignment with other catalogues and repositories (Tarallo et al., 2024). In addition, the integration of the Ontology FAIRness Evaluator (O'FAIRe) tool enables the assessment of the FAIRness level of semantic artefacts through a metadata-based automatic FAIRness assessment methodology (Amdouni et al., 2022).

Challenges and implemented solutions

We aim to implement a standardized annotation process in the LifeWatch Italy Data Portal and Metadata Catalogue for several metadata attributes, including keywords and dataset variables. The first requirement is to enable data exchange between these repositories and EcoPortal, where users can find semantic artefact terms to use for unambiguous annotation.

The second requirement is to develop a friendly User Interface (UI) for integrating property and value of metadata attributes with semantic artefact term URIs, labels and definitions. We also want to ensure that the semantic artefacts in use are FAIR, meaning they are findable, accessible, interoperable, and reusable by both humans and machines.

The integration involves implementing a REST API connector, using HTTP GET calls, to link the metadata wizard of the Data Portal and Metadata Catalogue with EcoPortal. This will improve (meta)data management to ensure FAIR compliance. The integration allows (meta)data providers to select attribute values for the metadata schema directly from semantic artefact terms published on EcoPortal by autocomplete features to enhance usability. The semantic artefacts are continuously updated on a regular basis, ensuring they remain accurate and relevant over time.

Figure 1 shows the semantic annotation within the Data Portal for the "Keywords" attribute. The same process is available for other attributes like "Software" and "Protocol". Users can access (Fig. 1a, b) and search (Fig. 1c) concepts and classes from semantic artefacts published within EcoPortal through a wizard. Users can select concepts and classes that are then reported, with their URI, inside the "Keywords" element (Fig. 1d).

A similar approach is followed for the attributes of the data table but it differs from the previous one because the definition of concepts or classes is also retrieved and is reported inside the element set as shown in Figure 2.

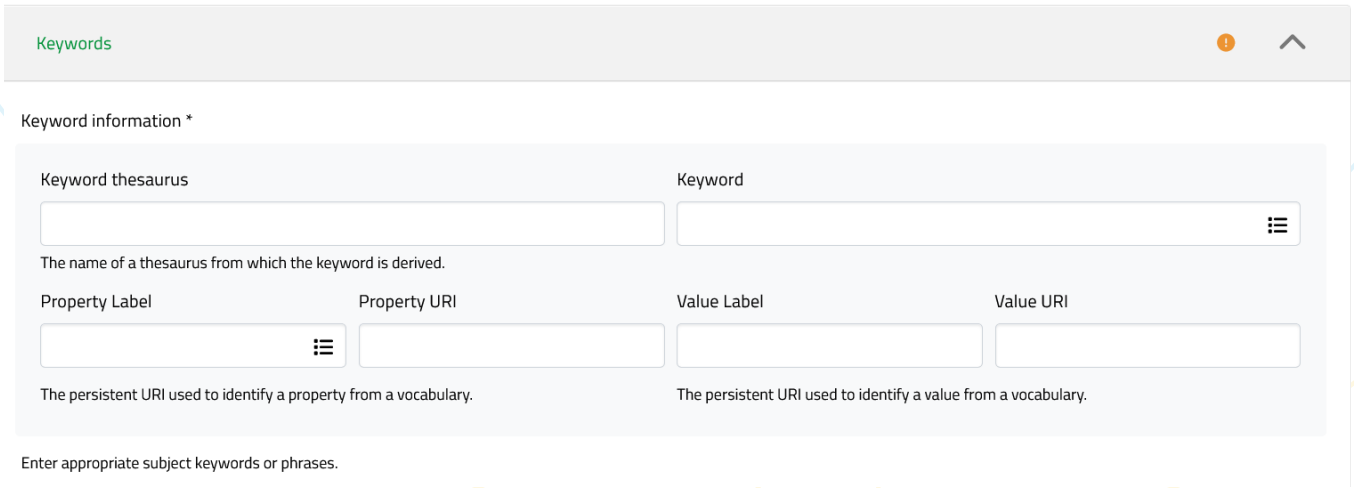


Figure 1a. Data Portal wizard. Keywords to compile.

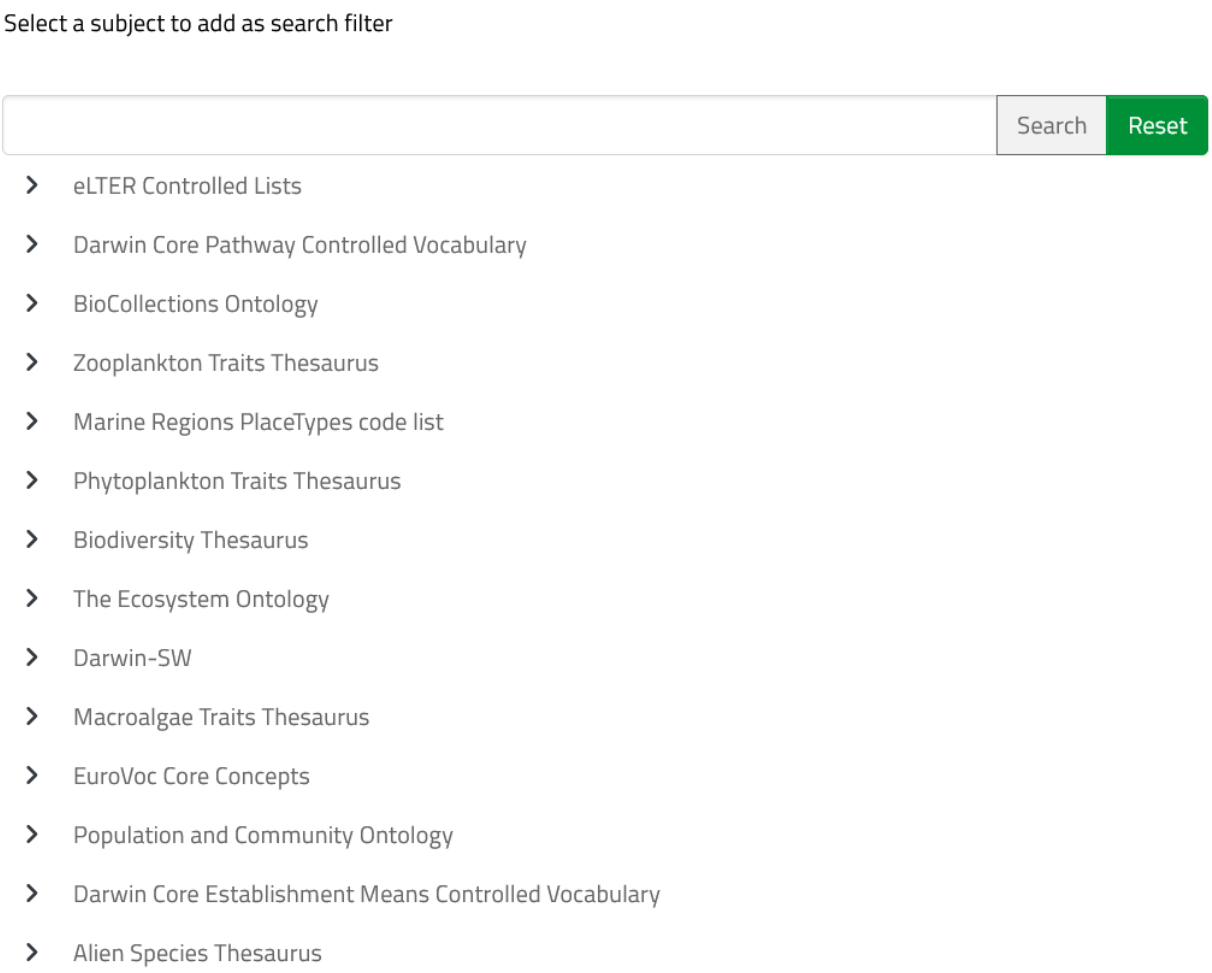


Figure 1b. Data Portal wizard. Point of access for semantic artefacts published within EcoPortal.

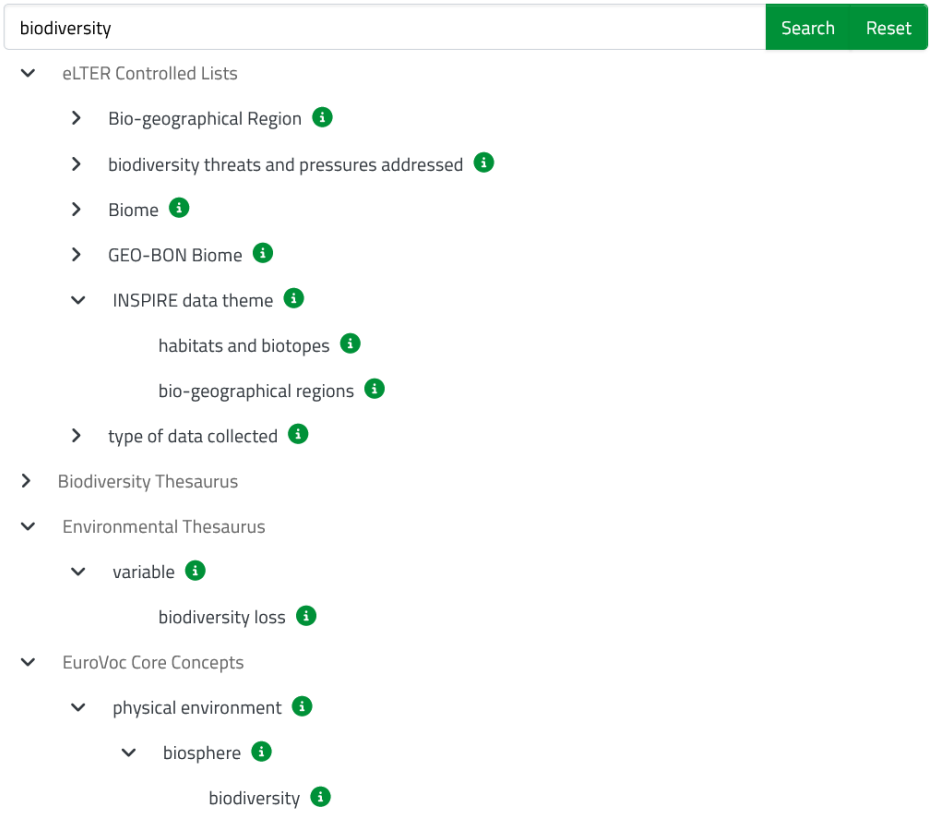


Figure 1c. Data Portal wizard. Example of a keyword search.

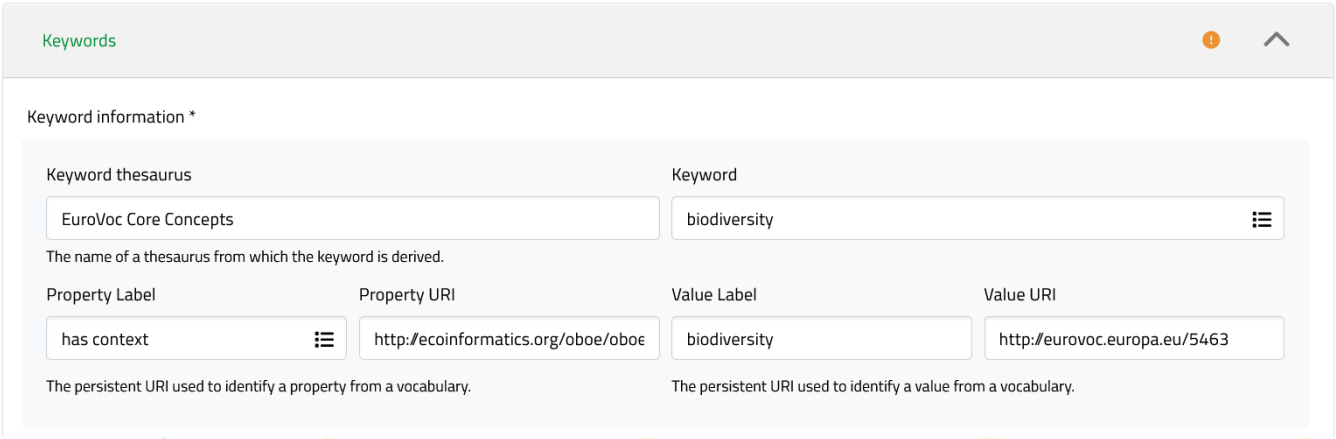


Figure 1d. Data Portal wizard. Keywords filled in.

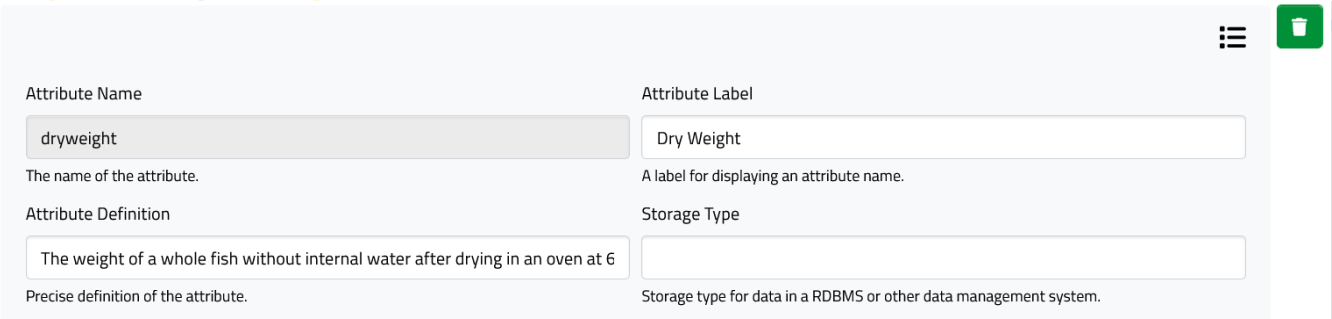


Figure 2. Example of data table's attribute compilation within the Data Portal. The "attribute definition" is retrieved directly from EcoPortal.

The process to obtain the semantic annotation within the Metadata Catalogue is shown in Figure 3. Users can search for concepts and classes directly in the search box (Fig. 3a). These fields are configurable by administrators to retrieve data from external resources (i.e., EcoPortal), enabling the autocomplete feature for filtering the list of occurrences and the possible value to be used (Fig. 3b).

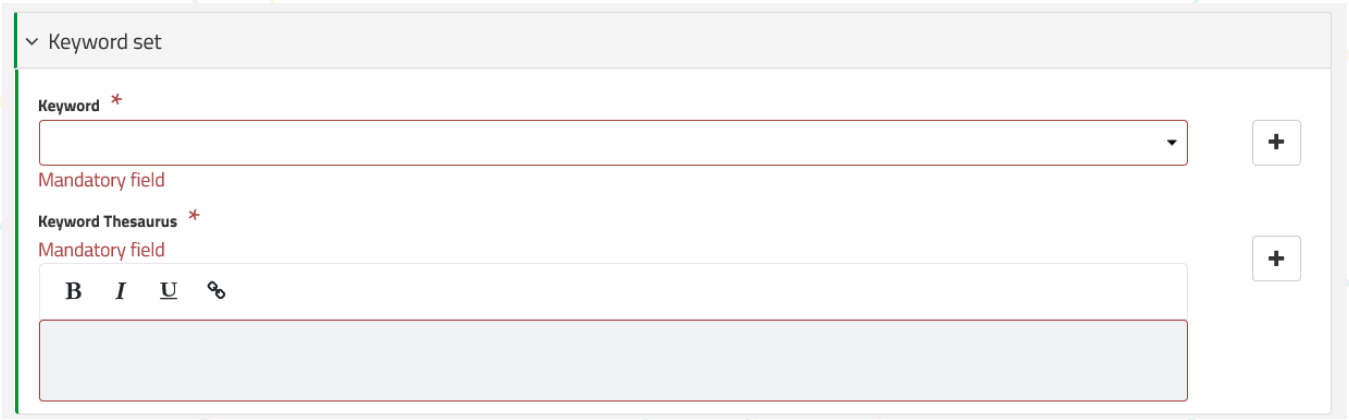


Figure 3a. Metadata Catalogue wizard. Keywords to compile.

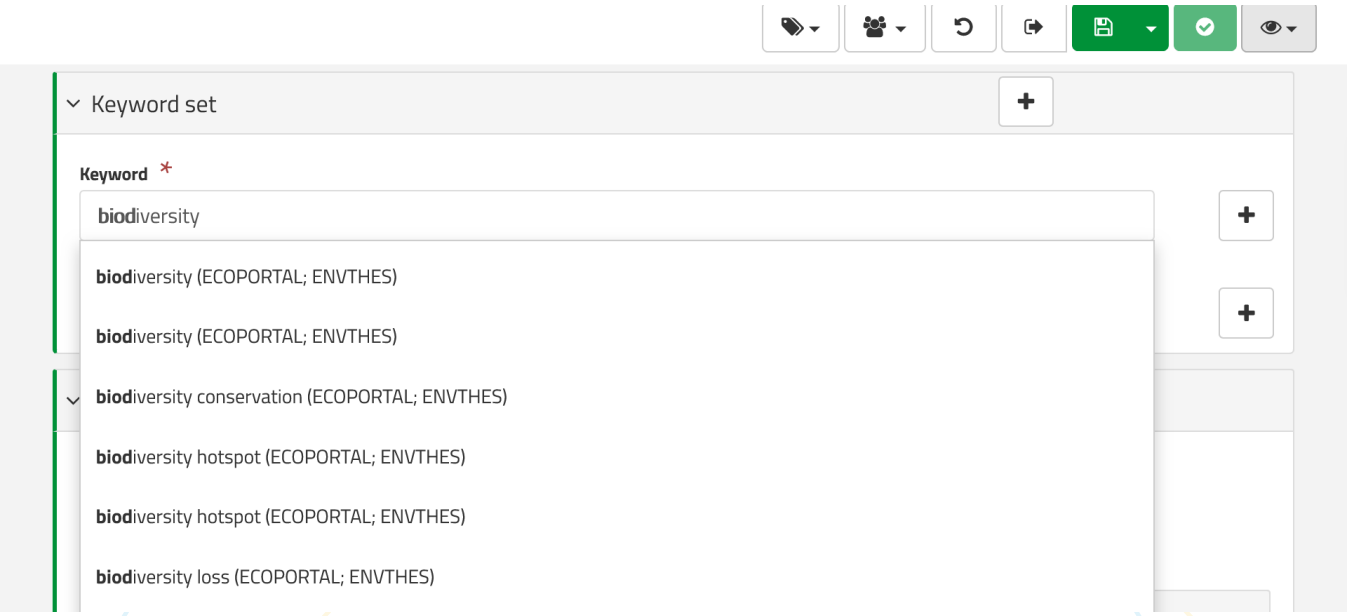


Figure 3b. Metadata Catalogue wizard. Drop-down list of concepts and classes.

Expected/measured impacts

The implementation of API connectors in the Data Portal and Metadata Catalogue, which support data providers to search for SA terms and retrieve URIs and labels directly within metadata submission forms, brings several key advantages to LifeWatch's data governance and management:

- **Improved Synergy:** This integration enhances collaboration across multiple platforms that are crucial to LifeWatch Italy's data management cycle.
- **FAIRness Improvement:** Expecting increased scores in key FAIRness assessment metrics used by the F-UJI tool integrated into the Metadata Catalogue, specifically:
  - FsF-I2-01M: Metadata uses semantic resources
  - FsF-I3-01M: Metadata includes links between the data and its related entities
  - FsF-R1-01MD: Metadata specifies the content of the data
- **Increased Reuse of SA Terms:** Anticipating a rise in the reuse of SA terms from EcoPortal, promoting more consistency and interoperability across (meta)data.
- **Enhanced User Satisfaction:** Simplifying the annotation process is expected to improve satisfaction rates among data providers and managers by offering a more efficient workflow.
- **Better EcoPortal Visibility:** Anticipating increased visibility and reuse of EcoPortal's SAs, driving wider adoption across the LifeWatch ecosystem.

With this new functionality, we can plan to harmonise and standardise the annotation process to enhance interoperability across all (meta)data assets.

Reference materials

- Amdouni E., Bouazzouni S., Jonquet C. (2022). O'FAIRe makes you an offer: Metadata-based Automatic FAIRness Assessment for Ontologies and Semantic Resources. International Journal of Metadata, Semantics and Ontologies, 16 (1), 16-46. https://hal-lirmm.ccsd.cnrs.fr/lirmm-03630233
- Di Muri, C., Pulieri, M., Raho, D. Muresan A.N., Tarallo A., Titocci J., Nestola E., Basset A., Mazzon S., Rosati I. Assessing semantic interoperability in environmental sciences: variety of approaches and semantic artefacts. Sci Data 11, 1055 (2024). https://doi.org/10.1038/s41597-024-03669-3

Tarallo, A., Pulieri, M., Ramezani, P., & Rosati, I. (2024). Advancements in EcoPortal: Enhancing functionalities for the ecological domain semantic artefacts repository. FAIR Connect, 2(1), 1-7. DOI: 10.3233/FC-240002

Vaira, L., Fiore, N., & Rosati, I. (2022). LifeWatch ERIC Application Profiles (Version 1). LifeWatch ERIC. https://doi.org/10.48372/8528-9Z45

PIDs as a cornerstone in actualising the FAIR principles within the LifeWatch infrastructure

- **Key topic** PIDs
- **Leading organisation** LifeWatch ERIC
- **Contributors** Parham Ramezani - LifeWatch ERIC, Nicola Fiore - LifeWatch ERIC

Overview

LifeWatch ERIC is a European Research Infrastructure Consortium dedicated to advancing e-Science research in biodiversity and ecosystem studies, supporting global sustainability challenges. Our mission involves uniting diverse scientific communities and creating a cutting-edge e-Science Research Infrastructure by connecting distributed observatories and research centers into unified accessible online platforms. A main component of our infrastructure, the LifeWatch ERIC Metadata Catalogue, underpinned by GeoNetwork, streamlines the management of metadata for various resource types, including Datasets, Research Sites, Services, Virtual Research Environments, and Workflows. It offers advanced search and user management functionalities, facilitating resource discovery and access control. Furthermore, we leverage modern semantic technologies, offering a transformative approach to comprehensively describe and interconnect diverse data sources, reducing barriers to data exchange among researchers. To achieve this, LifeWatch ERIC's EcoPortal plays a crucial role in ecological research, providing a Semantic Artefact Repository that consolidates core ontologies, domain-specific vocabularies, and reference lists. It offers essential services to facilitate seamless discovery and integration, exemplifying our commitment to advancing scientific collaboration and knowledge management.

Persistent identifiers (PIDs) are fundamental in actualising the FAIR principles within our infrastructure. EcoPortal facilitates the acquisition of Digital Object Identifiers (DOIs) for hosted resources via Datacite services, concurrently offering the capability to specify authors and contributors through ORCID iD integration. Additionally, PIDs are instrumental in affiliating institutions through the Research Organization Registry (ROR). Within the LW ERIC Metadata Catalogue, registered users seeking to create new resources (Datasets, Virtual Research Environments, etc.) must select the resource type and template, and provide mandatory metadata. This Catalogue facilitates the generation of DOIs for resources lacking them, leveraging the DataCite connection, following validation and verification. When it comes to workflow submissions, the challenge arises not only in assigning PIDs to the workflows themselves but also in allocating unique identifiers to each constituent service. This granular aspect of PIDs presents a pivotal consideration.

Challenges that need to be addressed

Managing dataset granularity in the Virtual Research Environment involves a critical question: how to handle the provenance of newly composed datasets? This challenge centers on deciding whether to assign DOI/PIDs to the entire dataset or its constituent provenance subsets. Another intricate issue is version granularity, which encompasses versioning of the entire artefact as well as individual entities, necessitating a clear linkage between deprecated and valid entities. Furthermore, we must establish a PID policy that aligns seamlessly with the EOSC Policy, ensuring compliance with essential standards.In response to these challenges, we're actively developing LifeBlock, a blockchain-based prototype. This innovative solution tracks all activities related to specific research objects, aiming to automate PID management through blockchain technologies. With LifeBlock, we're committed to enhancing our infrastructure's robustness, enabling more efficient management of complex datasets and their associated PIDs

Expected impact of the Use Case

PIDs are essential for tackling challenges associated with name changes, ensuring unambiguous references to individuals, and accurately attributing credit throughout a researcher's career. They also eliminate location dependency for digital objects. PIDs not only provide unique identification but also establish vital links between research entities, creators, and institutions, enriching the interconnection through machine-readable metadata in our systems. They enhance the discoverability, accessibility, and usability of research entities, enabling precise referencing of specific resource versions. PIDs also improve resource intelligibility by revealing their origins, enhancing accuracy and information flow. Moreover, they promote interoperability, bolstering trustworthiness through transparency and provenance. This interconnected network of specifically identified entities forms a robust foundation for assessment and evaluation within our infrastructure.

Expected outputs

While the original aim of PID services was to offer persistence, LifeWatch goes beyond the basic idea of just having persistent identifiers for content. Instead, with the ongoing mission to address our challenges and gain a better understanding of how PIDs relate to each other and the broader context, we aspire to establish a holistic research ecosystem

Change triggers impacting PID generation for sensitive data within the Social Sciences

- **Key topic** PIDs
- **Scientific domain** Social Sciences and Humanities
- **Leading organisation** CESSDA | UESSEX-UKDS
- **Contributors** Hervé L'Hours - UESSEX-UKDS/CESSDA, Josefine Nordling - CSC (T3.2 lead)

Overview

The UK Data Service (UKDS) is a partnership between the Universities of Essex, Manchester, UCL, Edinburgh and Jisc and the UK service provider to the Consortium of Social Science Data Archives (CESSDA).

The focus of this use case is on the research 'studies' of sensitive nature deposited at the UK Data Service and the subsequent derived data products. The sensitivity issue is the handling of information containing directly identifiable personal data or data that has the capacity to lead to reidentification, either through related variables within the dataset or through linkage with other data. We need to consider the implications involved in handling sensitive data for PID management and associated kernel metadata of a related identifier/object, e.g. for a repository identifier ("is currently approved to hold sensitive data") or a researcher identifier ("has current credentials for accessing sensitive data").

We will ensure the inclusiveness of metadata by considering the different requirements of PID management for digital objects beyond datasets. In CoreTrustSeal terms, the focus is Digital Object Management. Version changes to a digital object can trigger many different outcomes and the presence of sensitive data sets additional conditions on the digital object management process. Maintaining the provenance information is important throughout the process.

Sensitive digital object may or may not be assigned a persistent identifier and associated sensitivity metadata during the Conceive, Create and Collect phase. At the point of deposit, this sensitivity-related information may already exist and need to be integrated, or the repository may be entirely responsible for the assignment and handling of the identifier and related metadata. Throughout the following work on curation, quality, compliance discovery, identification, access and reuse the sensitivity metadata may need to be updated as digital objects are copied, changed, versioned and linked.

Description

Dealing with sensitive data requires careful consideration in terms of sensitivity assessment and versioning management. This use case will define and identify all the possible scenarios for 'change triggers' that could impact PID generation, PID metadata changes (change logs) and the 'declaration' of a new object during the various phases of the research data life cycle. The domain for exploration is Social Sciences, which comprise varied data types, formats and sources. The data is of both qualitative and quantitative in nature (statistical and survey data) and DDI is the predominant form of metadata. The immediate scoping of this use case is on controlled data (for data that may be disclosive) held in deposited digital objects. However, there is a need to explore and clarify the relationship between Sensitivity, Confidentiality and Disclosure.

Once the change triggers have been defined, there is a call for expansion of views. As we expand and explore these issues the focus is on defining criteria when a new persistent identifier is required, when it is sufficient to update metadata without a new PID and when that metadata should form part of the metadata kernel. This stage is followed by review and alignment with the PID policy and any implementation guidance. There is also

a clear call for alignment with the issues around complex data citation and data production workflows in PID management.

For file based objects we have to consider the dependencies between files and the potential for granular version changes that can be triggered by changes within a hierarchy, such as a series of questions or variables. We also need to consider whether a change/version/new identifier for a child has implications for a 'parent'. For semantic data, the network of version and identifier changes could be influenced by third parties who have control over some part of the linked data 'set' (e.g. a semantic artefact such as a controlled vocabulary). The level of granularity to which PIDs are provided impacts the possible granularity of citation.

Later plans include looking in more detail at the implication of different levels of responsibility for the data and/or metadata and the rights management (including machine-actionable rights). We will also identify whether previous identifiers exist, whether these identifiers will be maintained, whether they will be updated by the repository or whether a new identifier (and version/provenance model) will be applied. Lastly, we will define the handling of multiple deposit events and the impact on any identifiers.

Challenges that need to be addressed

Communication and clarification of how different approaches to copies, changes and versions are handled across different environments and domains are a foundational challenge to any best practices around digital object lifecycle management. We will consider whether some initial general-purpose documentation on this topic is required to support a more brief and digestible approach to the specific issues surrounding sensitivity. Sensitive data about people faces issues of both perceived and actual risk which must be addressed for human subjects, researchers, repositories, funders and the wider public. Communicating disclosure risk and mitigation processes can be complicated and technical. Overall transparency of practice delivered through safe and trustworthy organisations are critical. For this reason, the metadata about digital objects and metadata about the research projects, repositories and reuse environments that care for them must be aligned. We will seek to provide a generally applicable set of approaches while also addressing the particular challenges of linked data.

Expected Impact of the Use Case

Leveraging on the expertise of a large international consortium among social science data archives, where we have access to use case specifications from across service providers (CESSDA) through the UK Data Service. A well-designed guidance of PID usage for sensitive data (access and management), will provide better-aligned and documented practices that support the interoperability of organisations and digital objects, across secure environment borders. Furthermore, more efficient use of PIDs for sensitive data will benefit research, and thus have great societal and economic impact.

Expected outputs

Best practice documentation.

INRAE - Providing a recommendations document on PIDs usages

- **Key topic** PIDs
- **Scientific domain** Earth and environmental sciences
- **Contributor** François-Xavier Sennesal, INRAE

Overview

INRAE is the French National Research Institute for Agriculture, Food and the Environment. Through research, innovation and support for public policies, it proposes new directions to support the emergence of sustainable agricultural and food systems.

INRAE is the first French institute to have a Department for Open Science.

The objective is to respond to the challenges linked to the opening of scientific research in the context of digital development and increasingly strong expectations from society.

The main goal of this use case is the production of a recommendations document on PIDs, to be adopted and applied in the Institute. These recommendations concern different resource types, such as people, structures, events, sensors, documents and data. A specific effort has to be made on the versioning of PIDs and resources, especially for evolving data.

This recommendations document should lead research teams to adopt FAIR principles in their data management. Moreover, it should allow the Institute to propose new software services to implement these principles, as the Institute is responsible for the registration of PIDs and related resolutions.

Description

Each day, INRAE research teams produce plenty of data concerning animals, plants, landscapes and documents. All the data, for the majority coming from observations and experiences, have to be stored properly, then published in order to be valued and reused. It's moreover a necessity to make this data easily findable, thanks to software systems.

Finally, every data object must be uniquely identified to be potentially cross-referenced with other data produced anywhere in the world.

However, research teams do not currently use the same identification system to represent data. This situation is mainly due to the differences we observe between the methods used to store the data. We can see, for example, lab data stored in Excel files; in relational databases; in flat files. The teams which are more advanced in technology associate triple stores to their databases, to allow cross-usage of their data and favour key-words usage.

In other words, the technological differences, as well as the lack of information on the use of known identification systems, lead to a disparity in the manner to uniquely identify data.

Challenges that need to be addressed

The main challenge lies in composing a recommendations document, which adequately explains what kind of identifier should be used to represent this or that type of data. This document must be as precise as possible to explain how to organize data versioning and identifier versioning, as well as evolving dataset identification. Moreover, this document must be nearly from today's end-users usages to optimise adoption by research teams.

Expected Impact of the Use Case

The recommendations document must inform the choice of researchers in the use of this or that PID. It should make it possible to harmonize these choices within the institute. Researchers and engineers who wish/can will be able to implement technical solutions by implementing the recommendations themselves. Finally, these recommendations will allow the Institute to offer a PID management service guaranteeing their storage and resolution.

Expected outputs

PID management software system offered by the Institute to its research teams.

Advancing access interoperability with ODRL

- **Key topic** Interoperability | Metadata & Ontologies
- **Scientific domain** Social Sciences and Humanities
- **Leading organisation** UESSEX-UKDS
- **Contributors** Darren Bell - UESSEX-UKDS/CESSDA, Hervé L'Hours - UESSEX-UKDS/CESSDA

Overview

The UK Data Service (UKDS) is a partnership between the Universities of Essex, Manchester, UCL, Edinburgh and Jisc and the UK service provider to the Consortium of Social Science Data Archives (CESSDA).

The focus of this use case is on enabling machine actors to better interpret the currently ambiguous semantics of digital objects' access and usage conditions and secondly, to provide more specific guidance on how to encapsulate the definition and execution of access and usage conditions in FAIR signposting practices. The latter reference "license" as a link type but apart from referencing natural language license statements, this mechanism currently provides little scope for subsequent machine-actionable negotiation and execution of access/usage conditions for a digital object.

Access and usage conditions are typically specified, asserted and managed locally. Beyond classifying these with some shared, loosely understood categories such as "Open" and "Closed", such categories are largely bespoke to a particular repository. For example, UKDS has three top-level categories: Open, Safeguarded and Controlled and also supports embargoes. By way of contrast, OpenAire has openAccess, restrictedAccess, embargoedAccess and closedAccess. Such access categories are pivotal for FAIR however they encapsulate and signify a set of complex attributes, constraints and workflows but in a currently non-normative way. For the purposes of long-term global interoperability, such locally-defined high-level access categories are currently of little practical use beyond simple discovery, as they only signify precise meaning locally to the repository that assigns them.

An access category is normally assigned as the end result of a (typically human) assessment of the intersection of (a) the regulatory/legal context, (b) rights and usage prescriptions of the data owner, and (c) the disclosure risk of the data (itself a function of inherent properties of the data in isolation as well as emergent properties when the data is combined with other data). In most cases, these assessments are often non-deterministic.

W3C standards such as ODRL (Open Digital Rights Language) have emerged, which allows natural language rights statements to be formally represented as structured RDF data. This use case will create the first comprehensive coverage of ODRL statements for a national collection in the landing pages of UKDS "studies" (the primary object that acts as a container for datasets and documentation). This is a foundational first step to providing a machine-actionable corollary to hitherto natural language-based artefacts, such as licenses and data-sharing agreements.

Description

For researchers, access to data, particularly sensitive data, is too complex and takes much more time than it should. Much effort has been devoted to machine-actionable implementations of the FAIR principles but in the access arena, less progress has been made. Access and usage conditions are derived from the intersection of a number of factors: legal & statutory obligations, rights management assertions, external prescriptions from data owners, and intrinsic properties of the digital object e.g. more disclosive data will inevitably require more stringent access protocols. With the global recognition that interoperability will lead to better global services for researchers, access is no longer a second-order problem. Mediating researcher access to data has become a topic we can no longer leave primarily to humans' best administrative efforts, still largely informed by natural language license artifacts. Rights statements, legal obligations and access workflows need to be systematically

modelled and implemented in metadata and code, in order to be executed at scale by machines. ODRL, while not complete in its coverage of all aspects of rights and access management, is currently the most practicable way forward to deliver better access interoperability.

Challenges that need to be addressed

Attempting to harmonize top-level access categories across domains and repositories is unlikely to be a fruitful course of action: considering that CESSDA's data access policy took several years to reach an agreement on the most coarse-grained access categories. We will pursue a more granular, bottom-up approach that establishes core vocabularies for the key ODRL classes i.e. Parties, Permissions, Obligations, Prohibitions and Actions and best practices for representing this in ODRL policies. In practice, there are a finite number of items in these core vocabularies for the majority of access-related repository activities. Once they are available to deploy in ODRL policies, this will be a significant step forward in effectively modelling the definition of traditionally prose-based access/usage conditions statement. It is a precursor to a future goal (not in the scope of this use case) of connecting ODRL Actions to machine-actionable workflow definitions modelled in, for example, Common Workflow Language, leading to full end-to-end machine-actionable messaging and process choreography between repositories.

The machine-actionable access arena is relatively immature – compared to practices around discovery, for example. Simply communicating why this is important is not a trivial task to communities administratively and culturally accustomed to dealing with researchers' access to data as a largely human-mediated activity.

As well as providing a real-world production implementation of ODRL, we will provide guidance, both technical and more governance-related: the terminological overlaps and relationship between licensing, rights management, access and usage (among others) remain a barrier to precise articulation of problem statements and the design of solutions in response.

Expected Impact of the Use Case

Working with our partners in CESSDA in the newly created Sensitive Data Working Group, UKDS will endeavour to be the exemplar for an initial real-world implementation of ODRL and will encourage and advocate for the uptake of similar practice by other Service Providers in CESSDA.

We expect the benefits for data consumers to include:

- Medium-term:
  - transparency and efficiency in requesting data
  - consistency of access experience across different service providers
- Longer term:
  - automated processes and services across service providers
  - foundational infrastructure for future B2B federation of access workflows

For service providers, the standardized and structured approach to access through ODRL and associated controlled vocabularies will provide:

- Medium-term
  - Guidance on minimal best practices and design patterns for new systems development
  - Equity, and transparency in processing access requests
- Longer term
  - The ability to track and evaluate access requests more systematically helps provide more robust evidence to inform improvements to access management practices.
  - Opportunities for service providers to participate in multi-organisational and cross-domain collaborations

Expected outputs

Tangible outcomes/solutions

- ☐ Best practice documentation for embedding machine-actionable ODRL statements in resource landing pages and how this interacts with current FAIR signposting practice.
- ☐ Production implementation in UKDS catalogue.

Reference Materials

- ☐ CESSDA
- ☐ UK Data Service

Referencing software source code artifacts: identifiers for digital object

- **Key topic** PIDs
- **Scientific domain** Earth and environmental sciences | Life science | Photon & Neutron science | Social Sciences and Humanities
- **Leading organisation** Software Heritage
- **Contributors** Sabrina Granger, Morane Gruenpeter - INRIA, Josefine Nordling - CSC

Overview

Software identification refers to multiple practices depending if you focus more on describing (i.e. attributing credit to authors) or on referencing software. PIDs used to reference data sets, such as DOIs, are useful to reference a software as a project (i.e. the software as a concept, not a digital object). But referencing software artifacts (i.e. digital objects) with different levels of granularity calls for specific identifiers. Therefore identifiers in Software Heritage allow to reference a specific version of the source code of a project, at different levels of granularity: a snapshot, a release, a directory, down to a single file.

SWHID are unique identifiers intrinsically bound to the software components. The difference between extrinsic and intrinsic identifiers lies in the way the relation between identifier and designated object is created and maintained. SWHID don't rely on an external register. Thus, end-users can recompute identifiers on retrieved objects and verify the match.

An agreement on a standard therefore, in June 2023, a first stable version of the SWHID specification was published: it describes precisely how SWHIDs are computed. A working group gathering experts from different institutions had been launched in March 2023. The relevance of SWHIDs goes way beyond source code and Software Heritage.

The attribution of SWHID raises the fact that deduplication has to be built-in, the database of the archive itself has to implement an ad hoc data model. Thus, any software artifacts encountered in the wild gets added to Software Heritage only if a corresponding node with a matching intrinsic identifier is not already available in the graph—file content, commits, entire directories or project snapshots are all deduplicated, incurring storage costs only once.

The FAIR-IMPACT project is an opportunity to elaborate compelling use cases in order to help the stakeholders adopt best practices for software referencing.

Challenges that need to be addressed

The related challenges vary somewhat depending on the stakeholder at hand. We have defined separate sets of challenges for end-users, service and infrastructure providers and policy-makers.

The end-user oriented challenges lie in understanding why software calls for a specific PID, as software and data are distinct concerns, and in understanding the type of PID that is compatible with a particular need. For example, an end-user may need to cite a software as a project or to cite a fragment of source code. Yet, finding the exact matching code can be quite difficult, as the code excerpt is often edited a bit with respect to the original, e.g. to drop details that are not relevant for the discussion or due to space limitations.

The challenges of service and infrastructure (archives, aggregators, catalogs) providers relate to ensuring precise identification of software artifacts for reuse and reproducibility and in implementing workflows that ease the use of SWHIDs, e.g. the French repository HAL provides a deposit service that allows to archive a software via a SWHID and the

Image Processing On Line journal (IPOL) has decided to deposit systematically in the Software Heritage archive all the software artifacts associated to the articles it publishes. Other identified challenges is monitoring the adoption of SWHID among users via dedicated indicators, offering guidance to end-users, making the use of SWHID part of “current” science in all the academic fields, not only in computing science, and dealing with a posteriori curation for non-referenced software artifacts.

Policy-makers oriented challenges relate to ensuring that proper support actions are offered which promote the adoption of SWHID for software artifacts as early as possible in the software lifecycle and which assist in software archiving, as a prerequisite to reference software artifacts. In addition, regular monitoring efforts are needed to gauge the gap between current practices and recommendations.

Expected impact of the Use Case

Dissemination of software referencing good practices among stakeholders, partners and end-users from different scientific communities.

Expected outputs

It will result in a better understanding of the different use-cases related to SWHID.

Reference Materials

- ☐ Guidelines for researchers
- ☐ SWHID tutorial

Implementation of EOSC Interoperability Framework at STFC

- **Key topic** Interoperability
- **Scientific domain** Photon & Neutron science
- **Leading organisation** STFC
- **Contributors** Simon Lambert - STFC, Esteban Gonzalez & Oscar Corcho - UP

Overview

The EOSC Interoperability Framework (EOSC IF) is a generic framework that can be used by all the entities participating in the development and deployment of EOSC, providing a common understanding of the requirements, challenges and recommendations that they should take into account, as well as a general set of principles on how these recommendations may be addressed.

The ambition of this use case is to create a **questionnaire to compile and understand semantic practices in the use case institution**. This first analysis will be done from the technical and semantic perspective defined in the EOSC IF and with dimensions inspired by its recommendations.

Challenges that need to be addressed

- ☐ Identify a collection of dimensions to recognize the implantation of the different recommendations given by the EOSC IF.
- ☐ Allow non expert users to Identify semantic elements in an institution.

Expected impact of the Use Case

Achieving this use case will allow to: i) understand semantic practices in an institution, ii) monitor the recommendations of the EOSC IF adopted and iii) improve the cataloging of semantic practices in other domains. Results can be used in WP4, specially in T4.2 Semantic artifact lifecycle and catalogs.

Results will be used to generate guidelines about usage of the components identified in the EOSC InteroperabilityFramework (EOSC IF) and that are, or will be, implemented and deployed in EOSC-related projects.

Expected outputs

A questionnaire to gather information about interoperability practices in the use case. This questionnaire will be sent to other use cases in order to detect semantic practices in other domains.

Reference Materials

- ☐ European Commission, Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., et al., EOSC interoperability framework : report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021

PIDs for instruments in photon and neutron facilities science. Use case by STFC

- **Key topic** PIDs
- **Scientific domain** Photon & Neutron science
- **Leading organisation** STFC
- **Contributor** Simon Lambert - STFC

Overview

The proposal of this use case is to advance the use of PIDs for instruments, such as instruments, devices, softwares and services, in the context of photon and neutron facilities (research infrastructures). This is an area of great interest in the community, for example assisting with assessing the impact of individual instruments. The RDA has a Working Group on Persistent Identification of Instruments that takes a cross-domain approach, but its specific application to PaN has not been fully explored.

In the field of photon and neutron science, an important European project ExPaNDS (https://www.expands.eu) has just come to an end after three years of work. ExPaNDS is the European Open Science Cloud (EOSC) Photon and Neutron Data Service, and is a collaboration between ten national Photon and Neutron Research Infrastructures (PaN RIs) as well as EGI. ExPaNDS was founded on the ideals of FAIR data, and has produced many high-quality results for the PaN community. There were also some areas of great interest to the community that were not yet pursued, including PIDs for instruments (in this context an instrument is a beamline on a facility, and its associated equipment used for conducting analyses of particular types). The FAIR-IMPACT project is an opportunity to take this further to advance FAIR in the photon and neutron community.

Challenges that need to be addressed

- ☐ Investigating through engagement with ISIS staff how PIDs for instruments would be used in practice by different stakeholders, what changes of practices might be needed, and what the implications are for the adoption of PIDs.
- ☐ Examining and elaborating use cases especially assessing impact of instruments.
- ☐ Appropriately representing versioning as instruments are modified over time.

Expected impact of the Use Case

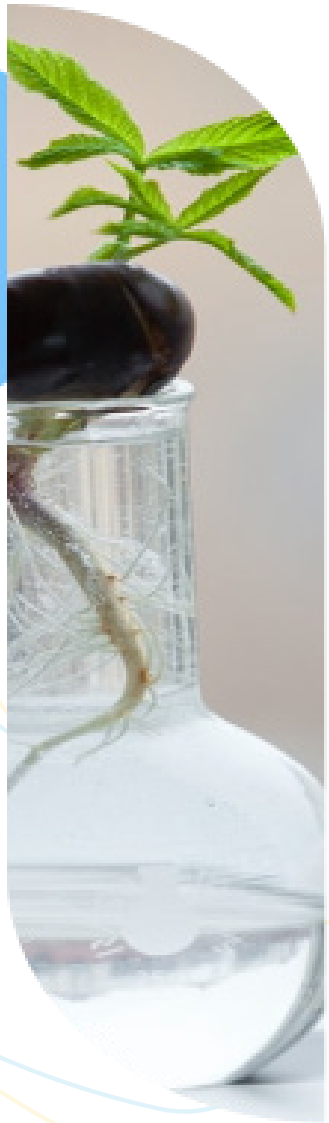
There is growing interest in assessing impact of research funding from different perspectives, including impact of whole facilities and of the instruments within them. The ability to refer to an instrument unambiguously through a PID would facilitate this assessment. Furthermore, it would enable credit to be clearly assigned to facilities staff responsible for particular instruments.

Expected outputs

It is not expected that within the FAIR-IMPACT project the work will lead to an implementation of PIDs for instruments. However it will result in a **better understanding of the processes around the generation and use of such PIDs and how they could fit into the practices of photon and neutron facilities**, as well as the use cases that motivate them and the implications for future implementation.

Reference Materials

- ☐ Persistent Identifiers as a crucial building block in the FAIR principles. Interview with Jessica Parland, CSC



Assessing FAIRness for Earth and Environmental Data. Use case by Dataterra and PANGAEA

- **Key topic** Interoperability | Metadata & Ontologies
- **Scientific domain** Earth and environmental sciences
- **Leading organisation** Dataterra (CNRS) | Uni Bremen (PANGAEA)
- **Contributors** Robert Huber - Uni Bremen (PANGAEA), Christelle Pierkot - Dataterra Research Infrastructure (CNRS)

Overview

In this use case, we intend to discuss potential FAIR metrics for the Earth and Environmental Sciences community (here: solid earth and oceans excluding atmosphere and biosphere) in collaboration with projects such as FAIR-EASE, Blue-Cloud and ENVRI-FAIR. For this purpose we want to analyze the FAIR habits of this community to find out if there are similarities in the use of e.g. identifiers, standards and vocabularies that justify deriving their own FAIR metrics from the existing FAIRsFAIR metrics. We will investigate existing technical interfaces for metadata exchange and use FAIR implementation profiles of relevant data archives.

Initiatives such as GEOSS or OGC have contributed in recent years to the fact that the level of standardization of earth and environmental science data repositories is generally quite high. In addition, there are a number of ongoing EU projects dealing with the implementation of the FAIR principles in this community. However, the Earth science community is quite diverse and there is no common understanding of FAIR so far. Using the example of some established research infrastructures from different fields of earth and environmental sciences, we will first investigate how homogeneous the use of FAIR implementation resources such as metadata standards or vocabularies is and try to develop metrics for the community or for appropriate parts of this community.

Challenges that need to be addressed

Within the earth and environmental sciences, there are a number of sub-communities that have already established functioning infrastructures for sharing data and metadata. Although there are many intersections in the use of standards and formats, there is no overview or appropriate database for this. Furthermore, there is a lack of information on how these communities intend to or already have implemented the FAIR principles. It will also be challenging to propose and agree on a common set of FAIR metrics for this community.

Expected impact of the Use Case

The main impact of the use-case will be to create synergies between ongoing initiatives on FAIR and to reuse existing preliminary research on FAIR from the FAIR-EASE, Blue-Cloud and ENVRI-FAIR projects. In particular, we will identify FAIR convergences between these projects and the data repositories involved in them. The resulting FAIR metrics will be used to derive tests in F-UJI that will be of significant help to the community in assessing the FAIRness of individual datasets.

Expected outputs

- ☐ Collection of FAIR Implementation Profiles from this community
- ☐ Draft FAIR metrics for the earth & environmental sciences
- ☐ Draft community specific FAIR assessments implemented in F-UJI

Providing documentation on harmonised and citable PIDs for subsets of protected data. Use case by EMBL-EBI

- Key topic PIDs
- Scientific domain Life science
- Leading organisation EMBL-EBI
- Contributors Henning Hermjakob - EMBL-EBI, Renato Juacaba Neto - EMBL-EBI, Josefine Nordling - CSC

Overview

European Bioinformatics Institute is Europe's largest provider of public biomolecular data resources. The institute is co-located with Elixir Hub and partnered up in many relevant EU projects, among others EOSC-Life, FREYA, and BY-COVID. This use case will explore PID practices in relation to complex data citation and sensitive data for the life science domain, and provide documentation on best practices to be adopted across domains. In addition to supporting life sciences, EMBL-EBI is increasingly also collaborating with other domains, e.g. social sciences in the context of Covid-19 research. EMBL-EBI provides consistent access to life science data by leveraging compact identifiers through the Identifiers.org resolution service. This service will be fine-tuned during the course of the project to ensure alignment with community FAIR practices and the broader EOSC context.

Description

The ambition of EMBL-EBI in the PID use case is to curate and update components of Identifiers.org. The updates will be aligned with community standards and needs in accordance with FAIR practices. Next step in this process will be to implement tombstone records in the Identifiers.org registry. Aligning the tombstone entry points following EOSC guidelines would ensure that all necessary entry points are included to enable verification of the contents' last residing place correctly and to ensure the persistence of the PID. However, some modifications might have to be made to the tombstone practices to avoid the system and its maintenance being overburdened. The intention is to create a proposal internally on how Identifiers.org tombstones should be implemented and then open it up for community review and discussion.

The ambition is also to study the possibility of facilitating the automated use of the Identifiers.org registry by including support for kernel information profiles and Digital Object interface protocol. This effort, however, requires some further deliberations to be able to define the exact approach and scope.

Identifiers.org resolution service provides PIDs to data hosted by several repositories. The API endpoints of the service can already provide metadata information on the referenced objects. The entries of the registry are actively curated by a team of specialists to ensure correct behaviour and normalised information. The service currently provides minimal support for PIDs intended for complex data in cases where the target repository identifies these objects with local IDs and provides a valid URL pattern for redirection to these objects. Providing additional support in such cases is another action point for the use case to address.

Challenges that need to be addressed

There is a need to find solutions to overcome some challenges related to PID practices, especially in cases where billions of data objects are included in a large number of resources, where data resources manage their own PIDs, where frequent data updates occur, and where there is a high barrier for adoption of global PID systems. PID practice alignment will require discussions to take place between relevant EOSC stakeholders, and this will require some time and effort.

Expected Impact of the Use Case

More efficient use of PIDs in sensitive data will benefit especially research within health and medicine, and thus have great positive societal and economic value. Documented complex data citation practices in relation to PIDs support the generation of more qualitative data and effects positively on the reusability and reproducibility of research data. Co-designing the PID practices through EOSC alignment provides a research environment which is responsive to the needs of the various research communities. EMBL-EBI is experienced in working with metadata standards, integration, discoverability, and display through its work with the Identifier.org service. Furthermore, EMBL-EBI is able to bring its valuable expertise from its partnership with a large international consortium, the European Bioinformatics Institute, which is a public biomolecular data resources provider. EMBL-EBI also has vast expertise gained from its partnership in many relevant EU projects, among others EOSC-Life, FREYA, and BY-COVID.

Expected outputs

Documentation on harmonised, citable PIDs for subsets of protected data. The use case will bring evolving Identifiers.org practices into a broader EOSC context and provide solutions to overcome some challenges related to PID practices.

Reference Materials

- EBI main website
- Identifiers.org Resolution Service

Encouraging and supporting researchers in producing FAIR computational workflows. Use case by University of Manchester

- **Key topic** PIDs
- **Scientific domain** Life science
- **Leading organisation** University of Manchester - UNIMAN
- **Contributors** Stian Soiland-Reyes - UNIMAN, Nick Juty - UNIMAN, Josefine Nordling - CSC

Overview

This use case is based around the University of Manchester's work with Persistent Identifiers in data production workflows via its involvement in the WorkflowHub - a registry of computational FAIR workflows. WorkflowHub is sponsored by the European RI Cluster EOSC-Life, the European Research Infrastructure ELIXIR and multiple EOSC projects (BY-COVID, BioDT and EuroScienceGateway). Its initial users were from within the life sciences working with COVID-19 workflows, but is now used by over 140 research groups and projects across disciplines.

The overall goal of this use case is to encourage and support FAIR Computational Workflows, where workflow systems help researchers in producing FAIR data and recording provenance of their analysis, but also where workflows themselves become FAIR scholarly objects in their own right, appear in the scholarly knowledge graph, gets cited in academic papers, and so on.

Workflows of any type (e.g. Galaxy, CWL, Nextflow, Jupyter Notebook) are registered in WorkflowHub from existing repositories like GitHub, or can be deposited as a direct upload. Metadata is extracted from the workflow and augmented by the user. This is archived in the form of an RO-Crate that also contains a snapshot of the executable workflow definition. The metadata uses JSON-LD and schema.org vocabulary for Dataset, together with a Bioschemas profile for computational workflows. WorkflowHub also uses the standard Common Workflow Language (CWL) as a way to describe the workflow structure and detailed annotations such as tools and containers required.

Workflows can be composed of various types of research objects which need to be formally and persistently identified to enable their reuse by other researchers. There are various challenges that need to be addressed resulting from the diverse types of identifiers of the various workflow components. In FAIR-IMPACT we are therefore following several strands to improve persistent identifiers for computational workflows:

- [ ] Improve and document explicit identification and linking (FAIR Signposting) from WorkflowHub to PIDs, metadata and RO-Crate downloads
- [ ] Enable an automatic request and recording function of Software Heritage identifiers (SWHID) when archiving a Git-based workflow
- [ ] Capture and expose PIDs for tools used by workflows (e.g. bio.tools, Bioconda) from Galaxy
- [ ] Generate location-independent identifiers (RFC6920) for data generated by workflow runs, potentially large/sensitive, to be included in workflow provenance
- [ ] Leverage RO-Crate to capture and propagate workflow provenance outputs and related PIDs
- [ ] Create RO-Crate profiles for capturing the provenance of an execution of a computational workflow with increasing granularity

Challenges that need to be addressed

There are several challenges related to multiple identifier types being used for various workflow components, such as software and data and their various inter-relations, for instance input and output files, and the details of the runs and types, such as workflow or process. Several different identifiers are already being used, e.g. EDAM terms and biotools identifiers, but inconsistently and are not propagated through workflow systems to the final results.

Additional challenges result from workflow identification across various repositories that may store the same workflow, or versions thereof, as well as from how to avoid identifier proliferation, for example where these offer various routes for exposure of workflows in the EOSC catalogue. It would also be useful to be able to identify components of workflows, and their related outputs. However, these may vary over time making it difficult to persistently refer back to.

Further challenges are manifested when considering other workflow systems, for instance Galaxy, which while highly used, makes identification of individual workflows difficult.

Expected impact of the Use Case

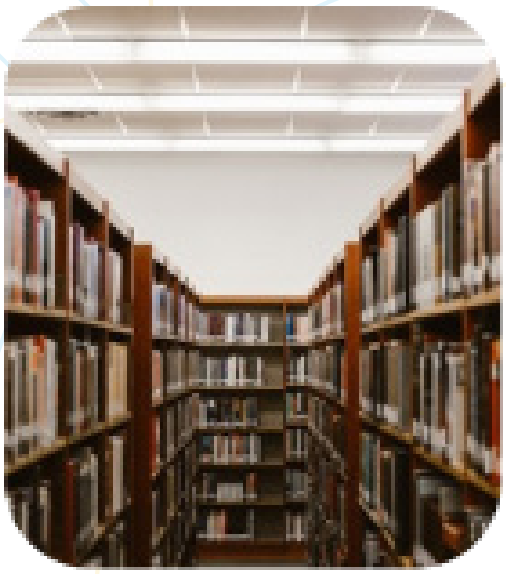
Through this use case work, we hope to achieve the necessary guidance for researchers to be able to get all relevant components of their computational workflows formally and persistently identified and thus be findable, reusable and citable by other researchers. This also makes it possible to capture provenance information. Furthermore, entire workflows can also become FAIR scholarly objects, equipped with their own identifiers and citing opportunities. All of these advancements significantly improve the traceability, transparency, reliability, and reproducibility of research.

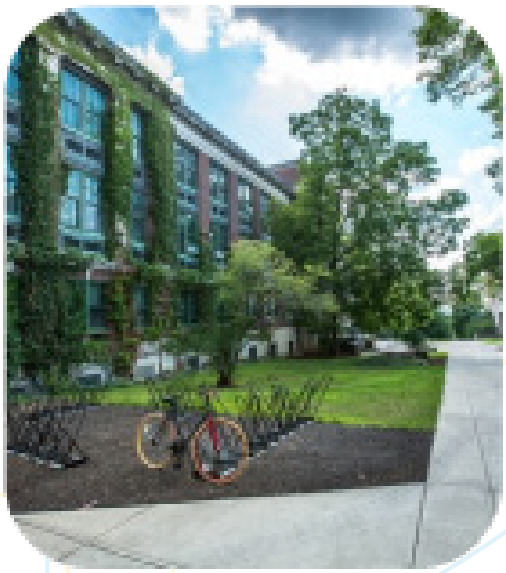
Expected outputs

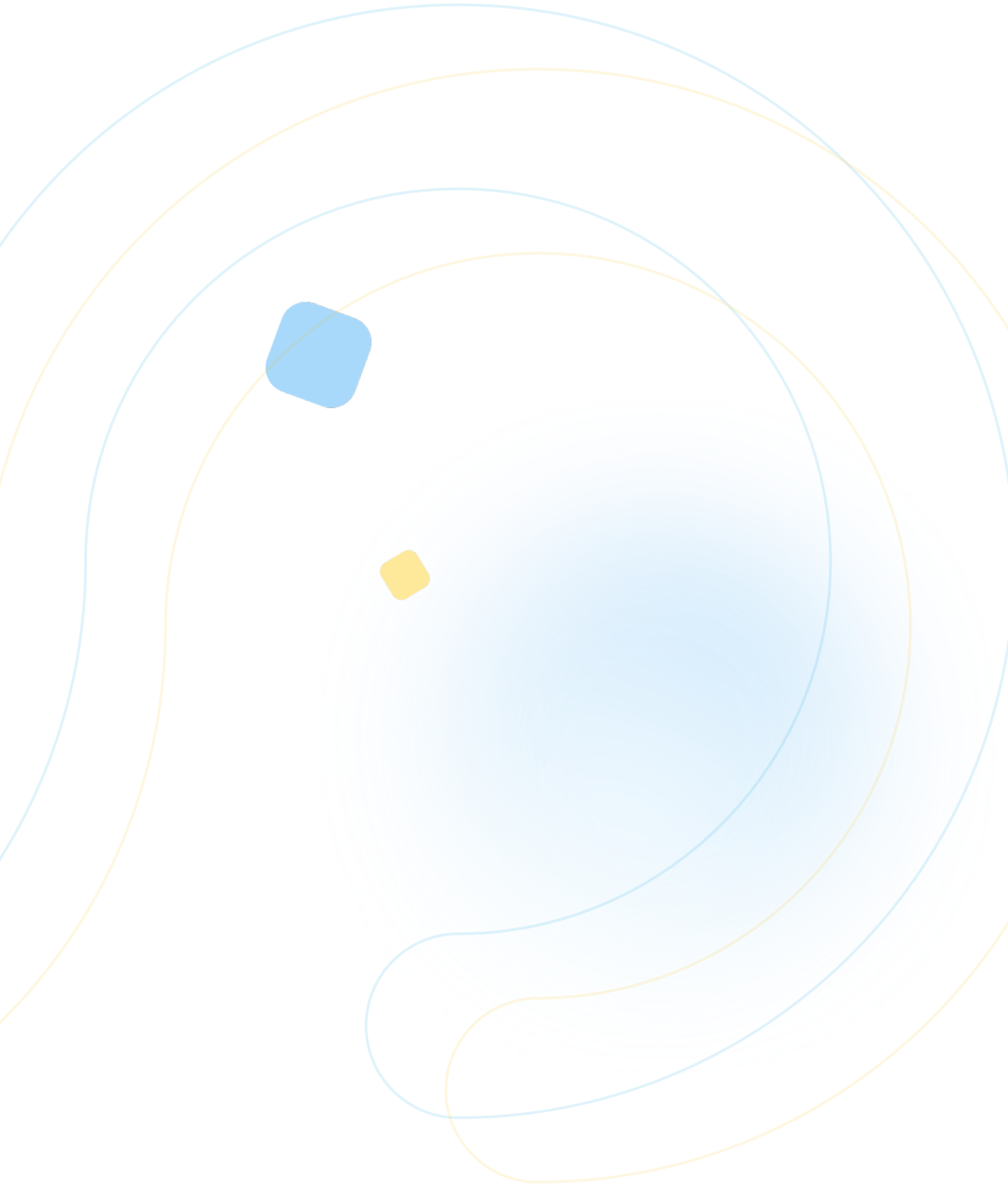
Improved documentation on identification of workflows and ways of linking workflows or parts of workflows to other research objects.

Reference Materials

- ☐ WorkflowHub page
- ☐ Research Object Crate (RO-Crate)







@fairimpact_eu

company/fair-impact-eu-project/

fair-impact.eu

Funded by the European Union