Semantic Linking of Research Infrastructure Metadata

Paul Martin[1], Barbara Magagna[2], Xiaofeng Liao[1], and Zhiming Zhao[1](✉)

[1] Multiscale Networked Systems, University of Amsterdam, 1098XH Amsterdam, The Netherlands
paulmartin.research@gmail.com, {x.liao,z.zhao}@uva.nl
[2] Environment Agency Austria, Vienna, Austria
barbara.magagna@umweltbundesamt.at

Abstract. The use of metadata to characterise scientific datasets, making data easier to discover and use directly by researchers and via various online data services, is one of the primary concerns of research infrastructures (RIs); also, of concern is the use of metadata to describe equipment, facilities, services and other research assets. Metadata models and terminology differ greatly between different communities and infrastructures however, and so make synthesising complex interdisciplinary scientific workflows involving assets from multiple RIs very challenging.

'Semantic linking' addresses the need to enhance the interoperability of RI services and data by bridging metadata schemes, ontologies and vocabularies used by different research communities, whether by standardising the terminologies and schemes used by those communities, or by dynamically transforming metadata from one standard to another when retrieved by services on behalf of researchers executing their scientific workflows.

Multiple techniques for and modes of semantic linking have been investigated in the context of the ENVRI community cluster of environmental and Earth science RIs, including top-down modelling of entities and activities within a standard reference model, enrichment of existing metadata records with shared terminology, full transformation of metadata records from one standard to another, and the generation of additional links to existing online data. We review some of these activities and their application to the promotion of semantic interoperability between RIs, and discuss other possibilities and recent developments that may also be useful for enhancing interdisciplinary data science.

Keywords: Metadata · Semantics · Linking

1 Introduction

The adoption and use of metadata for characterising and cataloguing scientific data and other research assets is one of the primary concerns of modern scientific research infrastructures (RIs). The production and maintenance of good metadata has bearing on the

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 226–246, 2020.
https://doi.org/10.1007/978-3-030-52829-4_13

entire research lifecycle, from acquisition and curation through to publishing, processing and use. The adoption of standard protocols, metadata schemes and controlled vocabularies for use in scientific data and their associated metadata by a given research community is supposed to expedite data sharing and the development of interoperable data services within a scientific discipline. The increasing need for interdisciplinary research makes such standardisation more challenging however, as the range and diversity of scientific products that should be normalised grows ever greater. Even mature standards do not always meet all community requirements, or else have ambiguous semantics that lead to variation in how they are applied. In addition, many communities have already adopted and adapted to their own preferred standards independently, and have their own established best practices and legacy systems. It therefore seems unavoidable that there will always be variation in metadata schemes, vocabularies and protocols, and thus a need to be able to translate information between different semantic contexts, as represented by specific data models and terminology, whether on request or performed dynamically out of sight of researchers. Regardless of how it is carried out, we refer to this kind of translation as *semantic linking*; techniques for bridging the gap between two or more semantic domains to permit cross-domain data science.

Semantic linking is of great importance in the development of an interdisciplinary 'data science commons' for researchers—a common environment for getting access to and contributing scientific data. The ideal scenario is that researchers can retrieve data, tools, models and other services from different RIs based on scientific requirements without having to know which specific infrastructure serves which specific data, and can use them in complex workflows without having to manually rework data inputs at each step [41]. Specifically, the use of semantic linking is necessary in the development of joint catalogues or indexes of research assets (needed for cross-RI search and discovery), to export data and metadata into different operational contexts, and to glue together services with different input and output formats.

Semantic linking was thus identified as one of the three main cross-cutting activities of the 'Data for Science' theme of the ENVRIplus project[1], alongside the development and exploitation of the ENVRI Reference Model (ENVRI RM) [1, 2] and the specification of common abstract architecture for the construction of interoperable services. One of the results of this activity was the development of Open Information Linking for Environmental Research Infrastructures (OIL-E) [3] as a kind of architectural hub ontology for RI descriptions. Using OIL-E as our baseline semantic model, we surveyed four different kinds of semantic linking during the project; in this chapter we review these four kinds in turn and consider how they reflect on the challenge of achieving semantic interoperability in data science research in general and within the environmental and earth sciences in particular.

In the next section (Sect. 2), we examine more closely the background and motivation for the investigation of semantics in environmental and Earth science RIs. We describe the methodology applied in ENVRIplus for surveying and rationalising the semantic landscape of RIs involved in the project (Sect. 3), before then moving on to discussing the four semantic linking scenarios we proceeded to investigate (Sect. 4). We discuss some of the technological developments that might have bearing on RI semantics and

[1] https://www.envriplus.eu/.

metadata and on semantic linking activities in general (Sect. 4) before finally drawing our conclusions (Sect. 6).

2 Background

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. In this context, the role of research infrastructures (RIs) 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 [4]. Consequently, today there is a host of different research infrastructures, each with their own intersecting speciality areas, but more broadly sharing many common scientific, technical, political and governance-oriented 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 different research infrastructures, with the intention that they are allowing them to collaborate and conduct their research more effectively than ever was possible before. This is the challenge that initiatives such as the Research Data Alliance[2] and proposals for FAIR (Findable, Accessibility, Interoperable and Reusable) data [5] seek to address, and it is one that fundamentally relies on the proper elicitation and application of semantics in research data in general.

Data semantics are provided by the various schemas produced for datasets and metadata and are embedded in the choice of vocabulary used to describe different data elements. For metadata in particular, having well-defined and rigorous descriptions in a machine-actionable format confers a number of advantages to both the provider and user of the data or other resources being described. Publishing metadata about the resources (not only data, but also services, tools and facilities) that RIs offer online (indicating such information as the type of resource and their provenance) allows them to advertise their offerings and allows researchers to browse and discover resources (including data, models, tools, services and other kinds of resources both digital and physical) that could be useful to their research. It also permits comparison and the integration of resources into larger workflows or toolchains. More fundamentally however, it also ensures that the resource (and this is especially vital for scientific datasets) is and continues to be correctly understood, and not subject to confusion regarding the exact thing being measured or observed, the units used, or the time and location when/where a measurement or observation was made. Semantic rigour is thus vital for well-grounded, reproducible and accountable research.

In this space there are many metadata standards, old and new; some of which are *de facto* standards long adopted by particular communities, while others have achieved *de jure* status as recommendations by certain community institutions such as the International Organization for Standardization (ISO) and the Open Geospatial Consortium

[2] https://www.rd-alliance.org/.

(OGC). For example, in the geospatial area, which concerns many environmental and Earth science RIs, there exist established standards such as ISOs 19115 [6] (for geospatial data) and 19139 [7] (the accompanying XML profile), 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 idiosyncratic across communities, with resulting variations in how metadata elements are realised or terms applied. There are also standard protocols for accessing catalogues of metadata records used to describe data collections via the Web; standards such as DCAT [8] describe how data catalogues should be structured, and protocols such as CSW [9] and OAI-PMH [10] describe how they ought to be accessed. Many RIs use these established protocols, but some RIs also use Semantic Web [11] technologies such as OWL [12] and SKOS [13] to describe their resources and use SPARQL [14] to access them. These RIs adapt ontologies such as OBOE [15] (for observations) and vocabularies such as EnvThes [16] (for ecology) to meet their own community's needs while building upon the semantic harmonisation work of other neighbouring communities. Continuing harmonisation of vocabulary and metadata between research infrastructures thus remains an on-going concern; for example, the European Open Science Cloud initiative (EOSC) [17] considers it a major priority to integrate existing terminological resources with the services provided by European RIs to realise its goals for better cross-disciplinary open science, and a similar urgency can be seen in other open science initiatives around the world.

The integration of resources requires alignment of data formats and content. One of the roles of an RI within the context of its target community is to facilitate standardisation, and as such RIs are very useful vehicles for aligning the use of semantics within a community. Nevertheless, such standardisation activity becomes very difficult once boundaries between communities (even within the same scientific discipline) are crossed. This is because intrinsically, the requirements and usage of data products can be very different between communities. This means that the metadata models used, and indeed how the very datasets being described are even structured for use by researchers, likewise differ considerably between communities. A simple example would be how some communities gather all data related to a given location into a single dataset that might then partitioned by time period, while other communities may gather all of a single kind of observation into one dataset with the locality of each observation reduced to a single field within each row of data. Thus, it remains necessary, even in the presence of initiatives such as RDA (which provides a forum for discussion of best practices for addressing various data science challenges) and initiatives such as Copernicus[4] and GEOSS[5] (which act as aggregators for specific classes of data and thus promote certain standards for such data), to consider how to transform metadata between models in order to allow different data services and tools to work together as part of a cohesive operational workflow.

The semantic linking work of ENVRIplus was intended to guide the harmonisation of semantics across environmental science research infrastructures by providing both

[3] https://inspire.ec.europa.eu/.
[4] https://www.copernicus.eu/.
[5] https://www.earthobservations.org/geoss.php.

contextualisation and a standard 'connective' upper ontology for the different kinds of entities and activities commonly found in those infrastructures. Notably, there is no 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 be expended in future. Instead, there exist many tools and frameworks for handling such mappings and a great body of research. Our concern then is rather with providing some baseline support for analysing the diversity of such schemes and mappings where they exist, and so help research infrastructure developers to focus their efforts on specific problem areas.

3 Semantic Linking in ENVRIplus

To even approach the topic of semantic linking, there is a need to understand the *semantic landscape* of research infrastructure at large. By 'landscape', we essentially mean information about not just which metadata schemes, ontologies and vocabularies are in use by different RIs, but also how they are used and for what purpose. Without an understanding of the landscape of the use of semantic instruments and standards, it is impossible to identify where to target semantic linking activities—to determine where it is needed, and which models/terminologies need alignment in order to facilitate some otherwise hypothetical workflow. The semantic linking activity in the ENVRI environmental and Earth science RI cluster[6] was carried out in several stages:

1. We collected information from environmental and Earth science RIs and communities, regarding their requirements, adopted technologies and the current state of the art; much of the results of this process appear in Chapter 3 of this book.
2. We used the requirements gathered in the previous step to refine the ENVRI RM (described in detail in Chapter 4), which importantly (for the purpose of semantics and shared terminology) provided a common vocabulary for describing various kinds of component and activity deployed in RIs, and helped us to identify the most important interactions typically facilitated (or needed) by environmental and Earth science RIs.
3. Concurrently, we also began gathering information about the community standards, protocols, and semantic/terminological resources used by RIs and in various aspects of environmental research, data and process specification. This was performed mainly via direct interactions with technical experts involved in RI development.
4. We developed Open Information Linking for Environmental Research Infrastructures (OIL-E, described more completely in Chapter 6) to capture the stereotypical elements of environmental and Earth science RIs as identified by ENVRI RM, and define the necessary relationships between those stereotypes across different views of science, information, computation, engineering and technology. One of the roles of OIL-E, aside from allowing for various RI descriptions based on ENVRI RM to be transformed into a format that can be uploaded into an ENVRI Knowledge Base [39] and programmatically queried, was to act as a connective 'hub' ontology

[6] https://www.envri.eu/.

for RI architecture. This ontology allows specifications of specific concepts to be extended with other, more specific ontologies and taxonomies used by the scientific community.
5. Using the OIL-E ontology to structure the data, we began mapping the semantic landscape of environment science by 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 (also described in Chapter 6) to contain all the formally-encoded data, and to provide a service with which architects and developers can investigate and contribute descriptions of RIs.
7. We further investigated specific approaches for linking data encoded using OIL-E with other (meta)data sources of interest to researchers or to the RIs that support their activities. The next part of this chapter goes into these investigations in further detail.
8. Within the framework of successor projects such as ENVRI-FAIR[7], we can now focus 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 and the ENVRI Knowledge Base.

Figure 1 provides a pictorial overview of the relationship between the various parts of the semantic landscape mapping in ENVRI, which was also used in various dissemination materials produced by the project.

4 Semantic Linking Scenarios

'Semantic linking' in the context of the cluster of environmental science research infrastructures is fundamentally concerned with how to contextualise (meta)data regarding research datasets, tools, methods and infrastructure such that they can be interpreted in accordance with a particular model of reality, are meaningfully comparable with similar metadata, and can be understood as part of a wider semantic landscape. This is so that (for example) we can determine the role of certain data in specific processes within a particular infrastructure. We therefore need to consider how to 'elevate' existing data semantically (by providing additional context needed to do more with the data), and how to transform those data where necessary (so that we can use them elsewhere). We need to consider what new data must be created to provide additional context to the entities we wish to model, as well as to describe the relationships between entities.

There are four semantic linking scenarios that need to be considered in the context of environmental science and environmental science research infrastructure, that we chose in the context of ENVRIplus to explore in more depth:

1. The creation of a new model for an existing artefact or process based on a formal ontology. This could be in addition to existing semantic metadata for that artefact or process, providing additional contextual information that could allow for multiple means of interaction with a given research asset, for example by creating multiple

[7] https://envri.eu/envri-fair/.

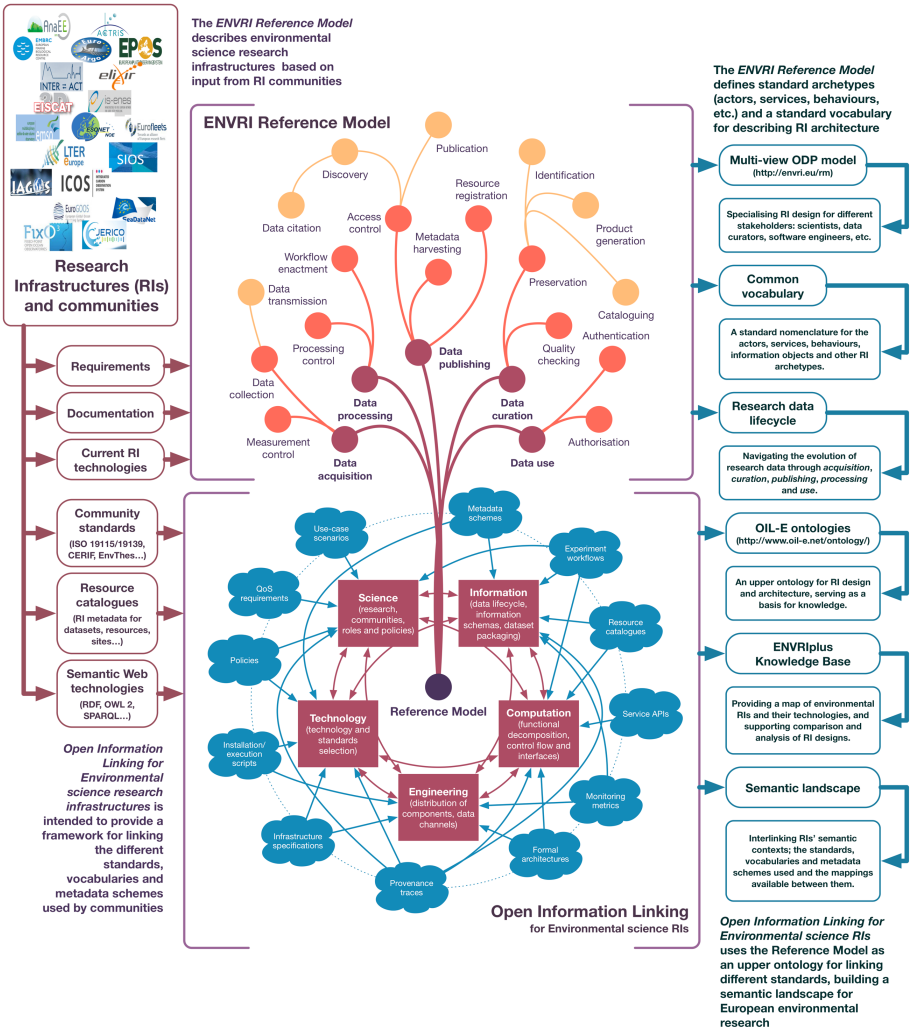

Fig. 1. The vision of semantic survey and linking over the course of the ENVRI projects.

metadata records in different schemes for the same data product for retrieval and use by different services with different protocols.

2. The enrichment of an existing model using controlled vocabulary extracted from an ontology or other formal terminological resource. In this case, the additional vocabulary provides additional metadata by which services (e.g. for search and discovery) can differentiate and classify research assets already described using a set metadata scheme and protocol.
3. The translation of an existing model from one semantic context to another. Rather than augmenting or linking to existing semantic metadata, this is the scenario where entirely new metadata is generated from existing metadata, generally for inclusion

in another metadata catalogue or repository which requires a different scheme for describing research assets.

4. The linking of two models for the same entity (or conceptually overlapping entities) by generating additional 'bridging' metadata between existing metadata records. This is the linked open data approach, whereby information existing independently in multiple contexts about the same or similar entities is somehow made connected such that an external query service can navigate between contexts and aggregate the results from each.

All of these four scenarios have overlaps in their objective and concerns to the extent that it is not always clear to which scenario a given semantic linking operation belongs (and in many cases an operation could justifiably belong to more than one), but nonetheless it is useful to consider how semantic technologies might be used to address each case in turn.

4.1 Semantic Contextualization

The most basic form of 'semantic linking' is the (re-)contextualisation of data already somehow modelled using some ontology or metadata scheme. Typically, this involves describing and classifying entities using a new ontology or other metadata scheme, which provides new metadata that can be used to discover and retrieve information about those entities (or the entities themselves if they exist as data). Doing this for multiple ontologies/schemes binds the data in question to two or more different semantic domains, and so allows the data to be examined in either context; this is most appropriate in the case of multiple systems that might want to query the data, but where each system supports a different schema. A benefit of this kind of 'multiple classification' is that it creates sample data for constructing more formal semantic mappings between two different semantic models should it later be determined that all data in one model needs to be transformed into the other. The main benefit, however, and the distinguishing factor from the scenario where the second model is simply generated from the first model automatically, is that the second model may capture information not representable by the first model, thus increasing the amount of information about a data entity available. For example, one model might not capture procedural aspects of how a dataset is created, while another does; thus, it is not possible to simply generate the metadata required by the latter model from the former model. It may be useful however to simultaneously describe the dataset using *both* models for the additional flexibility such multi-modelling grants, such as support for two different querying systems that each expect a specific model to be used.

In the context of the ENVRIplus project, different kinds of entities with semantic connotations (datasets, metadata schemes, vocabularies, etc.) were described using the OIL-E ontology and so 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 data acquisition, specifically the collection of data regarding phytoplankton.

In Fig. 2, concepts from four of the five viewpoints defined in OIL-E are used (though actually only one concept is used from the technology view). In addition, a number of

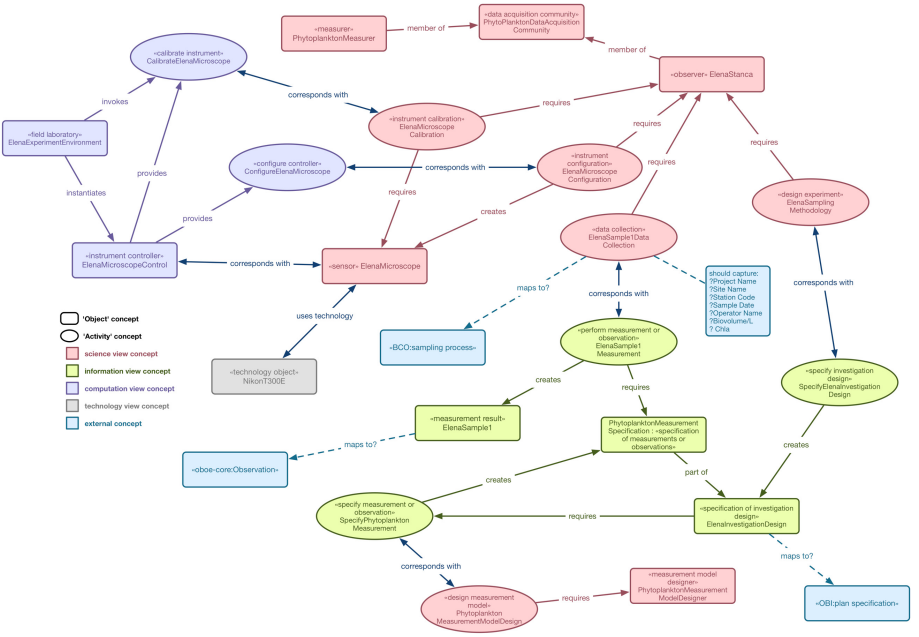

Fig. 2. Modelling the acquisition of data regarding phytoplankton across multiple views in OIL-E.

points at which entities might be further explicated using other ontologies are highlighted (using the OBOE, OBI or BCO ontologies). This allows us to describe the activity of data collection (answering who, what and how), the data being created, and the processes involved. Each of these views could be elaborated upon or linked to a larger dataset in OIL-E or, using one of the linking methods described in the following sections, translated into another ontological model.

The ENVRI Knowledge Base was the primary vehicle for exploring this kind of semantic linking within the ENVRIplus project: by collecting information about different RIs using the terminology of ENVRI RM and the framework of OIL-E, we were able to explore and visualise the resulting knowledge network and perform some fundamental comparative analyses.

4.2 Semantic Enrichment

Often, it is not necessary to create new descriptions of entity data from scratch. While some aspects of research infrastructure (particularly processes) are rarely formally described in any machine-actionable representation, other things (particularly datasets or services) already have descriptive metadata based on some formal model. The issue then becomes not that of how to (re-)model the entity in question, but how to 'plug in' the existing model into a wider semantic context such that the information within the model can be made better use of by a greater variety of knowledge-driven services. One approach is to transform the existing model into a new model that is somehow more 'semantically interoperable'; we address this in the next section. Another approach is to

enrich the existing model 'externally', by creating linking data hosted outwith the model that allows an external actor to find and retrieve the information in the current model; we address this in the section on semantic bridging. A third approach is to enrich the existing model *internally*, by taking controlled vocabulary from an external ontology or thesaurus and annotating the model where it permits the insertion of such vocabulary, allowing for external services to harvest that information and thus 'comprehend' the context in which the model is applied. For example, we can take observation data from an RI such as the Integrated Carbon Observation System (ICOS)[8] and annotate the datasets with terms from the EnvThes thesaurus for ecosystem observations in order to better identify the scientific context of each observation set—e.g. that it pertains to the North Atlantic Oscillation[9], or to snow accumulation[10].

We consider here the example of CERIF (Common European Research Information Format) [18]. CERIF is a recommendation for the contextualisation of research activity, relating people to organisations, to projects, to equipment, to datasets and other research products. Investigated as a possible base scheme for cross-RI joint research asset catalogues, CERIF is notable for 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. Almost every entity and relation can be assigned a classification however that indicates a particular semantic interpretation (e.g. that the relationship between a *Person* and a *Product* is that of a creator and their creation), allowing a CERIF database to be enriched with concepts from an external semantic model (or several linked models). In this respect, the vocabulary provided by OIL-E was investigated as a means to further classify objects in CERIF in terms of their role in a research infrastructure, e.g. classifying individuals and facilities by the roles they play in research activities, datasets in terms of the research data lifecycle, or computational services by the functions they enable. This can provide additional operational context for faceted search—for example to identify which processes generated a data product, or to search for quality-assured datasets only.

Some examples of classifications based on ENVRI RM stereotypes defined in OIL-E are given in Table 1. Classifying CERIF entity classes such as *Person*, *Facility*, *Result Entity* or *Service* using OIL-E concepts such as *environmental scientist*, *data provider*, *persistent dataset* and *virtual laboratory* is simple enough, but OIL-E can also be used to classify various classes of RI activity involving interactions between instances of CERIF entity in a way that is particularly suitable for describing time-bounded events involving those entities. For example, given a CERIF relation between a Person and the Result Entity that the person in question annotated, that relation can be classified using the 'annotate data' information action concept in OIL-E, with CERIF also capturing the time of annotation.

Semantic enrichment of this kind need not be limited to one particular semantic context. Providing additional information about the *scientific* context for datasets (e.g. categorising the experimental method applied to generate the data or the branch of science to which the data belong) is also important, and there exist many vocabularies to

[8] https://www.icos-ri.eu/.

[9] http://vocabs.lter-europe.net/EnvThes/20403.

[10] http://vocabs.lter-europe.net/EnvThes/20949.

Table 1. Example classifications of CERIF entities based on ENVRI RM stereotypes.

CERIF entity	OIL-E concept	Example classifications
'Event'	'behaviour'	'data collection [behaviour]', 'data replication [behaviour]'
'Equipment'	'resource'	'sensor network', 'storage system'
'Facility'	'resource'	'data repository', 'research infrastructure'
'Organisation Unit'	'actor'	'data publisher', 'semantic mediator'
'Person'	'actor'	'environmental scientist', 'engineer'
'Result Entity'	'persistent data'	'QA-assessed data', 'annotated data'
'Service'	'computational object'	'catalogue service', 'data broker'

do this (and indeed many are already in use for just this purpose). Aside from the prescribed code-lists of ISO 19115, environmental science research infrastructures such as AnaEE[11] and LTER-Europe[12] are actively developing better vocabularies for describing ecosystem and biodiversity research data, building upon existing SKOS vocabularies (such as EnvThes, referenced above).

There is no need to restrict annotation of metadata to one specific controlled vocabulary, especially if links between terms in different vocabularies can be established [40]. The identification of synonymous, subsuming and intersecting terms (and the publication of such links in a machine-accessible way such as on the Semantic Web) can provide the basis for better semantic search whereby a greater range of data products with similar characteristics can be retrieved on query without necessarily sharing precisely the same controlled vocabulary for their metadata. Making use of such linked vocabulary would simplify the task of integrating resource metadata from multiple catalogues as it would reduce the need to map all metadata values into a single master vocabulary (with the likely resulting loss of nuance), while still retaining the benefits of cross-RI search and discovery. A number of environmental and Earth science RIs such as AnaEE, LTER-Europe, LifeWatch and ICOS are now investigating such linking of vocabularies as part of an effort to make their respective resource catalogues more interoperable.

4.3 Semantic Mapping

Semantic mapping concerns the full mapping of data from one semantic context to another, with all the necessary structural transformation that entails. Such mapping might be applied on a targeted basis to specific metadata records, or to the results of specific queries retrieved from metadata servers, or there may be a mass translation of an entire catalogue. In general however, full semantic mapping is performed when integrating data from multiple sources into a single corpus with a single ontology and vocabulary. In the context of environmental science research infrastructure, this most typically arises

[11] https://www.anaee.com/.
[12] http://www.lter-europe.net/lter-europe.

in the construction of joint catalogues combining metadata records from heterogeneous sources. In order to facilitate search and discovery over the entire joint catalogue, the metadata gathered must of course be aligned to the greatest extent possible, which means one standard scheme, though cross-RI search and discovery can be further enhanced by identifying links between synonymous or related vocabulary terms, which can be seen as a kind of semantic bridging between controlled vocabularies (see next section).

A mapping agent will access the source of the data, apply the mapping, and record the mapped data in some target resource. The mapped data are 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 is therefore needed to trigger such updates or to regularly poll the source for changes. Various tools exist for defining mappings between different ontologies or metadata schemes. An example of such a tool is the 3M Mapping Memory Manager[13], which implements the X3ML framework [19] for specifying translations from XML-based metadata schemes to RDF.

In addition to the enrichment activity described above, we have also explored mapping from various different common metadata schemes into CERIF RDF, which applies the CERIF 1.6 standard to RDF [20]. Figure 3 shows a snapshot of semantic mapping, in this case defining mapping rules from the metadata scheme used by the CKAN-based EUDAT B2FIND service[14] to CERIF RDF using 3M.

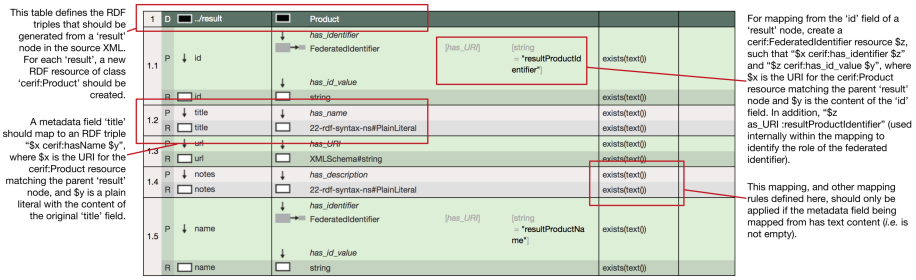

Fig. 3. Example of mapping rules generated in 3M: XML harvested from CKAN to CERIF RDF.

Mappings in 3M are described by X3ML mapping rules relating elements found in the source (XML-based) scheme to RDF triples in the target scheme, subject to various syntactic conditions (e.g. element type, parent hierarchy and internal content). Thus, 3M interprets the information fields in the source scheme based on the RDF *subject-predicate-object* model; each parent node in the source document is associated with a given RDF resource (the subject), and the content of each mapped field is used to generate triples linking that resource to other RDF resources (objects) via predicates derived from the field in question. For example in Fig. 3 above, each result node is mapped to a 'Product' resource, and the contents of each result node's 'id' field is associated with the product via a triple linking the product to a 'FederatedIdentifier' object via the predicate 'has_identifier'. 3M supports the specification of generators to

[13] https://github.com/isl/Mapping-Memory-Manager.
[14] https://eudat.eu/services/b2find.

produce unique identifiers for new RDF resources constructed during mapping of terms, and also provides various test and analytics facilities by which to evaluate (for example) the completeness of a given mapping. Examples of mappings into CERIF RDF, including mapping from OIL-E to CERIF have been published online [21] as part of the technical output of the VRE4EIC project[15], which the ENVRI community participated in.

Regarding the schema mapping between XML and RDF triples, we developed another work which provides insights in two folds [43]. Firstly, testify the validity of single matcher in a column based manner for the semantic data types. Secondly, testify the validity of a highly configurable framework that utilises hierarchical classification in order to construct a composable pipeline. Based on this vision, a Reconfigurable pipeline for Semi-Automatic Schema Matching (REPSASM)[16], was implemented to solve the customizability of the matching problem by providing an environment in which a user can create, configure and experiment with their own schema-matching procedure.

Other tools exist for transformation of data records, particularly between formats and models. For example, the derivation of RDF from relational database tables can be done quite naively by treating each table as the subject of an RDF triple, each column as the predicate, and each cell as the object, but this rarely creates a good representation of the source data. Instead, the use of tools such as Ontop[17], which applies the R2RML standard [22] for mapping relational database schemes to RDF, can allow relational databases to be queried as if they were RDF (see Fig. 4).

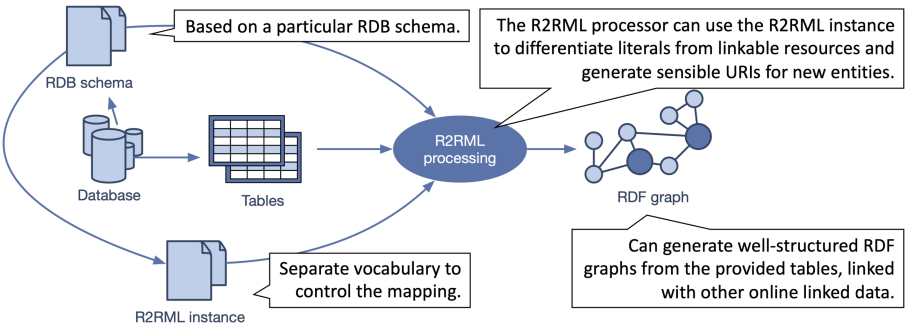

Fig. 4. Using R2RML to generate RDF from relational database tables.

This essentially performs the desired mapping on the fly, allowing for the benefits of mature relational database management systems to be retained; such an approach is being applied by RIs such as AnaEE[18] to extract metadata for semantic annotation using a standard ecosystem ontology built on the OBOE ontology. One thing to consider is that because this type of mapping can be performed at query time, it is not necessary to actually fully transform all the content of a relational database into RDF in advance, or

15 https://www.vre4eic.eu/.
16 https://github.com/JordyBottelier/arpsas.
17 https://ontop.inf.unibz.it/.
18 https://www.anaee.com/.

indeed at all. Instead, transformation can be performed on the results of queries on the database, presenting those results as native RDF without any indication that the source data exist in a different format or schema. Based on the types of query and retrieval operations performed on a data corpus, this kind of on-the-fly mapping might be more performant than transforming the entire data corpus in advance, especially if changes to the corpus might later need to be propagated to the mapped data. The decision on whether to map everything in advance to create a unified data source, or to map on demand just the information extracted from queries, is an important one to make when carrying out semantic mapping; balancing stability, performance, liveness and other concerns against one another. It is not a binary choice however. Certain key metadata (used for locating data for example) could be mapped in advance to create an 'upper' database via which to query individual data sources, or the results of recent or recurring queries for which mapping has already been performed could be cached in a central locale, reducing data retrieval time. All these approaches to mapping can be automated, and so be used 'under the surface' to improve the interoperability of RI systems and create the appearance of standardisation from the researcher perspective.

4.4 Semantic Bridging

Sometimes, the main barrier to interoperability is not the format of the metadata describing the data or service of interest, nor is it the vocabulary used within the metadata record, but simply the inability to find and access the metadata in question in an efficient, seamless way. RIs work diligently to provide portals via which researchers can find and access the data they are responsible for curating, but often this still carries the requirement to visit the RI's specific data portal and manually make the relevant request. Many RIs do contribute specific classes of data to aggregators (such as Copernicus), but often data is still kept in specific silos, retrievable yet isolated.

There are a number of ways to address this problem, including the construction of more cross-RI joint catalogues to expose RI resources to broader communities, but here we focus on a single approach, which is that of linked data [23]. The linked data approach is to leverage Semantic Web technologies to publish resource metadata in an open, retrievable way that can easily be cross-referenced by others in their own published (meta)data, so creating a wide-spanning distributed knowledge graph that can be navigated programmatically by discovery and query services. If RI resource metadata is available online as linked data of this (or a functionally-equivalent) form, then semantic linking might be reducible to simply creating more links between local knowledge graphs to build or add to a global, cross-RI knowledge graph. We refer to this approach to semantic linking as *semantic bridging*.

Semantic bridging is mainly applicable where there is a commonality of data format, but there is a need for additional semantic context for computational services to be able to infer that information from two or more sources actually relates to the same artefact. One case might involve relating entities referred to in the description of an RI process or subsystem with existing metadata regarding those entities, perhaps hosted by the very RI being described—for example the ENVRI Knowledge Base might refer to datasets provided by the ICOS RI, for which RDF information is provided by the

ICOS Carbon Portal[19]. Simply including the URI links used by the Carbon Portal in the metadata provided by the ENVRI Knowledge Base would allow any system querying the knowledge base to follow through to the Carbon Portal without manual intercession by a human investigator.

Another example involves the bridging between online provenance data structured according to the W3C PROV standard [24] with an OIL-E description of an infrastructure process such that there are direct links between a provenance dataset and a reference to that dataset in the ENVRI Knowledge Base, allowing queries to be distributed across both datasets. We can use SHACL rules [25] to describe how to generate additional RDF triples classifying entities in the provenance graph using OIL-E, and then automatically assert them in the knowledge base, with pointers back to the provenance data. Figure 5 provides a (simplified) example of such a rule, for relating a PROV activity to an OIL-E behaviour.

```
:ProvActivityMappingShape a sh:NodeShape ;
  sh:targetClass prov:Activity ;
  sh:rule [
    a sh:SPARQLRule ;
    rdfs:label "Map PROV entities onto OIL-E science view.";
    sh:prefixes prov: , oil: ;
    sh:construct """
      PREFIX oil: <http://www.oil-e.net/ontology/oil-base.owl#>
      PREFIX prov: <http://www.w3.org/ns/prov#>
      CONSTRUCT {
        $this a oil:Behaviour .
      } WHERE {
        $this a prov:Activity .
        …
      }"""
  ] .
```

Fig. 5. Sample SHACL rule for mapping PROV-O activities to OIL-E science view behaviours.

SHACL allows us to define the conditions under which to produce new data (via SPARQL construct queries) that can be inserted into the ENVRI Knowledge Base and used by a distributed query broker to find and retrieve information from the provenance store as if it were an extension of the RI description in the knowledge base. The main challenge is the construction of 'conditional' rules that allow for the different kinds of provenance graph, as even within the PROV standard there are various ways to build a provenance trace depending on the primary concerns of the developer.

In this case, the linking of PROV data to RI specifications in OIL-E confers another benefit, which is that we can validate whether the structure of PROV traces (involving interactions between Agents, Activities and Events) matches the form of the RI provenance tracking behaviour as defined using ENVRI RM. Thus such bridging allows for possible validation of the provenance graph based on OIL-E definitions, and allows

[19] https://www.icos-cp.eu/.

for a distributed query broker to potentially access the provenance data directly via the bridging data in the knowledge base.

5 Discussion

Semantics in heterogeneous distributed systems are plagued by many of the problems of knowledge representation in general, such as how to achieve adequate computability, consistency and completeness in data coming from various sources produced in various different ways. The Semantic Web provides one means to represent and publish information in a lightweight, machine-actionable way, but it does not remove the necessity to deal with these problems, adding to them further issues of data redundancy, unreliability and limited performance versus more tightly integrated data models such as used in relational databases. Considerable attention has been given to the openness, extensibility and computability of Semantic Web standards, weighing different options (e.g. the use of SKOS over OWL [26, 27] to reduce the complexity of specifying controlled terminologies and their relationships). The use of linked data for describing resources (of all kinds) is already well-established, with research now focusing on different approaches to generating linked data from various sources and with how to navigate and query distributed information. Examples of such recent research include the generation of a navigable Graph of Things from an array of live IoT data sources [28] and the use of crowdsourcing to provide real-time transport data in rural areas [29], both topics with relevance to how RIs gather and expose field observations acquired via sensors or human experts. On the topic of distributed query, various frameworks have been proposed such as LDQL [30] and LILAC [31], which may make linked data based search over distributed metadata catalogues more practical and efficient than is currently the case.

Most geospatial technologies currently used by environmental and Earth science RIs have been developed independently of the Semantic Web, with recommendations such as INSPIRE[20] being mostly technically (albeit not conceptually) disjoint from it. Instead, bodies such as OGC have produced a number of open standards for Web access of metadata which are in common use by many RIs, usually brokered via software such as GeoNetwork[21]. This poses a barrier for integration of geospatial catalogues published via technologies such as CSW or OAI-PMH into the Semantic Web, and adaptors are still needed to query such data sources and present responses in RDF format (e.g. [32]), though there are also unifying technology proposals such as OGC's GeoSPARQL[22] to at least partially address this gap.

For mapping between a modest set of standards, manual mapping with tool support remains most practical, but automation may help to accelerate the construction of new mappings, provided that the precision and recall of such mappings can be made sufficient (most likely at present by mixing machine learning techniques with expert supervision and refinement). While how best to map metadata between different terminologies and

[20] https://inspire.ec.europa.eu/.
[21] https://geonetwork-opensource.org/.
[22] http://www.opengeospatial.org/standards/geosparql.

models remains an open question, automated mapping techniques can at least be (somewhat) objectively evaluated by comparing performance against human-crafted ontology sets covering the same domain (e.g. OntoFarm for conference organisation [33]). Given that syntactic mapping is still a big part in building semantic mapping, it is necessary to consider not only synonymous and otherwise-related terms in English, but also multilingual support; Bella et al. [34] provide an example of how to conduct mapping not rooted solely in measuring against a base English syntax.

Metadata descriptions of research assets are not limited to 'characteristic' information; provenance data (which might be structured according to a standard such as PROV-O) for data products and processes are also an important target for semantic linking, especially for creating unified (or at least *unifiable*) records of how research assets are used and where they came from; such records may be generated from scientific workflow management systems with provenance support [35, 42]. Such systems remain important for reproducible data science; most scientific investigations must follow a clear workflow, and there have been a number of workflow management systems developed with different characteristics and target applications [36], several of which have been applied to data science [37]. The use of ontologies for verification and validation of workflows has already been explored (e.g. [38]), and the ability to construct and validate such workflow specifications using metadata from service catalogues demonstrates that the cataloguing problem is not wholly centred on datasets.

The need to use controlled vocabulary within scientific datasets is self-evident, as is the need for standard schemes to describe such datasets, but it is still difficult for researchers, particularly researchers working independently, to even identify the best terminologies to use with their data (e.g. to use in particular data fields or to annotate their data), let alone to apply them in order to make it easy to integrate and interpret as part of a larger data corpus. For example, various repository services now exist that host controlled vocabularies and ontologies for use by researchers (e.g. BioPortal[23] and AgroPortal[24]), but there is a lack of standard tools for discovering these terminological resources and evaluating their appropriateness to researchers' own needs and those of their communities. This represents a fundamental problem that must also be addressed when considering approaches to semantic linking—there is not much value in harmonising standards that researchers themselves are not fully aware of, nor is it useful if the mappings, translation services and other products of harmonisation are themselves invisible to the scientific community. This is another area in which community-driven initiatives such as ENVRI and RDA might prove invaluable.

6 Conclusion

Semantic linking is a topic of considerable importance for the effective realisation of seamless interoperability between research infrastructure, needed to achieve the kind of open data and open science research commons being now promoted by initiatives such as

[23] https://bioportal.bioontology.org/.
[24] http://agroportal.lirmm.fr/.

DataONE[25] and EOSC. While standardisation of metadata schemes, protocols and terminology across different areas of domain science can and does enhance interoperability between different data and resource providers (and can be considered the main driver of such interoperability in practice), it is clear that there will remain necessary disparities between communities driven by their need to attend to the specific requirements of their own researchers and as a byproduct of legacy technology choices. As long as these disparities exist, there will be a need for some kind of translation of data between two or more data models, executed at the intersection between different services operating in different semantic domains. Thus, the examination of different techniques and the adoption of specific technologies to perform these translations on demand remains an important facet in the promotion of interoperability within and across research infrastructure.

There are various ways to enhance the semantic interoperability of data and services provided by RIs. In this chapter we have provided an overview of some techniques that were investigated in the context of the Horizon 2020 ENVRIplus project:

- **Semantic contextualisation**, where we increase the body of contextual information available about the resources and data that already exists by applying ontologies and other meta-models to describe those resources and data in different ways, increasing the number of facets by which we can explore them.
- **Semantic enrichment**, where we use controlled vocabularies to further classify and annotate existing metadata records to make search and discovery easier.
- **Semantic mapping**, where we develop transformation models by which to fully convert information described in one data model into another, minimising information loss.
- **Semantic bridging**, where we generate additional linking data to 'bridge' between two online data sources, leveraging the power of linked data to permit distributed querying of a wider network of knowledge.

Our overview of these techniques only scratches the surface of what is required to improve semantic interoperability and what is currently being done by various communities and community initiatives. Practical semantic alignment requires considerable attention on the part of semantic modellers and RI developers. In particular, it is necessary to identify where such attention should be focused: the specific standards, protocols, models and terminologies that would provide the greatest benefit if linked; as well as the specific intermediary transformation services which, if deployed in the right place, would expedite data integration and service composition for the most relevant scientific use-cases. To make these judgements, it's important to understand exactly how these semantic resources are being used already by RIs and research communities, and where interdisciplinary research is being stymied by a lack of standardisation or interoperability.

Acknowledgements. This work was supported by the European Union's Horizon 2020 research and innovation programme via the ENVRIplus project under grant agreement No 654182.

[25] https://www.dataone.org/.

References

1. Zhao, Z., et al.: Reference model guided system design and implementation for interoperable environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science, Munich, Germany, pp. 551–556. IEEE (2015). https://doi.org/10.1109/eScience.2015.41
2. Nieva de la Hidalga, A., et al.: The ENVRI Reference Model (ENVRI RM) version 2.2, November 2017. https://doi.org/10.5281/zenodo.1050349
3. Martin, P., et al.: Open information linking for environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science (e-Science), pp. 513–520. IEEE (2015). https://doi.org/10.1109/eScience.2015.66
4. Martin, P., Chen, Y., Hardisty, A., Jeffery, K., Zhao, Z.: Computational challenges in global environmental research infrastructures, chap. 12. In: Chabbi, A., Loescher, H.W. (eds.) Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities, pp. 305–340. CRC Press (2017). https://zenodo.org/record/3361569
5. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**, 1–9 (2016)
6. ISO 19115-1:2014: Geographic information—Metadata—Part 1: Fundamentals. ISO standard, International Organization for Standardization (2014)
7. ISO 19139:2007: Geographic information—Metadata—XML schema implementation. ISO/TS standard, International Organization for Standardization (2007)
8. Erickson, J., Maali, F.: Data catalogue vocabulary (DCAT). W3C recommendation. W3C (2014). http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
9. Nebert, D., Voges, U., Bigagli, L.: OGC catalogue services 3.0—general model. OGC implementation standard. Open Geospatial Consortium (2016). http://docs.opengeospatial.org/is/12-168r6/12-168r6.html
10. Lagoze, C., Van de Sompel, H.: The making of the open archives initiative protocol for metadata harvesting. Libr. Hi-tech **21**(2), 118–128 (2003)
11. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
12. W3C OWL Working Group: OWL 2 web ontology language. W3C recommendation. W3C (2012). https://www.w3.org/TR/2012/REC-owl2-overview-20121211/
13. Bechhofer, S., Miles, A.: SKOS simple knowledge organization system reference. W3C recommendation. W3C (2009). http://www.w3.org/TR/2009/REC-SKOS-reference-20090818/
14. W3C SPARQL Working Group: SPARQL overview. W3C recommendation. W3C (2103). http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/
15. Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F.: An ontology for describing and synthesizing ecological observation data. Ecol. Inform. **2**(3), 279–296 (2007)
16. Schentz, H., Peterseil, J., Bertrand, N.: EnvThes—interlinked thesaurus for long term ecological research, monitoring, and experiments. In: EnviroInfo, pp. 824–832 (2013)
17. The European Commission: Realising the European Open Science Cloud. The European Commission (2016). https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf
18. Jörg, B.: CERIF: the common European research information format model. Data Sci. J. **9**, 24–31 (2010)
19. Marketakis, Y., et al.: X3ML mapping framework for information integration in cultural heritage and beyond. Int. J. Digital Libr. **18**(4), 301–319 (2016). https://doi.org/10.1007/s00799-016-0179-1

20. Martin, P., Remy, L., Theodoridou, M., Jeffery, K., Zhao, Z.: Mapping heterogeneous research infrastructure metadata into a unified catalogue for use in a generic virtual research environment. Future Gener. Comput. Syst. **101**, 1–13 (2019). https://doi.org/10.1016/j.future.2019.05.076
21. Theodoridou, M., Ivanovic, D., Martin, P., Remy, L., Muckensturm, M.: X3ML mappings from common metadata schemes to CERIF RDF (2019). https://doi.org/10.5281/zenodo.2548732
22. Sundara, S., Das, S., Cyganiak, R.: R2RML: RDB to RDF mapping language. W3C recommendation. W3C (2012). http://www.w3.org/TR/2012/REC-r2rml-20120927/
23. Berners-Lee, T.: Linked data. W3C Design Issues (2006). https://www.w3.org/DesignIssues/LinkedData.html. Accessed 26 Feb 2018
24. Groth, P., Moreau, L.: PROV-overview. W3C note. W3C (2013). http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
25. Kontokostas, D., Knublauch, H.: Shapes constraint language (SHACL). W3C recommendation. W3C, July 2017. https://www.w3.org/TR/2017/REC-shacl-20170720/
26. Stellato, A.: Dictionary, thesaurus or ontology? Disentangling our choices in the semantic web jungle. J. Integr. Agric. **11**(5), 710–719 (2012)
27. Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., Summers, E.: Key choices in the design of simple knowledge organization system (SKOS). Web Semant.: Sci. Serv. Agents World Wide Web **20**, 35–49 (2013)
28. Le-Phuoc, D., Quoc, H.N.M., Quoc, H.N., Nhat, T.T., Hauswirth, M.: The graph of things: a step towards the live knowledge graph of connected things. Web Semant.: Sci. Serv. Agents World Wide Web **37**, 25–35 (2016)
29. Corsar, D., Edwards, P., Nelson, J., Baillie, C., Papangelis, K., Velaga, N.: Linking open data and the crowd for real-time passenger information. Web Semant.: Sci. Serv. Agents World Wide Web **43**, 18–24 (2017)
30. Hartig, O., Pérez, J.: LDQL: a query language for the web of linked data. Web Semant.: Sci. Serv. Agents World Wide Web **41**, 9–29 (2016)
31. Montoya, G., Skaf-Molli, H., Molli, P., Vidal, M.E.: Decomposing federated queries in presence of replicated fragments. Web Semant.: Sci. Serv. Agents World Wide Web **42**, 1–18 (2017)
32. Patroumpas, K., Georgomanolis, N., Stratiotis, T., Alexakis, M., Athanasiou, S.: Exposing INSPIRE on the semantic web. Web Semant.: Sci. Serv. Agents World Wide Web **35**, 53–62 (2015)
33. Zamazal, O., Svátek, V.: The ten-year OntoFarm and its fertilization within the onto-sphere. Web Semant.: Sci. Serv. Agents World Wide Web **43**, 46–53 (2017)
34. Bella, G., Giunchiglia, F., McNeill, F.: Language and domain aware lightweight ontology matching. Web Semant.: Sci. Serv. Agents World Wide Web **43**, 1–17 (2017)
35. Altintas, I., Barney, O., Jaeger-Frank, E.: Provenance collection support in the Kepler scientific workflow system. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 118–132. Springer, Heidelberg (2006). https://doi.org/10.1007/11890850_14
36. Liew, C.S., Atkinson, M.P., Galea, M., Ang, T.F., Martin, P., Hemert, J.I.V.: Scientific workflows: moving across paradigms. ACM Comput. Surv. **49**(4), 66:1–66:39 (2016). https://doi.org/10.1145/3012429, http://doi.acm.org/10.1145/3012429
37. Mork, R., Martin, P., Zhao, Z.: Contemporary challenges for data-intensive scientific workflow management systems. In: Proceedings of the 10th Workshop on Workflows in Support of Large-Scale Science. ACM (2015). https://doi.org/10.1145/2822332.2822336
38. Miksa, T., Rauber, A.: Using ontologies for verification and validation of workflow-based experiments. Web Semant.: Sci. Serv. Agents World Wide Web **43**, 25–45 (2017)

39. Zhao, Z., et al.: Knowledge-as-a-Service: a community knowledge base for research infrastructures in environmental and earth sciences. In: 2019 IEEE World Congress on Services (SERVICES), Milan, Italy, pp. 127–132. IEEE (2019). https://doi.org/10.1109/SERVICES.2019.00041
40. Liao, X., Zhao, Z.: Unsupervised approaches for textual semantic annotation, a survey. ACM Comput. Surv. **52**, 1–45 (2019). https://doi.org/10.1145/3324473
41. Zhao, Z., et al.: Scientific workflow management: between generality and applicability. In: Fifth International Conference on Quality Software (QSIC 2005), Melbourne, Australia, pp. 357–364. IEEE (2005). https://doi.org/10.1109/QSIC.2005.56
42. el Khaldi Ahanach, E., Koulouzis, S., Zhao, Z.: Contextual linking between workflow provenance and system performance logs. In: 2019 15th International Conference on eScience (eScience), San Diego, CA, USA, pp. 634–635. IEEE (2019). https://doi.org/10.1109/eScience.2019.00093
43. Liao, X., Bottelier, J., Zhao, Z.: A column styled composable schema matcher for semantic data-types. Data Sci. J. 18–25 (2019). https://doi.org/10.5334/dsj-2019-025