ENVRI^{PLUS} DELIVERABLE

D5.3

A DEFINITION OF THE ENVRIPLUS SEMANTIC LINKING FRAMEWORK AT CONCEPTUAL AND FORMAL LEVELS

WORK PACKAGE 5—REFERENCE MODEL GUIDED RI DESIGN

LEADING BENEFICIARY: UNIVERSITY OF AMSTERDAM

Author(s)	**Beneficiary/Institution**
Paul Martin, Zhiming Zhao	University of Amsterdam (UvA)
Barbara Magagna	Umweltbundesamt (EAA)

Accepted by: Paola Grosso

Deliverable type: OTHER

Dissemination level: PUBLIC

Deliverable due date: 30.4.2018/M36

Actual Date of Submission: 30.4.2018/M36

A document of the ENVRI^{PLUS} project—http://www.envriplus.eu
This document has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 654182.

Abstract

This document is part of the deliverable for **Task 5.3, "Semantic description and linking between RI architecture and technologies"**, which addresses the semantic linking cross-cutting activity of the ENVRIplus **Data for Science** theme.

The complete deliverable consists of three parts. The first part is *Open Information Linking for Environmental Science Research Infrastructures (OIL-E)*, a set of OWL ontologies (based on the ENVRI Reference Model) that acts as an upper ontology for different entities and activities attributable to environmental science research infrastructures and which serves as the basis for semantic linking in ENVRIplus between different semantic descriptions. The second part is an online knowledge base for the ENVRI cluster that provides access to information about research infrastructure design and resources, structured using OIL-E to demonstrate its use. The third and final part is this report, which summarises the activity of the task and provides both links to and essential context for the first two components.

Project internal reviewer(s)

Project internal reviewer(s)	Beneficiary/Institution
Keith Jeffery	National Environment Research Council (NERC)
Christian Pichot	Institut National de la Recherche Agronomique (INRA)

Document history

Date	Version
30/8/17	Outline of deliverable produced to assist work package planning.
30/3/18	First complete draft produced for internal review.
23/4/18	Revised draft based on reviewer comments.
30/4/18	Final draft for submission.

Document amendment procedure

Amendments, comments and suggestions should be sent to Paul Martin (p.w.martin@uva.nl).

Terminology

This deliverable uses terminology based on the ENVRI Reference Model [15], which is published online as an ontology: `http://www.oil-e.net/ontology/envri-rm.owl`.

Project summary

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonize policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance transdisciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

Table of contents

1 Introduction

Semantic linking is one of the three main cross-cutting activities in the 'Data for Science' theme of ENVRIplus, alongside the development and exploitation of the ENVRI Reference Model and the common architecture for interoperable services. The deliverable for Task 5.3 *'Semantic description and linking between RI architecture and technologies'* is a technology/specification deliverable that provides the main development outputs of the task. As such, there are provided three components:

1. **Open Information Linking for Environmental Science Research Infrastructures (OIL-E)**[1], a set of OWL ontologies based on the ENVRI Reference Model that acts as an upper ontology for different entities and activities attributable to environmental science research infrastructures, and which provides the framework for semantic linking based on contextualisation of different kinds of resource metadata provided by research infrastructures.
2. The **ENVRI Knowledge Base**[2], an online knowledge base for the ENVRI cluster that provides access to information about RI design and RI resources, based on OIL-E, and which serves as a practical demonstrator of the kind of semantic search and query that OIL-E can facilitate.
3. This report, which summarises the activity of the task and provides both links to and essential context for the first two components.

In this section, we review the motivation and requirements of the task, introducing the remainder of the report.

1.1 Motivation

Modern day environmental research depends on the collection and analysis of large volumes of data gathered via sensors, field observations, controlled experiments, simulation and modelling. The role of research infrastructures (RIs) in this context is to support researchers with datasets, platforms and tools that allow them to engage effectively with the available data, but no single research infrastructure can hope to encompass fully the whole research ecosystem [12], and so, as in the case of ENVRIplus, we have a host of different research infrastructures, each with their own areas of speciality, but nevertheless sharing many common scientific, technical, governance and political interests.

Meanwhile, researchers are being called upon to address societal challenges that are inextricably tied to the stability of our native ecosystems. These challenges are intrinsically interdisciplinary in nature, requiring collaboration across traditional disciplinary boundaries. The challenge therefore is to help researchers to freely and effectively interact with the full range of research assets potentially available to them across many research infrastructures, allowing them to collaborate and conduct their research more effectively than ever before.

Publishing metadata about the resources they offer online (indicating type, coverage, provenance, *etc.*) allows research infrastructures to advertise their offerings and allows researchers to browse and discover data (as well as models, sites, tools and other resources both virtual and physical) that could be useful to their research. In this space there are many standards, old and new, some *de facto* standards long adopted by particular communities, while others have achieved *de jure* status as recommendations by governing institutions. For example, in the geospatial area in which many environmental science research infrastructures are concerned there exist established standards such as ISOs 19115 [5] and 19139 [6], which form the basis for the INSPIRE[3] recommendation for spatial metadata in Europe. In practice however, the implementation of these and other standards can sometimes be partial or haphazard, with variations in how metadata elements are realised or terms applied.

Harmonisation of vocabulary and metadata between research infrastructures thus remains an on-going

[1] http://www.oil-e.net/ontology/

[2] http://oil-e.vlan400.uvalight.net/

[3] https://inspire.ec.europa.eu/

concern; it is the role of cluster projects such as ENVRIplus[4] to work to promote common semantic models, and this deliverable focuses specifically on one part of this effort, as embodied by Task 5.3 of the ENVRIplus Description of Action.

1.2 Task review

Within ENVRIplus, Task 5.3 is concerned with the design and prototyping of a **semantic linking framework** for linking descriptions of RI architecture with RI metadata, controlled vocabularies used within RI domains, and e-infrastructure service specifications. The objective of the task can be seen as one of making the **semantic landscape** of environmental science research infrastructure in Europe easier to survey and navigate, and so aid in the development of interoperable services that rely on the coherent use of metadata and controlled vocabularies. The task therefore sits between Tasks 5.2 *'RI characterisation and ENVRIplus reference model'* and 5.4 *'Interoperation based architecture design'*, taking the ENVRI Reference Model (ENVRI RM) to produce an upper ontology for environmental science research infrastructure and then providing a knowledge base for future architecture and service design. According to the ENVRIplus Description of Action, Task 5.3 is required to help:

1. Establish a common semantic framework for RIs, describing the landscape of semantic standards and vocabularies used by RIs, and harmonise between them where possible.
2. Provide tools for describing data, services and technologies semantically.
3. Provide a framework for keeping descriptions accurate even as technical details change.
4. Provide tools to link and map data and services provided by different RIs both for conceptual representation and for data processing.

We have addressed these requirements by developing OIL-E as a kind of architectural upper ontology for RI description, and building a prototype knowledge base utilising OIL-E to collect information about standards and vocabularies which can then be queried in order to establish commonalities and gaps. This knowledge base also allows us to collect information about mappings and other harmonisation activities, and allows us to leverage the existing range of tools and services for dealing with Semantic Web based Linked Data, as well as expedite our own tool development for specific use-cases. ENVRI RM has been used as the basis for OIL-E, providing a semantic 'hub' for relating different kinds of specification (whether of data, processes or other resources) via the activities specified in the reference model; this confers a number of advantages, including a multi-view perspective on RI design and a means to validate provenance traces based on their adherence to specific activity models[5].

1.3 Layout

The remainder of this document is laid out as follows:

Section 2 "Semantic linking in ENVRIplus" concerns the theory and principles of semantic linking as defined within the ENVRIplus project, and describes the relationship the semantic linking task has with other activities in the project.

Section 3 "Open Information Linking for Environmental RIs" provides an overview of the OIL-E framework and discusses how it has been applied to RIs within ENVRIplus. It also describes the ENVRI Knowledge Base and how it can be queried.

Section 4 "Further development" summarises this report and discusses avenues for further development of the semantic linking work.

[4] http://www.envriplus.eu/

[5] This latter case will be investigated more fully for Deliverable 8.6 *"Data provenance and tracing for environmental sciences: prototype and deployment"*.

2 Semantic linking in ENVRIplus

The semantic linking framework of ENVRIplus is intended to guide the harmonisation of semantics across environmental science research infrastructures by providing contextualisation and a standard upper ontology for the different kinds of entities and activities commonly found in those infrastructures. Notably, it is not itself a catch-all solution to the problem of *mapping* between different metadata schemes used by RIs, for which there has been considerable effort already expended and for which considerable effort will still be expended. There exist many tools and frameworks for handling such mappings and a great body of research—our concern is rather with providing some baseline support for analysing the diversity of such schemes and mappings where they exist, and helping research infrastructure developers to focus their efforts on specific problem areas.

In this section we define the scope of the semantic linking framework at conceptual level and its relationship with Work Package 5 of the ENVRIplus project.

2.1 Building the linking framework in ENVRIplus

The semantic linking task in ENVRIplus was carried out in the context of the wider activities in Work Package 5, following on from the requirements gathering and modelling activities of Tasks 5.1 and 5.2 respectively, and based on the vision originally articulated in [13]. The basic procedure for the task can be summarised as follows:

1. We collected information from environmental RIs and communities: requirements, technologies and the current state of the art (Task 5.1).
2. We used these requirements to refine ENVRI RM (Task 5.2), which importantly provides a common vocabulary for describing various kinds of component and activity deployed in RIs.
3. Concurrently, we also began gathering information about community standards, semantic resources and vocabularies for other aspects of environmental research, data and process specification (Task 5.3).
4. We updated OIL-E, the OWL-based ontological adaptation of ENVRI RM to improve its usefulness as an upper ontology for RI architecture which we can use to link various standards and specifications used by different RI entities.
5. With OIL-E, we are now mapping the semantic landscape of environment science, encoding information about the different RIs, their component parts and their constituent processes, as well as associating standards and software to different entities where appropriate.
6. This has resulted in the creation of a knowledge base to contain all the OIL-E data, to provide an interface via which architects and developers can investigate descriptions of RIs, and to provide connecting links with external linked data where made available by RIs.
7. For the final phase of Task 5.3 and beyond, we are focusing on capturing mapping information for bridging between OIL-E and other RI knowledge representations, and on tools for semantic modelling and discovery using OIL-E.

The vision and interlinking between reference model and open information linking is illustrated by Figure 1, variants of which have been used in dissemination materials.

2.2 Reference model guided semantic linking

As with all activities within Work Package 5 *"Reference model guided RI design"*, the semantic linking framework has been developed based on the ENVRI Reference Model (ENVRI RM) [15].

ENVRI RM[6] is constructed using the Open Distributed Process (ODP) [4] for modelling complex distributed systems; ODP requires the modelling of a system from five different viewpoints (enterprise,

[6] http://envri.eu/rm

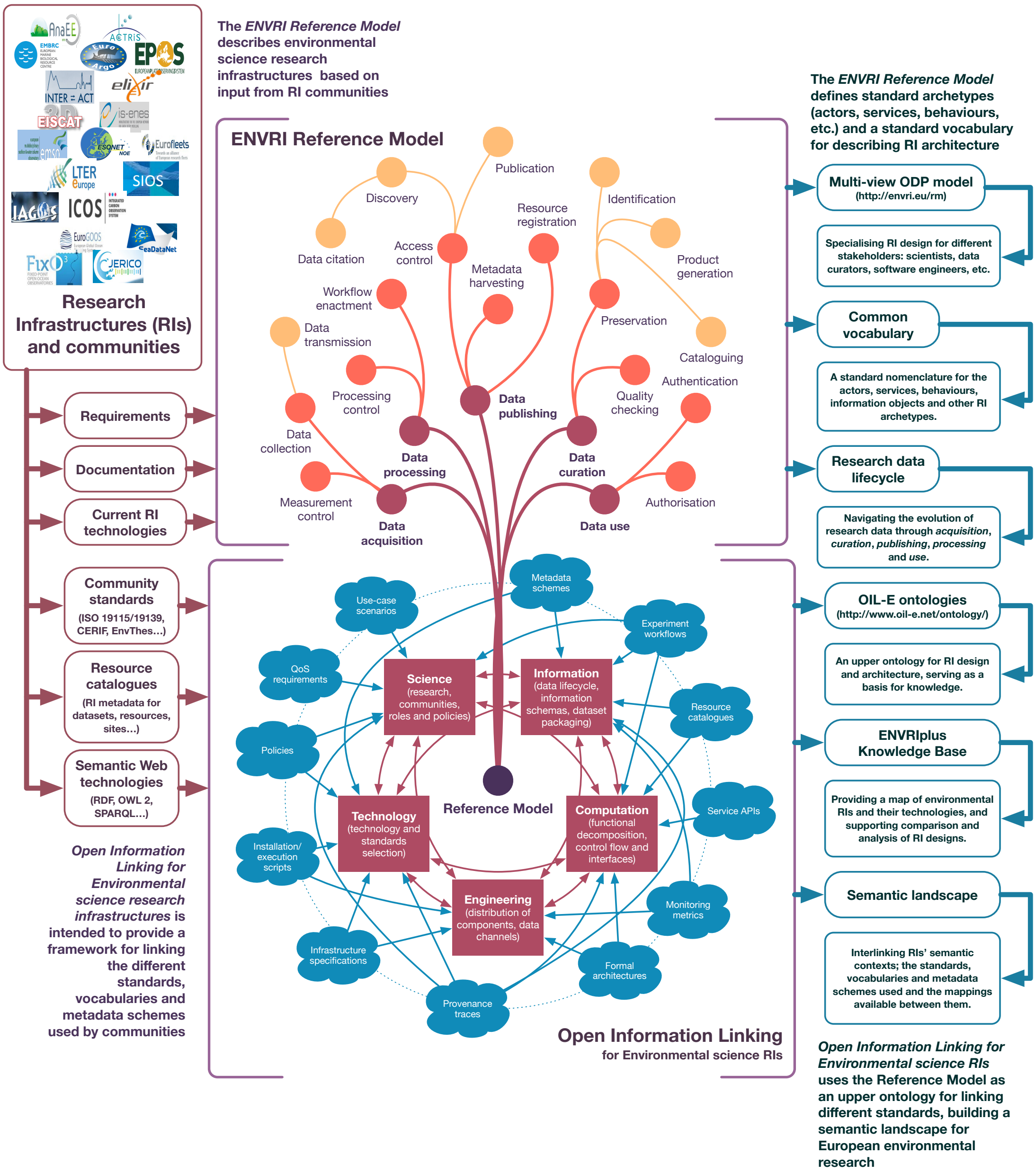

Research Infrastructures (RIs) and communities
The ENVRI Reference Model describes environmental science research infrastructures based on input from RI communities
ENVRI Reference Model
Publication
Discovery
Identification
Resource registration
Access control
Data citation
Product generation
Metadata harvesting
Workflow enactment
Preservation
Data transmission
Cataloguing
Authentication
Processing control
Quality checking
Data publishing
Data collection
Data processing
Data curation
Measurement control
Authorisation
Data acquisition
Data use
Requirements
Documentation
Current RI technologies
Community standards
(ISO 19115/19139, CERIF, EnvThes…)
Resource catalogues
(RI metadata for datasets, resources, sites…)
Semantic Web technologies
(RDF, OWL 2, SPARQL…)
Open Information Linking for Environmental science research infrastructures is intended to provide a framework for linking the different standards, vocabularies and metadata schemes used by communities
Metadata schemes
Use-case scenarios
Experiment workflows
QoS requirements
Science
(research, communities, roles and policies)
Information
(data lifecycle, information schemas, dataset packaging)
Resource catalogues
Policies
Service APIs
Reference Model
Technology
(technology and standards selection)
Computation
(functional decomposition, control flow and interfaces)
Installation/ execution scripts
Engineering
(distribution of components, data channels)
Monitoring metrics
Infrastructure specifications
Formal architectures
Provenance traces
Open Information Linking
for Environmental science RIs
The ENVRI Reference Model defines standard archetypes (actors, services, behaviours, etc.) and a standard vocabulary for describing RI architecture
Multi-view ODP model
(http://envri.eu/rm)
Specialising RI design for different stakeholders: scientists, data curators, software engineers, etc.
Common vocabulary
A standard nomenclature for the actors, services, behaviours, information objects and other RI archetypes.
Research data lifecycle
Navigating the evolution of research data through acquisition, curation, publishing, processing and use.
OIL-E ontologies
(http://www.oil-e.net/ontology/)
An upper ontology for RI design and architecture, serving as a basis for knowledge.
ENVRIplus Knowledge Base
Providing a map of environmental RIs and their technologies, and supporting comparison and analysis of RI designs.
Semantic landscape
Interlinking RIs' semantic contexts; the standards, vocabularies and metadata schemes used and the mappings available between them.
Open Information Linking for Environmental science RIs uses the Reference Model as an upper ontology for linking different standards, building a semantic landscape for European environmental research

Figure 1: The vision of semantic linking in the ENVRIplus project.

information, computation, engineering and technology) with the correspondences between the five resulting views ensuring their mutual validity. This viewpoint-based approach provides clarity to each 'facet' of the end model by reducing the number of competing elements to only those that match a particular set of concerns (such as the flow of information through the system), while still retaining the aggregate complexity needed to model any substantive distributed system. ENVRI RM uses three of the five views prescribed by ODP to capture the generic aspects common across all RIs: *enterprise* (renamed *science* in respect to the subject area), *information* and *computation*. ENVRI RM then uses the *engineering* and *technology* viewpoints to explore the more specific solutions and design patterns observed as being used by current RIs for the generic components prescribed in the three former views. The ENVRI RM ontology within the OIL-E semantic framework captures all the objects defined as of version 2.2 of ENVRI RM along with their essential relations and indeed extends those relations to better draw links between entities and facilitate a greater range of queries.

The use of methodology such as ODP helps guide the software engineering process by recognising the existence of different kinds of stakeholder in system development with different primary concerns, and providing a multi-faceted modelling context that addresses each while maintaining an overall coherent specification. This benefits all parties by providing distinct specifications of each facet of the system that adequately reveal the key characteristics of the system from one perspective while ignoring details that are less relevant to that perspective but which will be reliably addressed in one of the others. A similar benefit can be obtained for ontologies, where simple decompositions of systems with one particular perspective in mind often produce clearer, more usable ontologies in practice. Conversely, trying to do 'too much' within the framework of a single ontology can make it more difficult to use and more likely to contain errors or points of contention.

OIL-E's use of the Reference Model for ODP (RM-ODP) is not wholly new; RM-ODP has been expressed in ontology form as early as 2001 [18]. Applications of ODP have been studied extensively [9], and ODP has been applied to the design of various kinds of infrastructure, including in the Internet of Things (IoT) / Smart Cities context [16]. The applicability of ODP (a standard published in the previous century) to modern service-oriented architectures and Cloud has also been addressed [7].

The multi-viewpoint approach intrinsic to ENVRI RM and inherited from ODP informs the design of OIL-E in many ways however. Most notably, each viewpoint essentially provides its own micro-ontology, instances of the concepts defined in which can then be related to concepts in other views via correspondences (as defined by ODP). This allows OIL-E to operate as a 'hub' ontology, whereby specifications created by extending one view (*e.g.* computation) can be used to dictate requirements on another view (*e.g.* information).

2.3 Semantic linking scenarios

The semantic linking framework of ENVRIplus is conceived to support a number of different semantic linking scenarios. For each of these scenarios, we have defined the problem and conducted some investigation based on the use of OIL-E.

Contextualisation This scenario is one where different kinds of entity with semantic connotations (datasets, metadata schemes, vocabularies, *etc.*) are described within an OIL-E dataset and classified in terms of ENVRI RM, where possible with direct links to their respective access points (*e.g.* URLs for querying and retrieving metadata) or specifications (*e.g.* landing pages for ontologies) as appropriate. Figure 2 provides an example of such contextualisation in the context of data acquisition.

The ENVRI Knowledge Base is the primary vehicle for exploring this kind of semantic linking: by collecting information about different RIs using the terminology of ENVRI RM and the framework of OIL-E, we can explore and visualise the resulting network of information and perform comparative analyses.

Classification This scenario concerns using OIL-E as a controlled vocabulary for the classification

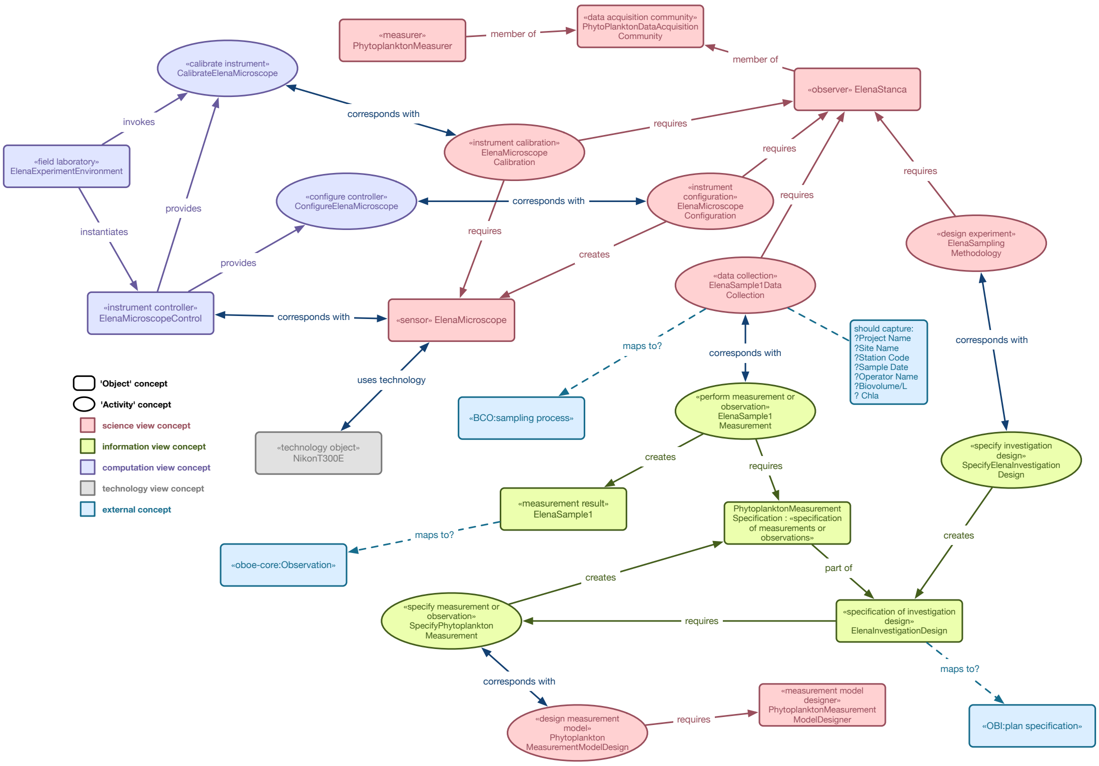

«measurer» PhytoplanktonMeasurer
member of
«data acquisition community» PhytoPlanktonDataAcquisition Community
member of
«observer» ElenaStanca
«calibrate instrument» CalibrateElenaMicroscope
corresponds with
«instrument calibration» ElenaMicroscope Calibration
requires
requires
requires
requires
invokes
«field laboratory» ElenaExperimentEnvironment
provides
instantiates
«configure controller» ConfigureElenaMicroscope
corresponds with
«instrument configuration» ElenaMicroscope Configuration
requires
creates
«design experiment» ElenaSampling Methodology
provides
«instrument controller» ElenaMicroscopeControl
corresponds with
«sensor» ElenaMicroscope
«data collection» ElenaSample1Data Collection
should capture:
?Project Name
?Site Name
?Station Code
?Sample Date
?Operator Name
?Biovolume/L
? Chla
maps to?
corresponds with
corresponds with
uses technology
«BCO:sampling process»
«perform measurement or observation» ElenaSample1 Measurement
«specify investigation design» SpecifyElenaInvestigation Design
'Object' concept
'Activity' concept
science view concept
information view concept
computation view concept
technology view concept
external concept
«technology object» NikonT300E
creates
requires
«measurement result» ElenaSample1
PhytoplanktonMeasurement Specification : «specification of measurements or observations»
creates
maps to?
«oboe-core:Observation»
part of
creates
«specify measurement or observation» SpecifyPhytoplankton Measurement
requires
«specification of investigation design» ElenaInvestigationDesign
maps to?
corresponds with
«measurement model designer» PhytoplanktonMeasurement ModelDesigner
«design measurement model» Phytoplankton MeasurementModelDesign
requires
«OBI:plan specification»

Figure 2: Modelling the acquisition of data regarding phytoplankton across multiple views in OIL-E.

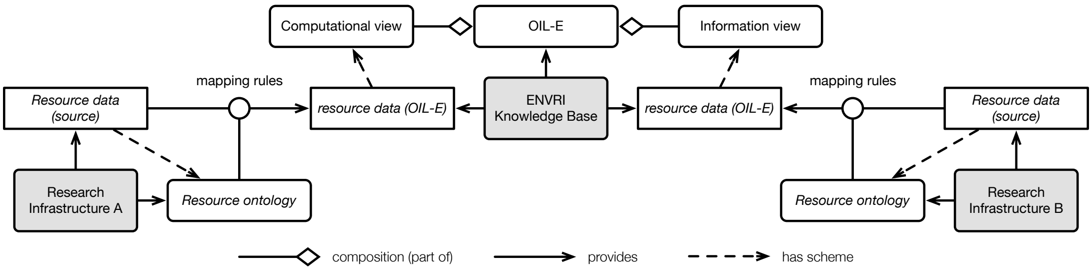


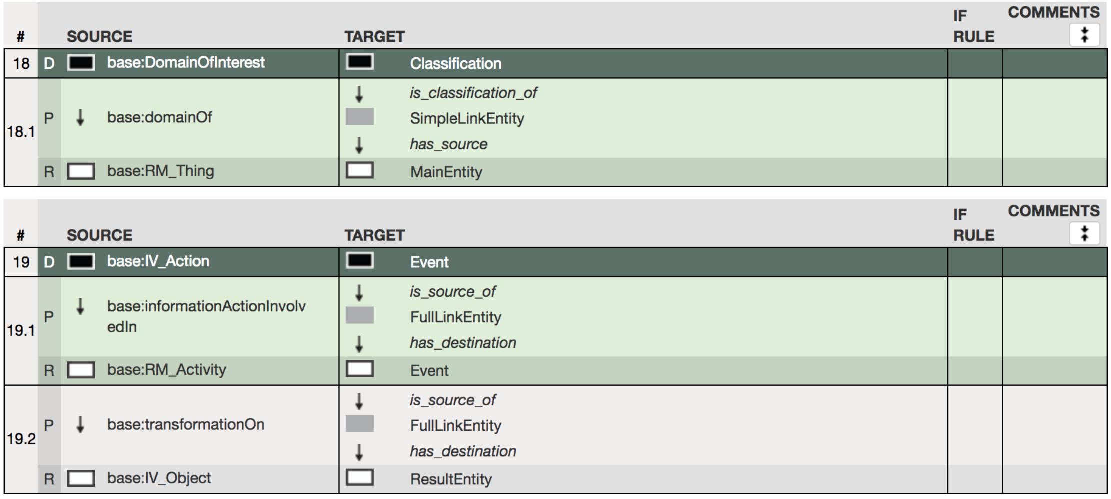
Figure 3: Using mappings from other metadata schemes into OIL-E.

of entities (datasets, services, tools and other resources that might have metadata records associated with them) within another metadata scheme.

The use of OIL-E as a classification scheme is being investigated in the context of CERIF [8], a scheme for research information systems that is also a recommendation within ENVRIplus for the ENVRI canonical metadata standard (see D8.3 'Interoperable cataloging and harmonization for environmental RI projects: system design'). A notable feature of CERIF is how it separates its semantic layer from its primary entity-relationship model. Most CERIF relations are semantically agnostic, lacking any particular interpretation beyond identifying a link between entities. Almost every entity and relation can be assigned though a classification that indicates a particular semantic interpretation, allowing a CERIF database to be enriched with concepts from external semantic models (such as OIL-E).

Semantic mapping The full semantic mapping scenario concerns cases where some data is to be fully translated into the OIL-E schema from some other schema or vice versa. In this case, a mapping scheme between OIL-E and the other target is needed to facilitate the mapping process.

A mapping agent will access the source of the data, apply the mapping, and record the mapped



#		SOURCE	TARGET	IF RULE	COMMENTS
18	D	■ base:DomainOfInterest	■ Classification		
18.1	P	↓ base:domainOf	↓ *is_classification_of* ▨ SimpleLinkEntity ↓ *has_source*		
	R	□ base:RM_Thing	□ MainEntity		

#		SOURCE	TARGET	IF RULE	COMMENTS
19	D	■ base:IV_Action	■ Event		
19.1	P	↓ base:informationActionInvolvedIn	↓ *is_source_of* ▨ FullLinkEntity ↓ *has_destination*		
	R	□ base:RM_Activity	□ Event		
19.2	P	↓ base:transformationOn	↓ *is_source_of* ▨ FullLinkEntity ↓ *has_destination*		
	R	□ base:IV_Object	□ ResultEntity		

Figure 4: Example of mapping rules generated in 3M: OIL-E to CERIF.

data in some target resource (*e.g.* the ENVRI knowledge base, as shown in Figure 3 for data from two different RIs, mapped into OIL-E for two different viewpoints). Such a full mapping is then independent of the original source, but this also means that the data may need to be updated at times if the source changes, and a process needed to trigger such updates or to regularly poll the source for changes. In addition to the classification activity, described earlier, we have also explored mapping between OIL-E and CERIF datasets directly. Figure 4 shows a snapshot of OIL-E/CERIF semantic linking (in this case defining mapping rules from OIL-E to CERIF) using the 3M Mapping Memory Manager[7].

Semantic bridging This scenario concerns cases where linkable data are present and accessible as linked data [2] online, and thus the requirement of semantic linking is not to replicate the existing accessible data, but to generate the necessary bridging data that will allow distributed reasoners to navigate between OIL-E and the target.

Essentially, mapping data is used to generate the additional classification data needed to relate entities defined in the target data source to concepts in OIL-E. That classification data can be added to an intermediary knowledge resource (*e.g.* the ENVRI Knowledge Base) and provide enough information to allow query systems to ‘follow through’ to the original source. An example of this kind of activity is bridging between the ENVRI Knowledge Base and provenance traces implemented in PROV-O [14]. We can use SHACL rules [10] to describe how to generate additional RDF triples classifying entities in the provenance graph using OIL-E, and then automatically assert them into the knowledge base, with pointers back to the provenance data. This allows for possible validation of the provenance graph based on OIL-E definitions, and allows for a distributed query broker to potentially access the provenance data directly via the bridging data in the knowledge base.

[7] https://github.com/isl/Mapping-Memory-Manager

3 Open Information Linking for Environmental RIs

Open Information Linking for Environmental RIs (OIL-E) is intended to assist with semantic harmonisation between different research infrastructures by providing an upper ontology for describing research infrastructure components and activities based on the archetypes defined by the ENVRI Reference Model.

The purpose of OIL-E is to provide a framework by which the semantics of different controlled vocabularies can be studied in order to allow translation and reasoning over heterogeneous datasets. This entails:

- Comparing different concept models for modelling research assets and data, and identifying commonalities and gaps.
- Building generic tools using existing technologies to handle the search and mapping of models related to RI architecture and specification.

The linking component of OIL-E glues concepts both inside ENVRI RM and between ENVRI RM and external vocabularies. In the latter case, external models can be classified in terms of ENVRI RM in order to help map the landscape of RI-related standards and models. The ENVRI RM ontology only contains a limited set of vocabularies derived from common RI functionality and design patterns, so linking the ENVRI RM ontology with external models will also enable domain-specific extensions to ENVRI RM itself. The internal correspondences between the different ENVRI RM views can potentially be used to indirectly draw associations between concept models with quite different focus areas (*e.g.* data versus services).

3.1 OIL-E ontologies

The OIL-E ontologies are implemented in OWL version 2 [17]. The current versions of the OIL-E ontologies can always be found at the following URL:

```
http://www.oil-e.net/ontology/
```

The indicated URL also provides links to previous major versions of the ontologies for reference; all known current uses of the ontology are based on the most recent ontology files however.

The OIL-E ontologies serve a number of purposes:

1. To capture notions of research infrastructure from perspectives of user interaction, data evolution, computation and physical infrastructure (*e.g.* sites and sensors).
2. To clearly separate these different views on infrastructure, but also establish their correspondences.
3. To capture the most significant interactions between different actors and resources, and to model the information objects that are both used and produced during such interactions.
4. To help establish the relationships between other standards and vocabularies in terms of the facets of infrastructure, resources and activity to which they apply.

The core of OIL-E is a single ontology, *oil-base*, which provides a set of abstract classes derived from the most common elements used by ENVRI RM from all the five viewpoints and acts as the basis for all further OIL-E extensions. It should be noted that *oil-base* is not a general-purpose upper ontology for describing scientific phenomena like BFO [1], but rather is a means to describe architectural and procedural aspects of RIs. Figure 5 illustrates *oil-base*'s root concept hierarchy and its subdivision into the top-level concepts for the five ENVRI RM viewpoints.

The simple categorisation of objects, activities and attributes independent of particular viewpoints is used as the basis for defining exclusion sets and restrictions on object properties while allowing certain concepts to exist in multiple views and to define many generic properties for use in and across multiple views. The distribution of specific concepts to specific views is then done via inference using *classifier* concepts. By default the five RM-ODP viewpoints are used, but alternative viewpoints can

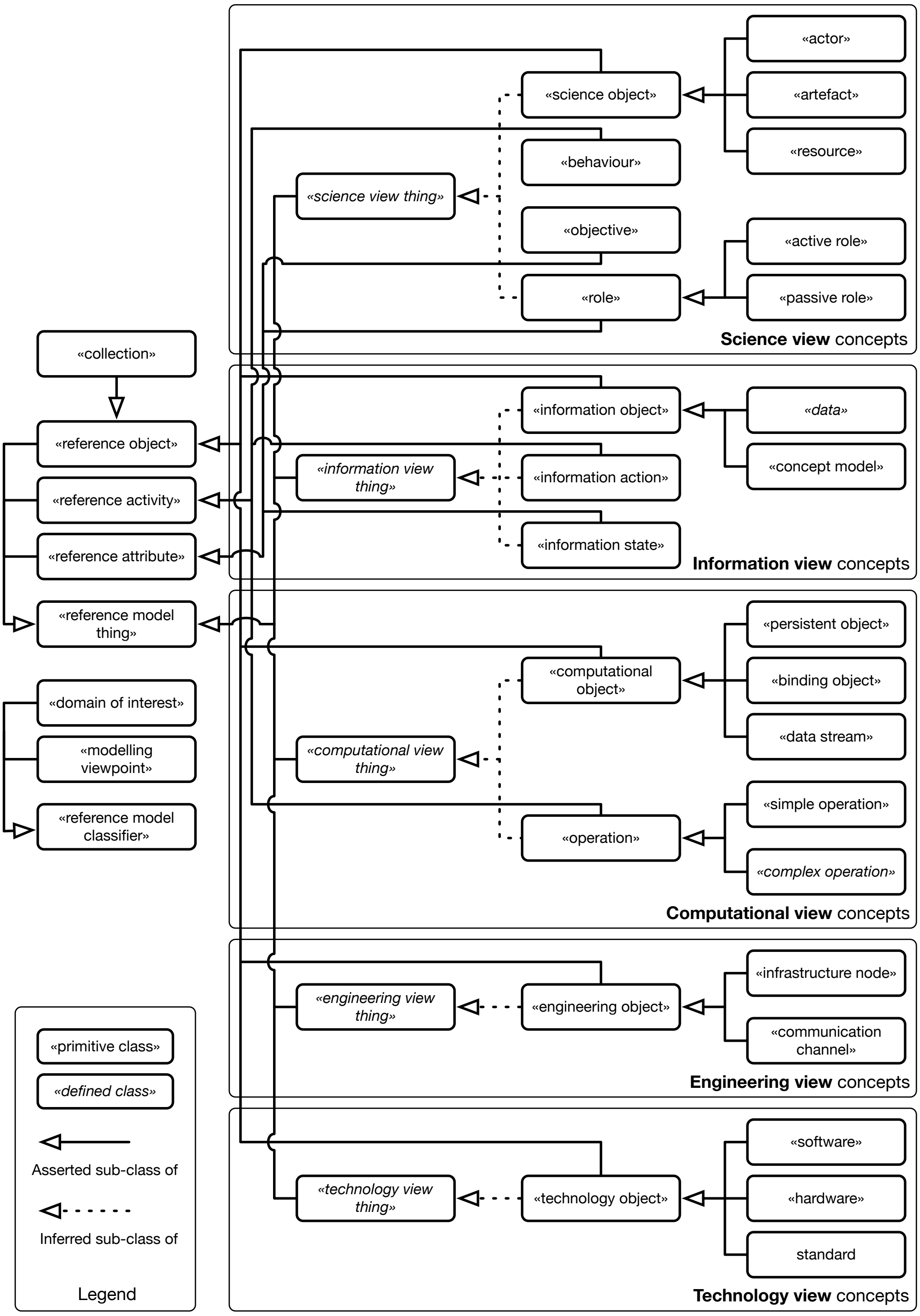


Figure 5: The base hierarchy of OIL-E across five viewpoints. Some concepts have been omitted for clarity.

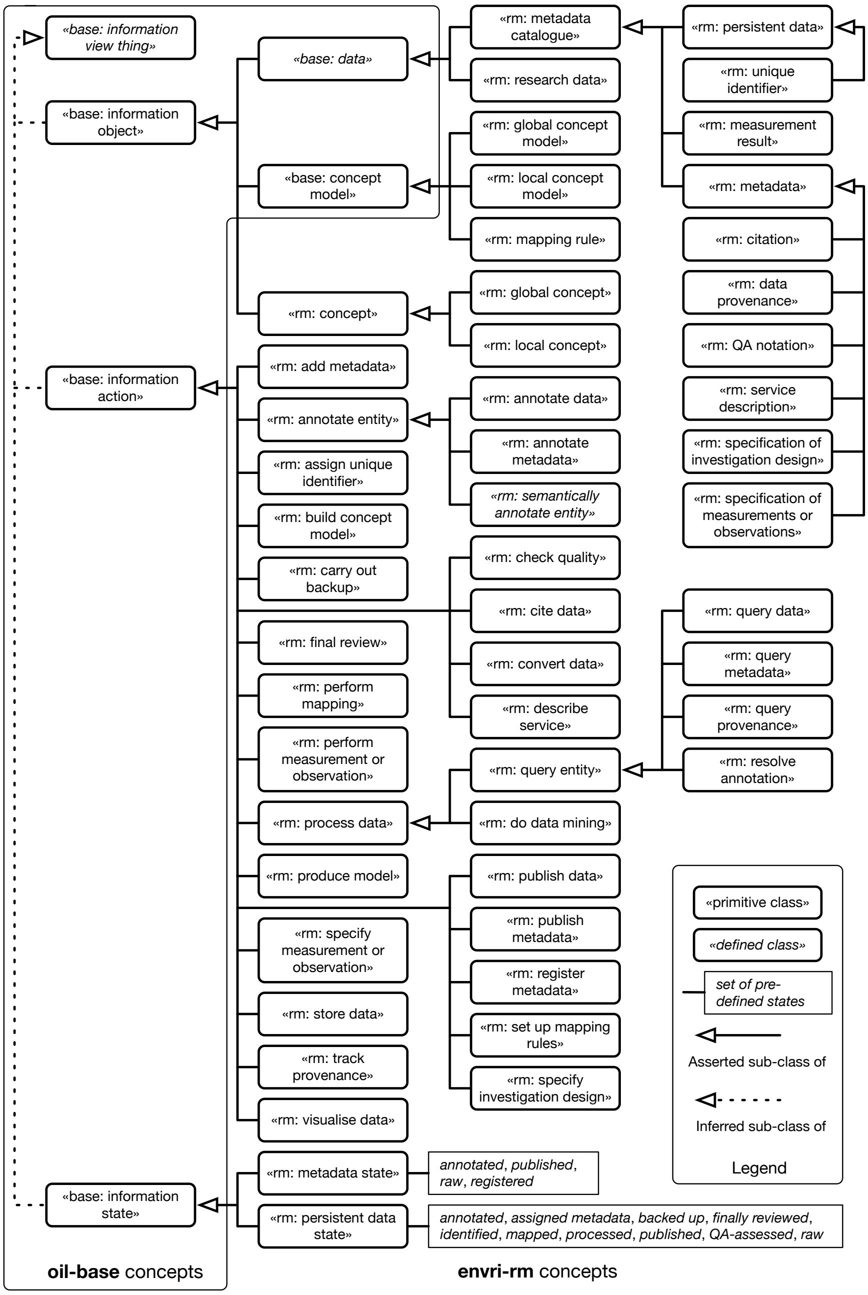
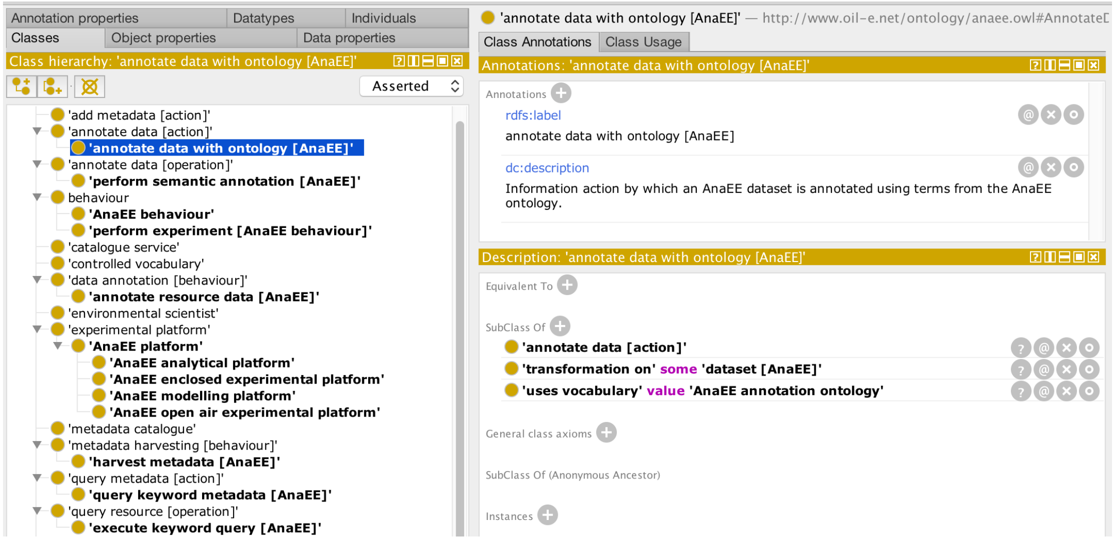


Figure 6: The hierarchy of OIL-E's information viewpoint, listing the information archetypes defined by ENVRI RM. Some concepts have been omitted for clarity.



Class hierarchy: 'annotate data with ontology [AnaEE]'

- 'add metadata [action]'
- 'annotate data [action]'
  - **'annotate data with ontology [AnaEE]'**
- 'annotate data [operation]'
  - **'perform semantic annotation [AnaEE]'**
- behaviour
  - **'AnaEE behaviour'**
  - **'perform experiment [AnaEE behaviour]'**
- 'catalogue service'
- 'controlled vocabulary'
- 'data annotation [behaviour]'
  - **'annotate resource data [AnaEE]'**
- 'environmental scientist'
- 'experimental platform'
  - **'AnaEE platform'**
    - **'AnaEE analytical platform'**
    - **'AnaEE enclosed experimental platform'**
    - **'AnaEE modelling platform'**
    - **'AnaEE open air experimental platform'**
- 'metadata catalogue'
- 'metadata harvesting [behaviour]'
  - **'harvest metadata [AnaEE]'**
- 'query metadata [action]'
  - **'query keyword metadata [AnaEE]'**
- 'query resource [operation]'
  - **'execute keyword query [AnaEE]'**

'annotate data with ontology [AnaEE]' — http://www.oil-e.net/ontology/anaee.owl#AnnotateD

Annotations: 'annotate data with ontology [AnaEE]'

rdfs:label
annotate data with ontology [AnaEE]

dc:description
Information action by which an AnaEE dataset is annotated using terms from the AnaEE ontology.

Description: 'annotate data with ontology [AnaEE]'

SubClass Of
- **'annotate data [action]'**
- **'transformation on' some 'dataset [AnaEE]'**
- **'uses vocabulary' value 'AnaEE annotation ontology'**

Figure 7: Extending OIL-E to model components and activities of AnaEE: modelling data annotation specifically in AnaEE (provisional data shown in Protégé [https://protege.stanford.edu/]).

be defined to meet the architectural needs of different infrastructures that are not perfectly catered for by the ODP approach. A particular problem of ODP is that it does not define concepts for linking outwards to other systems, however this does not prevent such concepts being defined in OIL-E.

The *oil-base* ontology is extended by the *envri-rm* ontology, which takes the archetypes defined by ENVRI RM and defines them all as concept classes with the prescribed relations between them, allowing better contextualised classification of RI entities like data products, data product catalogues, instruments, workflow components, *etc.* For example, Figure 6 shows the set of concepts defined by ENVRI RM for the information viewpoint. Similar hierarchies exist for each of the other viewpoints supported, with the science, information and computational viewpoints containing the most detail (in line with ENVRI RM and the focus on generic aspects of research infrastructure construction).

oil-base and *envri-rm* are the basis set for OIL-E; other ontologies or instance data sets extend either ontology to serve particular purposes, such as to define extensions for specific RIs or to map to other standards as described below.

3.2 Using OIL-E to model RIs and research activities

Information specific to individual RIs is created by extending *envri-rm* with concepts particular to the RI, as well as providing specific instances of RM archetypes implemented by the RI, for example as shown in Figure 7. These concepts may apply to any of the views defined by ENVRI RM, with OIL-E providing the vocabulary necessary to relate concepts within and between views. The technology viewpoint of OIL-E allows for the identification of specific technologies (software, standards, *etc.*) linked to types and instances of RI datasets and services, which can then be brought together to compare technology use between different RIs, for example as shown in Figure 8. We can also identify the context in which such technologies are used (*e.g.* for what kind of dataset or to implement what service), and provide information about where such technologies can be acquired.

Notably, OIL-E conflates two major classes of information regarding RIs: *schematic* information, about the general 'kinds' of element found in a given RI; and *instance* information, about actual services, datasets, technologies currently found in an RI. For example, "ICOS Level 1 data" concerns a general class of dataset found in the ICOS Carbon Portal, the properties of which apply to all instances of such datasets, while there may also be an individual defined for a specific Level 1 data product in ICOS. A description of the former is schematic information, while a description of the latter is instance information. In practice, most OIL-E data so far produced is a mix of schematic information and instance data about *invariant* parts of RIs; for example the "ICOS Carbon Portal"

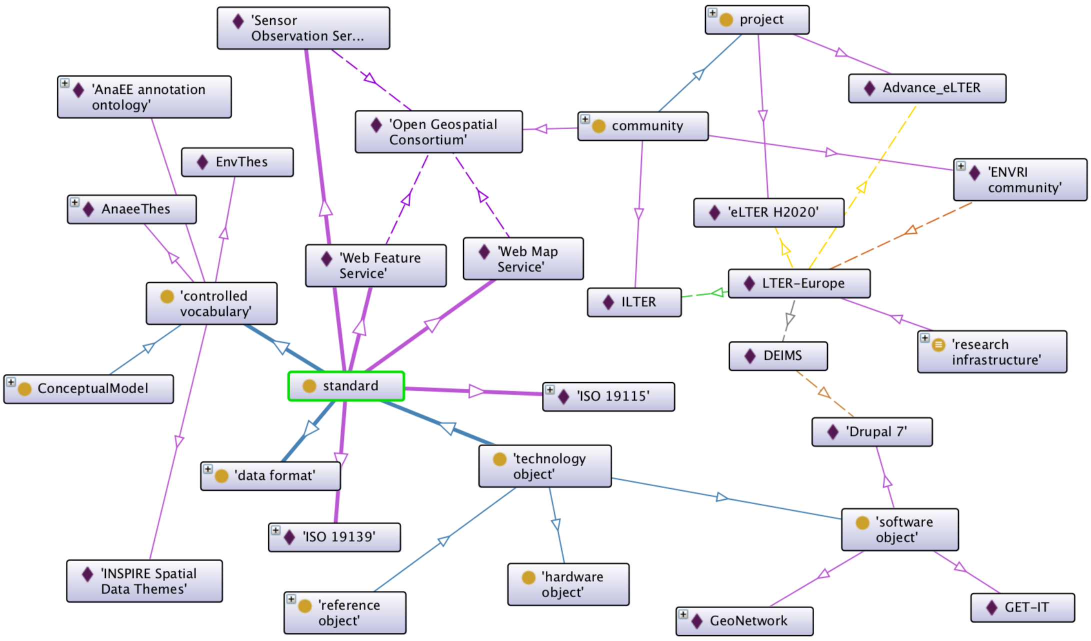


Figure 8: Linking technologies and standards: the use of different technologies by different RIs can be explored via the knowledge graph generated using RI data in OIL-E (sample data shown using OntoGraf [https://github.com/protegeproject/ontograf]). Nodes marked with a yellow circle are classes while nodes marked with a purple diamond are individuals; for clarity and brevity, not all nodes have been expanded.

is a specific component of the ICOS RI rather than a class of component and thus is instance data, as is the metadata standard "ISO 19139" used for data produced by many RIs (though there may also be a class of "ISO 19139 compliant datasets"). Whether schematic or instance information, the combination of this data provides a description for an RI that can be used to classify not only persistent RI entities such as datasets and services, but also transient events, which (for example) allows such extensions of OIL-E to be used to classify or validate provenance traces.

Over the course of the ENVRIplus project, a number of meetings have been held with RI developers in order to acquire information about their respective RIs that can then be modelled using OIL-E. All information gathered has been uploaded to the ENVRI Knowledge Base, and can be accessed and explored via its SPARQL endpoint.

3.3 ENVRI Knowledge Base

A key outcome in ENVRIplus resulting from the creation of OIL-E is a knowledge base for information about the research infrastructures in the ENVRI community and some of their activities. The need for such a knowledge base was motivated by the need to better map the semantic landscape of environmental science RIs in Europe, and in particular to gather information about the different metadata schemes, ontologies, thesauri and other controlled vocabularies used by RIs specifically in terms of their use in RIs (as opposed to simply providing another ontology portal).

The knowledge base also serves as a useful demonstration platform for OIL-E, allowing interested parties to directly interact with OIL-E data.

All information about research infrastructures found in the ENVRI knowledge base at present is provisional, and should not be considered to be an absolutely accurate representation of the infrastructures in question; rather, the knowledge base is an instrument to demonstrate the benefits of collecting such information for comparison and analysis.

The ENVRI Knowledge Base in its first iteration as a product of the ENVRIplus project has three basic purposes:

1. It provides an example of OIL-E in use, providing examples of RI-oriented data structured in accordance with the OIL-E ontologies.
2. It provides a repository for RI architectural information and 'design wisdom' encoded using ENVRI RM that can be programmatically queried and analysed.
3. It serves as a database of information about technologies and standards used by RIs.

The current knowledge base is hosted via a standalone instance of Apache Jena Fuseki[8], which provides a triple store for aggregated RDF data along with a service API and internal reasoning capabilities based on the OWL standard. The knowledge base contains the complete set of OIL-E ontologies along with a representative sample of RI-specific data for the purposes of demonstration and experimentation. Access to the knowledge base is achieved via a SPARQL endpoint, the address of which serialised SPARQL queries can be appended as HTTP requests:

`http://oil-e.vlan400.uvalight.net/rm/sparql?`

There is also a public landing page provided to allow for the testing and modification of sample queries by interested parties:

`http://oil-e.vlan400.uvalight.net`

For resolving queries, the knowledge base is able to apply the relations and classifications defined by OIL-E in order to infer results beyond those explicitly asserted in the internal triple store, allowing agents to use the full set of ENVRI RM archetypes to guide discovery and search over the RI data provided.

For the ENVRI Knowledge Base, we identified four key knowledge capabilities that application of the semantic linking framework can facilitate:

1. **A survey of the technical landscape.** The web of knowledge created by semantic linking should help us understand what technologies (including software, standards and vocabularies) are being used by environmental science RIs.
2. **Comparative solution analysis.** It should be possible to compare solutions developed by environmental science RIs—specifically, given the knowledge of how technologies are used in their proper context, we should be able to compare developments in equivalent contexts.
3. **Gap analysis and component recommendation.** Given a reference model for environmental science RIs, it should be possible to identify what is missing in the current development state of a given RI, and based on both that model and the solutions developed by other RIs, it should be possible to then make certain recommendations.
4. **Linked open research infrastructure.** The web of knowledge created by semantic linking should itself be publicly accessible, machine-navigable, and provide a gateway to the services and data held by the RIs, including where available data provenance and resource catalogues, and making use where appropriate the use of other ENVRI services such as the catalogue service for cross-RI search.

Queries demonstrating how the knowledge base can address each of these capabilities in turn are defined on the public landing page at `http://oil-e.vlan400.uvalight.net`.

[8] https://jena.apache.org/documentation/fuseki2/

4 Further development

The knowledge base and OIL-E are both the basis for more tools with which to support several useful functions. We can envisage a number of avenues of further development (or in most cases, alignment with existing developments for mutual benefit)—these include:

Cross-RI search and discovery. OIL-E provides a standard taxonomy for various entities and activities related to RIs, which can be used to classify different kinds of resource as part of a faceted search pipeline. An OIL-E knowledge base can hypothetically act directly as a catalogue service for multiple RIs, but this is not necessarily the best possible approach, as OIL-E is optimised for describing RI design and contextualising RIs' component parts, rather than providing a more traditional metadata scheme for (data) resources. In ENVRIplus, it is the flagship catalogue service[9] that provides this kind of joint catalogue functionality, with the ENVRI Knowledge Base providing ancillary knowledge facilities. For example, the knowledge base can be the basis for a discovery service for heterogeneous research assets (including other catalogues) based on its internal network of relationships based on ENVRI RM.

Faster RI specification using ENVRI RM. Detailed descriptions of RIs in terms of their architecture, core data products and processes allows for more in-depth investigations and comparisons of RI solutions to various technical problems. ENVRI RM provides the basis for such descriptions, but requires specialist expertise to use effectively, and has previously been used manually, resulting in the creation of a body of documentation for each RI modelled. OIL-E captures all the key concepts in ENVRI RM, and thus a tool based on OIL-E (and backed by a knowledge base) that would allow RI architects to more easily specify their RIs using ENVRI RM templates backed by OIL-E validation would accelerate the creation of RI data that can be used in comparative analyses.

Requirements recommendation. Using tools such as OIL-E and the ENVRI Knowledge Base, it is possible to do comparative analysis of the solutions provided by RIs in terms of technology and processes to address various common problems regarding the handling of research data (and other things). This requires a certain degree of constructive analysis of a number of queries. Tools which can interact with the knowledge base on behalf of a user, constructing and interpreting queries behind a more friendly interface, could be very useful for taking full advantage of the corpus of knowledge built up from RI modelling.

Provenance exploration. There are two notable ways in which OIL-E data can interact with provenance data, especially data encoded to the W3C PROV standard [3]. The first is as linking data to various provenance repositories, contextualising the role of the repositories and providing a reference to where the provenance is and how it can be extracted. The second is as a validation framework; given descriptions of RI processes encoded in OIL-E, provenance traces can be checked against those descriptions by mapping agents, entities and activities to the correct OIL-E concepts and then checking whether the relationships described in the provenance trace match those of prescribed by the process model.

Natural language based document analysis and annotation. A significant corpus of existing information about RIs exists in the form of written documentation produced by RI architects and developers. The ability to apply a framework such as OIL-E to annotate uploaded documents, identifying possible references to concepts defined in ENVRI RM in the text for example, would be useful both to contextualise documents automatically and provide initial descriptions for the RIs and RI components described by the documents. Such descriptions can be verified and extended by human experts, and also used as training data for producing better annotations in future, or perhaps even to identify possible extensions (e.g. new concepts or alternative synonyms for existing concepts) to ENVRI RM. Machine learning tools would thus provide a valuable additional source of data for the knowledge base, or to validate existing models of RIs.

The use of OIL-E as both a metadata scheme for architecturally-oriented RI data and as a classification scheme for research assets has been explored in the context of the VRE4EIC project[10]. VRE4EIC is

[9] See Deliverable 8.3 *"Interoperable cataloging and harmonization for environmental RI projects: system design"*.

[10] https://www.vre4eic.eu/

concerned with defining a generic architecture for virtual research environments (in a similar spirit to ENVRI RM), and with the implementation of some key building blocks of such environments, including a metadata service built around a cross-RI resource catalogue. To build this metadata service, metadata harvested from external sources is converted to CERIF RDF[11] using the X3ML mapping framework [11]. In addition to there being a mapping from OIL-E to CERIF (allowing OIL-E data from *e.g.* the knowledge base to be used to enrich a CERIF-based catalogue), investigations are also underway into using OIL-E as a classification scheme for CERIF objects created when mapping from other standards such as ISO 19139, where the additional contextual information provided by OIL-E can be used to enrich the base CERIF entities and relations.

One possible extension to OIL-E now being investigated is the integration of SHACL functions into OIL-E. The Shapes Constraint Language (SHACL) is a constraint language used to validate RDF graphs, and is a refinement of prior *de facto* standards such as SPIN and SWRL. Unlike OWL it performs closed world validation rather than open world classification, and also makes the unique name assumption that OWL explicitly does not. SHACL can be used to embed SPARQL queries into RDF graphs as part of rules or functions that can be applied on the content of the graph, providing a means for RI service developers to publish instructions for building (for example) parameterised HTTP requests to their services that other actors can retrieve from the knowledge fabric—this ties directly in the linked open research infrastructure capability described for the ENVRI Knowledge Base in the previous section. Such an approach allows interaction logic to be defined (and updated) in one place (*e.g.* the knowledge base or a successor system that may be distributed over several nodes perhaps directly curated by RIs). It also admits the possibility that other information in the linked knowledge graph can be used in a dynamic fashion to introduce some additional interstitial intelligence into the logic.

The activities described in this report are on-going, and extend beyond the bounds of Task 5.3 and the ENVRIplus project as a whole. The fragmented yet still very active semantic landscape of environmental science research infrastructure cannot be expected to cease evolving any time soon, and efforts must be made to evolve with that landscape. Only by continued interaction and monitoring will the ‘semantic linking’ of interdisciplinary environmental research be expected to achieve its full potential.

[11] https://www.eurocris.org/cerif/main-features-cerif

References

[1] Robert Arp, Barry Smith, and Andrew D Spear. *Building ontologies with Basic Formal Ontology*. The MIT Press, 2015.

[2] Tim Berners-Lee. Linked data. *W3C Design Issues*, 2006. Accessed 26th February 2018.

[3] Paul Groth and Luc Moreau. PROV-overview. W3C note, W3C, 2013. http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

[4] ISO 10746-1. Information technology—Open Distributed Processing—Reference model: Overview. ISO/IEC standard, International Organization for Standardization, 1998.

[5] ISO 19115-1:2014. Geographic information—Metadata—Part 1: Fundamentals. ISO standard, International Organization for Standardization, 2014.

[6] ISO 19139:2007. Geographic information—Metadata—XML schema implementation. ISO/TS standard, International Organization for Standardization, 2007.

[7] Mostafa Jebbar, Abderrahim Sekkaki, and Othmane Benamar. Integration of SOA and cloud computing in RM-ODP. In *Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on*, pages 97–105. IEEE, 2012.

[8] Brigitte Jörg. CERIF: The common european research information format model. *Data Science Journal*, 9:24–31, 2010.

[9] Haim Kilov, Peter F. Linington, José Raúl Romero, Akira Tanaka, and Antonio Vallecillo. The reference model of open distributed processing: Foundations, experience and applications. *Computer Standards & Interfaces*, 35(3):247 – 256, 2013.

[10] Dimitris Kontokostas and Holger Knublauch. Shapes constraint language (SHACL). W3C recommendation, W3C, July 2017. https://www.w3.org/TR/2017/REC-shacl-20170720/.

[11] Yannis Marketakis, Nikos Minadakis, Haridimos Kondylakis, Konstantina Konsolaki, Georgios Samaritakis, Maria Theodoridou, Giorgos Flouris, and Martin Doerr. X3ML mapping framework for information integration in cultural heritage and beyond. *International Journal on Digital Libraries*, pages 1–19, 2016.

[12] Paul Martin, Yin Chen, Alex Hardisty, Keith Jeffery, and Zhiming Zhao. Computational challenges in global environmental research infrastructures. In Abad Chabbi and Henry W Loescher, editors, *Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities*, chapter 12, pages 305–340. CRC Press, 2017.

[13] Paul Martin, Paola Grosso, Barbara Magagna, Herbert Schentz, Yin Chen, Alex Hardisty, Wouter Los, Keith Jeffery, Cees de Laat, and Zhiming Zhao. Open information linking for environmental research infrastructures. In *2015 IEEE 11th International Conference on e-Science (e-Science)*, pages 513–520. IEEE, 2015.

[14] Deborah McGuinness, Satya Sahoo, and Timothy Lebo. PROV-O: The PROV ontology. W3C recommendation, W3C, 2013. http://www.w3.org/TR/2013/REC-prov-o-20130430/.

[15] Abraham Nieva de la Hidalga, Barbara Magagna, Markus Stocker, Alex Hardisty, Paul Martin, Zhiming Zhao, Malcolm Atkinson, and Keith Jeffery. The ENVRI Reference Model (ENVRI RM) version 2.2, 30th October 2017, November 2017.

[16] I. Román, G. Madinabeitia, L. Jimenez, G.A. Molina, and J.A. Ternero. Experiences applying RM-ODP principles and techniques to intelligent transportation system architectures. *Computer Standards & Interfaces*, 35(3):338–347, 2013.

[17] W3C OWL Working Group. OWL 2 web ontology language. W3C recommendation, W3C, 2012. https://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

[18] Alain Wegmann and Andrey Naumenko. Conceptual modeling of complex systems using an rm-odp based ontology. In *Enterprise Distributed Object Computing Conference, 2001. EDOC'01. Proceedings. Fifth IEEE International*, pages 200–211. IEEE, 2001.