Mapping metadata from different research infrastructures into a unified framework for use in a virtual research environment

Paul Martin*, Laurent Remy†, Maria Theodoridou‡, Keith Jeffery§ and Zhiming Zhao*
*Institute for Informatics, University of Amsterdam, Amsterdam, Netherlands
†euroCRIS / IS4RI, France
‡Institute of Computer Science, Foundation for Research and Technology—Hellas, Heraklion, Greece
§Keith G Jeffery Consultants, United Kingdom
Emails: {p.w.martin, z.zhao}@uva.nl, lremy@is4ri.com, maria@ics.forth.gr, keith.jeffery@keithgjefferyconsultants.co.uk

***Abstract*—Virtual Research Environments (VREs) augment research activities by integrating tools for data discovery, data retrieval, workflow management and researcher collaboration, often coupled with a specific computing infrastructure. The drive towards open data science discourages 'walled garden' solutions however, and has led to the creation of dedicated research infrastructures (RIs) that gather data and provide services to particular research communities without prejudice towards any particular science gateway or virtual laboratory technology.**

There is a need for generic VREs that can be easily customised to the needs of specific communities and coupled with the services and resources of many different RIs, but the resource metadata produced by these RIs rarely adheres perfectly to any particular standard or vocabulary, making it difficult to search and discover resources independently of their provider. Cross-RI search can be expedited by metadata mapping services that can harvest metadata published under different standards to build unified resource catalogues—such an approach poses a number of challenges however. In this paper we take the example of the VRE4EIC e-VRE metadata service, which uses X3ML mappings to build a single CERIF catalogue for describing data products and other resources provided by multiple RIs. We consider the extent to which it addresses the challenge of cross-RI search, and we also discuss how it might take advantage of semantic harmonisation efforts in the environmental science domain.

***Keywords*—virtual research environment, research infrastructure, metadata catalogue, metadata mapping.**

I. Introduction

Virtual Research Environments (VREs) [1], also known as virtual laboratories or science gateways, are one of three types of science support environment developed to support researchers in data science [2], focusing on supporting research activities on a holistic rather than infrastructural or service level. VREs provide integrated environments that typically include tools for activities such as data discovery and retrieval, collaboration, process scheduling and workflow management, and many are coupled with a particular computational infrastructure, often making use of public e-infrastructures or the Cloud. Data are brought into that infrastructure and manipulated via a particular data processing platform or scientific workflow management system [3]—however this approach is contrary to the recent drive towards open science and open data, which discourages 'walled garden' solutions.

Increasingly, what we observe instead is the creation of dedicated research infrastructures (RIs) that aggregate and curate scientific data (including real-time observations) for a particular research community, which then provide access to these data via unified services [4], usually without prejudice towards any particular VRE. Complicating this matter, there is now a substantive push to better integrate these efforts into a cohesive multidisciplinary commons for open science and open research data, as embodied by initiatives such as the European Open Science Cloud (EOSC) [5].

Developing generic VREs that can be easily coupled with different RIs and customised for specific communities is a goal of many recent research projects, including VRE4EIC[1] and BlueBRIDGE[2], and is particularly challenging given the lack of conformity of standards and vocabularies in environmental science and similar domains. Significant software engineering effort is often required on the behalf of data scientists to build specific adaptors for such couplings, but even then it remains crucial to provide the capability to search across different RIs for similar data products or services to support integrative and transdisciplinary research. This entails a complex interaction between a VRE and multiple RIs, distributing queries through multiple adaptors and then aggregating the results—or else a prior harvesting of metadata from all providers to allow preliminary queries to be conducted on a single logical catalogue.

In this paper we investigate how the use of a flexible metadata mapping and publication service can expedite the coupling of a VRE with RI resources using different metadata schemes to provide cross-RI metadata search and discovery. As a case study, we take the VRE4EIC metadata service, developed as a building block for an RI-agnostic VRE, and we detail how X3ML mappings [6] from standards such as ISO 19139 [7] and DCAT [8] to CERIF [9] are used to automatically ingest metadata published by different RIs to

[1]https://www.vre4eic.eu/
[2]http://www.bluebridge-vres.eu/

produce a single resource catalogue. We weigh the benefits of this approach and discuss some ways in which such catalogues can be further augmented, for example to facilitate semantic search based on the harmonisation of vocabularies used for describing ecosystem and biodiversity data.

II. Background

Modern environmental research depends on the collection and analysis of large volumes of data gathered via sensors, observations, simulations and experimentation. Researchers are called upon to address societal challenges that are inextricably tied to the stability of our native ecosystems such as food security and climate management, challenges intrinsically interdisciplinary in nature, requiring collaboration across traditional disciplinary boundaries. The role of RIs in this context is to support researchers with data, platforms and tools, but no single RI can hope to encompass the full research ecosystem. 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 RIs, allowing them to collaborate and conduct their research more effectively.

Publishing metadata about resources online (indicating type, coverage, provenance, *etc.*) allows RIs to advertise their facilities and researchers to browse and discover data and other resources useful to their research. While there exist standards such as ISOs 19115 [10] and 19139 [7] for geospatial metadata however, the implementation of such standards by RIs can be somewhat idiosyncratic. Resource catalogues themselves can be described using standards such as DCAT [8] and harvested via CSW [11] or OAI-PMH [12], but many RIs also use Semantic Web [13] technologies such as OWL [14] and SKOS [15] to describe their resources, adapting ontologies such as OBOE [16] (for observations) and vocabularies such as EnvThes [17] (for ecology) to meet their own community's needs. Harmonisation of vocabulary and metadata between RIs thus remains a concern, with cluster projects such as ENVRIplus[3] working to promote common models. Concurrently, initiatives like RDA[4] address broader research data management issues such as metadata standards cataloguing, standards for data collections and interoperability between repositories, providing recommendations to such projects.

From the VRE perspective, it is necessary to be pragmatic when coupling with the services provided by RIs, a process that can also be assisted by the use of standard models and vocabularies. Jeffery et al. [18] define a reference architecture for enhanced VREs ('e-VREs') able to work with many different RIs and e-infrastructures. In this architecture, microservices are used to implement each of six key building blocks split across three tiers of operation, as shown in Figure 1 for the case of the metadata management. Meanwhile Nieva et al. [19] describe a reference model (ENVRI RM) for environmental science RIs, defining their archetypical elements in the context of the research data lifecycle. Being based on

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

[4]https://rd-alliance.org/

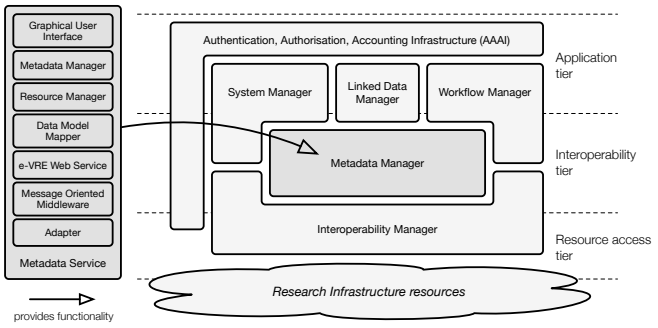

Fig. 1. Providing a metadata service: the recommended microservice stack to implement the metadata manager in the e-VRE reference architecture.

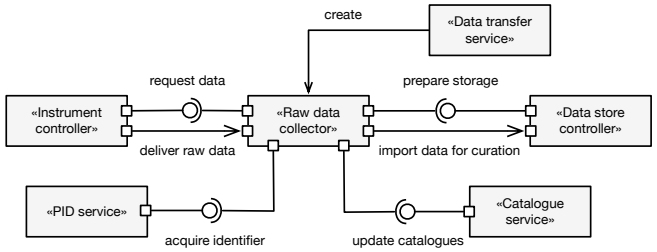

Fig. 2. A computational view of raw data acquisition: ENVRI RM specifies components and activities using UML (in this case, a component diagram).

RM-ODP [20], it models RIs from five viewpoints: science, information, computation, engineering and technology. Each view has its own concerns that correspond to those of the other views, and is able to describe various key RI activities (*e.g.* Figure 2). Open Information Linking for Environmental RIs (OIL-E) [21] is a small set of OWL specifications based on ENVRI RM that provide an upper ontology for RI descriptions and which can be used to contextualise different kinds of RI asset from an architectural or interaction-based perspective—as opposed to being a general-purpose ontology for describing scientific phenomena like BFO [22]. A conceptual model with a similar focus on the products and tools of research rather than on scientific classification itself is CERIF [9], a European standard for describing research information systems. CERIF provides a framework for describing relationships between people, projects, tools and research products (and more), and has been applied to describing solid earth science RIs [23].

These models provide both the means to talk about research support environments such as VREs and RIs in a standard way, but can also be leveraged as a means to better classify different kinds of resource as part of a faceted search mechanism, as we shall discuss later in Section IV. For now, we consider how VREs can be constructed that support rather than are hindered by the heterogeneity of RI resources and resource metadata, and how a VRE can facilitate cross-RI search and discovery.

III. Methodology and challenges

According to Jeffery et al. [18], VREs can retrieve descriptions of RIs' resources either via separate interfaces with each

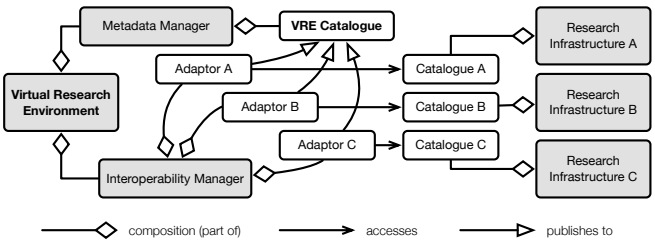

Fig. 3. An e-VRE produces adaptors to harvest and convert metadata from different catalogues, building a common metadata catalogue for its users.

RI's own resource catalogue, or via a joint resource catalogue that already encompasses all of the RIs' resources. The former approach relies on the construction of separate discovery and access interfaces with every RI, and makes it difficult to search over multiple RI resource catalogues simultaneously, requiring the translation and distribution of queries over every interface. Meanwhile, the latter approach simplifies search and discovery, but requires initial harvesting of metadata from all separate RI catalogues, translation of all metadata into a single common denominator standard, and careful management as the number of original data sources scales upwards.

In terms of the e-VRE reference architecture [18], there are a few needed steps to harvest resource metadata from an RI:

1) A resource catalogue provided by an RI is identified for harvesting. Identification might be performed by a discovery service, or be part of the manual configuration of a customised VRE metadata catalogue.
2) The VRE's *interoperability manager* must provide an adaptor for the given resource catalogue—essentially, the VRE must have the means to interact with the catalogue via the correct protocol (*e.g.* OAI-PMH or SPARQL [24]), but also have a model for (at least partially) *mapping* metadata retrieved from the source scheme to the scheme used internally by the VRE.
3) The adaptor can then be used to harvest metadata records from the source, mapping them into a format suitable for ingestion into the VRE's own metadata catalogue.
4) This ingested data is then made available to users of the VRE via its own search and query interface.

The main entities involved in this process are shown in Figure 3. In this example, the result is that metadata can now be harvested by the VRE's *metadata manager* using the adaptors provided by the interoperability manager. This activity may be a one-off event, but more likely the metadata harvested will need to be periodically updated.

Whatever the chosen approach however, any VRE cataloguing solution should try to address certain challenges:

1) How best to discover new resources—a VRE catalogue may be carefully curated for a given community, but even if automation is rejected, there should be a clear process for how to expand the catalogue.
2) How to ensure the freshness of catalogue data—ensuring that updates to source catalogues are propagated to VRE catalogues in reasonable time.
3) How to manage the underlying catalogue schema—given new vocabularies, standards or simply evolution in how standards are applied, how to update the model underlying a catalogue without losing existing data coherence.
4) How to manage ever larger quantities of data—whether by relying on more capable database technologies, distribution of the catalogue, or dynamic construction of the catalogue 'on demand' based on prior queries.

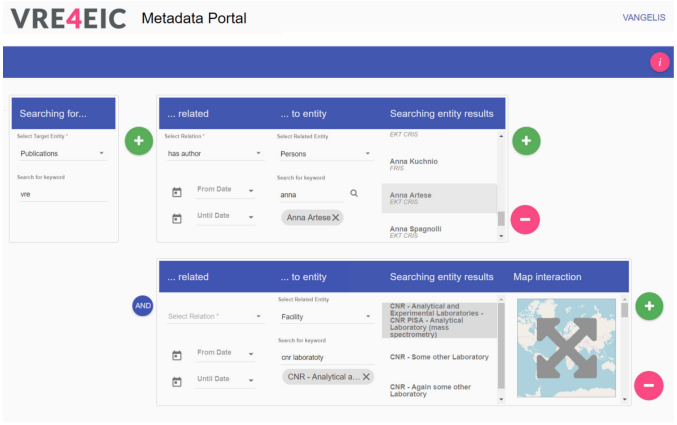

Fig. 4. The VRE4EIC metadata portal: searching for data publications published by Anna Artese through CNR Pisa's mass spectrometry analytical laboratory.

In light of these challenges, we consider a particular implementation of the resource metadata harvesting approach described above based on certain key technologies.

IV. Implementation

The VRE4EIC Metadata Portal has been developed in accordance with the e-VRE reference architecture, providing the necessary components to implement the *metadata manager* functionality. The purpose of the portal is to provide faceted search over catalogue data harvested from multiple RIs, aggregated within a single CERIF-based VRE catalogue. Search is based on the composition of queries based on the *context* of the research data, filtering by organisations, projects, sites, instruments, people, *etc.*, for example as shown in Figure 4. The portal supports map-based search, the export and storing of specific queries, and the export of results in various formats. The CERIF catalogue itself is implemented in RDF (based on an OWL ontology) as a Blazegraph[5] triple store and is structured according to CERIF version 1.6[6].

Metadata harvested from external sources is converted to CERIF RDF using the X3ML mapping framework [6]. The mapping process is as illustrated in Figure 5:

1) Sample metadata, along with their corresponding metadata schemes are retrieved for analysis.
2) Mappings are defined that dictate the transformation of the selected RDF and XML based schemas to CERIF.

[5]https://www.blazegraph.com/
[6]https://www.eurocris.org/cerif/main-features-cerif

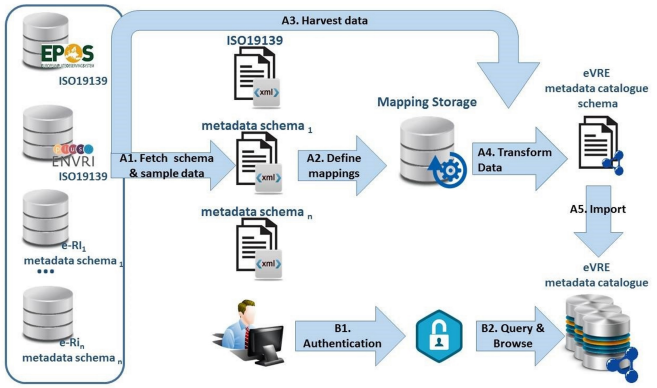

Fig. 5. e-VRE metadata acquisition and retrieval workflow: metadata records are acquired from multiple sources, mapped to CERIF RDF and stored in the VRE catalogue; authenticated VRE users query data via the e-VRE.

1	D	../result		Product		
1.1	P	id	has_identifier FederatedIdentifier has_id_value	[has_URI]	[string = "resultProductIdentifier"]	exists(text())
	R	id		string		exists(text())
1.2	P	title	has_name			exists(text())
	R	title	22-rdf-syntax-ns#PlainLiteral			exists(text())
1.3	P	url	has_URI			exists(text())
	R	url	XMLSchema#string			exists(text())
1.4	P	notes	has_description			exists(text())
	R	notes	22-rdf-syntax-ns#PlainLiteral			exists(text())
1.5	P	name	has_identifier FederatedIdentifier has_id_value	[has_URI]	[string = "resultProductName"]	exists(text())
	R	name	string			exists(text())

Fig. 6. Example of mapping rules generated in 3M: result metadata in CKAN is mapped to a CERIF product with data properties corresponding to each possible attribute in the original CKAN XML scheme.

3) Metadata is retrieved from different data sources in their native format, *e.g.* as ISO 19139 or CKAN[7] data.
4) These mappings are used to transform the source data into CERIF format.
5) The transformed data are ingested into the CERIF metadata catalogue.

Once ingested, these data become available to users of the metadata portal, who can query and browse data upon authentication by the front-end authentication/authorisation service.

X3ML mappings are described using the 3M Mapping Memory Manager[8]. Mappings are described by mapping rules relating *subject-property-object* triples from the source scheme to equivalent structures in the target scheme, subject to various syntactic conditions, as illustrated in Figure 6. 3M supports the specification of generators to produce identifiers for new concepts constructed during translation of terms, and provides test and analytics facilities. Mappings into CERIF RDF have been produced for Dublin Core, CKAN, DCAT-AP, and ISO 19139 metadata, as well as RI architecture descriptions in OIL-E, as part of the technical output of the VRE4EIC project[9].

[7]https://ckan.org/

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

[9]Mappings are accessible at http://www.ics.forth.gr/isl/3M-VRE4EIC, username 'vre4eicGuest' and password 'vre4eic'.

In summary, the Portal has many desirable characteristics: a flexible model in CERIF for integrating heterogeneous metadata, a tool-assisted metadata mapping pipeline to easily create or refine metadata mappings or refine existing mappings, and a mature technology base for unified VRE catalogues. What we foresee more development needed in is the discovery of new resources and the acquisition of updates. In this respect, RI-side services for advertisement of new resources or updates to which a VRE can subscribe to trigger automated ingestion of new or modified metadata would be particularly useful.

The VRE4EIC Metadata Portal has been provided as a demonstrator to the cluster of environmental science RIs in Europe via the ENVRIplus project as well as directly to the European Plate Observing System (EPOS)[10], with sample data harvested from a subset of those RIs. Evaluation of the demonstrator indicates a number of possible avenues of development, particularly with regard to supporting richer cross-RI search, the two most noteworthy here being:

1) Further exploitation of CERIF's semantic layer.
2) Integration of semantic search facilities.

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. Almost every entity and relation can be assigned though a classification that indicates a particular semantic interpretation (*e.g.* that the relationship between a *Person* and a *Product* is that of a creator), allowing a CERIF database to be enriched with concepts from an external semantic model (or several linked models).

The vocabulary provided by OIL-E[11] has been identified within VRE4EIC as a means to further classify objects in CERIF in terms of their role in an RI, *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 provides additional *operational* context for faceted search (*e.g.* identifying which processes generated a given data product), but providing additional context into the *scientific* context for data products (*e.g.* categorising the experimental method applied or the branch of science to which it belongs) is also necessary. Environmental science RIs such as AnaEE[12] and LTER-Europe[13] are actively developing better vocabularies for describing ecosystem and biodiversity research data, building upon existing SKOS vocabularies. The AnaEE data vocabulary (anaeeThes) [25] and LTER's environmental thesaurus EnvThes [17] have mappings to other established domain vocabularies such as Agrovoc[14] and GEMET[15]. These RIs are now collaborating with other RIs involved in ENVRIplus to harmonise their vocabularies in order to provide *semantic linking* between terms used in their respective sub-domains.

[10]https://www.epos-ip.org/

[11]http://oil-e.net/ontology/

[12]https://www.anaee.com/

[13]http://www.lter-europe.net/lter-europe

[14]http://aims.fao.org/standards/agrovoc

[15]http://www.eionet.europa.eu/gemet/

The identification of synonymous, subsuming and intersecting terms (and the publication of links on the Semantic Web) provides 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.

V. Discussion

The use of linked data [26] 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—for example, recent research includes the generation of a navigable Graph of Things from an array of live IoT data sources [27] and the use of crowdsourcing to provide real-time transport data in rural areas [28], 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 languages/frameworks have been proposed such as LDQL [29] and LILAC [30], which may make linked data based search over distributed catalogues more practical and efficient than is currently the case.

The Semantic Web is plagued by many of the problems of knowledge representation in AI including computability, inconsistency and incompleteness, adding data redundancy, unreliability and limited performance versus more tightly integrated data models. 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 [31], [32]). Most geospatial technologies used by environmental science RIs today have been developed independently of the Semantic Web however, with recommendations such as INSPIRE[16] being mostly disjoint from it, though technologies such as OGC's GeoSPARQL[17] attempt to address this. This poses a barrier for integration of geospatial catalogues published via 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.* [33]).

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. How to best map between ontologies (or other kinds of schema) remains an open question, but mapping techniques can be evaluated by comparing performance against ontology sets covering the same domain (*e.g.* OntoFarm for conference organisation [34]). Multi-lingual support is also important in collaboration; for example Bella et al. [35] address how to conduct mapping based on more than just English syntax.

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

[17]http://www.opengeospatial.org/standards/geosparql

It is not only resource metadata that can be usefully accessed via a VRE. Access to provenance data (which might be structured according to a standard such as PROV-O [36]) for data products and processes would also be useful to researchers, and VREs can also be contributors of provenance data via their own workflow systems (*e.g.* for Kepler [37]). CERIF is able to represent time-bounded role-based semantic relationships, but the source metadata provided by RIs still often lacks this kind of information; the adoption of standardised and ubiquitous provenance by RIs would address this either by enriching the basic metadata for resources, or by providing additional sources of provenance data that could be integrated with the base metadata when producing unified catalogues.

The e-VRE reference architecture also addresses the need for a *workflow manager* component, for composing processing tasks in series or parallel on available computational resources. Most scientific investigations do follow a clear workflow, and there have been a number of workflow management systems developed with different characteristics and target applications [38], several of which have been applied to science [39]. The use of ontologies for verification and validation of workflows has already been explored (*e.g.* [40]), 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.

VI. Conclusion

In this paper we linked the development of VREs (also science gateways and virtual laboratories) to the outgrowth of dedicated RIs in Europe and beyond, and argued the need for new VREs that can be freely coupled with different RI resources based on the requirements of researchers and the evolving data research environment. We asserted that metadata mapping is needed to facilitate cross-RI search and discovery due to the diversity of metadata schemes, vocabularies and protocols used to access resource catalogue data published by different RIs, and furthermore that it is useful to be able to aggregate distributed resource metadata into a single logical catalogue. We outlined a methodology for building such a catalogue based on the e-VRE reference architecture and the adoption of a robust metadata mapping pipeline for handling heterogeneous data sources. We provided an example in the VRE4EIC Metadata Portal of how the methodology is applied, using CERIF as a framework for aggregating resource metadata from different metadata catalogues provided by EPOS and ENVRIplus. We described the application of X3ML mappings, constructed using the 3M editor, to translate ISO 19139 XML, CKAN, Dublin Core, DCAT-AP and OIL-E data into CERIF RDF for ingestion into a CERIF catalogue. We considered how the CERIF semantic layer can be augmented with vocabulary from OIL-E to further contextualise research entities, and how recent semantic harmonisation work in environmental science RIs can further augment the capabilities of VREs as clients for semantic faceted search of RI resources. Finally, we discussed the role that some of the technologies identified have in other research literature, examined some related work, and suggested

future avenues of investigation for coupling VREs with other types of service provided by RIs, *e.g.* provenance services.

ACKNOWLEDGEMENTS

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreements 654182 (ENVRIplus project), 676247 (VRE4EIC project) and 643963 (SWITCH project).

REFERENCES

[1] L. Candela, D. Castelli, and P. Pagano, "Virtual research environments: an overview and a research agenda," *Data Science Journal*, vol. 12, pp. 75–81, 2013.

[2] Z. Zhao, P. Martin, C. de Laat, K. Jeffery, A. Jones, I. Taylor, A. Hardisty, M. Atkinson, A. Zuiderwijk, Y. Yin, and Y. Chen, "Time critical requirements and technical considerations for advanced support environments for data-intensive research," in *2nd International workshop on Interoperable infrastructures for interdisciplinary big data sciences (IT4RIs 16), in the context of IEEE Real-time System Symposium (RTSS), Porto, Portugal*, 2016.

[3] E. Deelman, D. Gannon, M. Shields, and I. Taylor, "Workflows and e-Science: An overview of workflow system features and capabilities," *Future Generation Computer Systems*, vol. 25, no. 5, pp. 528–540, 2009.

[4] P. Martin, Y. Chen, A. Hardisty, K. Jeffery, and Z. Zhao, "Computational challenges in global environmental research infrastructures," in *Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities*, A. Chabbi and H. W. Loescher, Eds. CRC Press, 2017, ch. 12, pp. 305–340.

[5] European Commission, "Realising the european open science cloud," 2016.

[6] Y. Marketakis, N. Minadakis, H. Kondylakis, K. Konsolaki, G. Samaritakis, M. Theodoridou, G. Flouris, and M. Doerr, "X3ML mapping framework for information integration in cultural heritage and beyond," *International Journal on Digital Libraries*, pp. 1–19, 2016.

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

[8] J. Erickson and F. Maali, "Data catalog vocabulary (DCAT)," W3C, W3C Recommendation, 2014, http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/.

[9] B. Jörg, "CERIF: The common european research information format model," *Data Science Journal*, vol. 9, pp. 24–31, 2010.

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

[11] D. Nebert, U. Voges, and L. Bigagli, "OGC catalogue services 3.0—general model," Open Geospatial Consortium, OGC Implementation Standard, 2016, http://docs.opengeospatial.org/is/12-168r6/12-168r6.html.

[12] C. Lagoze and H. Van de Sompel, "The making of the open archives initiative protocol for metadata harvesting," *Library hi tech*, vol. 21, no. 2, pp. 118–128, 2003.

[13] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

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

[15] S. Bechhofer and A. Miles, "SKOS simple knowledge organization system reference," W3C, W3C Recommendation, 2009, http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

[16] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa, "An ontology for describing and synthesizing ecological observation data," *Ecological informatics*, vol. 2, no. 3, pp. 279–296, 2007.

[17] H. Schentz, J. Peterseil, and N. Bertrand, "Envthes-interlinked thesaurus for long term ecological research, monitoring, and experiments." in *EnviroInfo*, 2013, pp. 824–832.

[18] K. G. Jeffery, C. Meghini, C. Concordia, T. Patkos, V. Brasse, J. v. Ossenbruck, Y. Marketakis, N. Minadakis, and E. Marchetti, "A reference architecture for virtual research environments," in *Proceedings of the 15th International Symposium of Information Science (ISI 2017)*. Verlag Werner Hulsbusch, 2017, pp. 76–88.

[19] A. Nieva de la Hidalga, B. Magagna, M. Stocker, A. Hardisty, P. Martin, Z. Zhao, M. Atkinson, and K. Jeffery, "The ENVRI Reference Model (ENVRI RM) version 2.2, 30th October 2017," Nov. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1050349

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

[21] P. Martin, P. Grosso, B. Magagna, H. Schentz, Y. Chen, A. Hardisty, W. Los, K. Jeffery, C. de Laat, and Z. Zhao, "Open information linking for environmental research infrastructures," in *2015 IEEE 11th International Conference on e-Science (e-Science)*. IEEE, 2015, pp. 513–520.

[22] R. Arp, B. Smith, and A. D. Spear, *Building ontologies with Basic Formal Ontology*. The MIT Press, 2015.

[23] D. Bailo, D. Ulbricht, M. L. Nayembil, L. Trani, A. Spinuso, and K. G. Jeffery, "Mapping solid earth data and research infrastructures to CERIF," *Procedia Computer Science*, vol. 106, pp. 112–121, 2017.

[24] W3C SPARQL Working Group, "SPARQL 1.1 overview," W3C, W3C Recommendation, 2013, http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

[25] Anaee-France semantic group, "AnaEE Thesaurus," 2016. [Online]. Available: http://dx.doi.org/10.15454/1.4894016754286177E12

[26] T. Berners-Lee, "Linked data," *W3C Design Issues*, 2006, accessed 26th February 2018. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html

[27] D. Le-Phuoc, H. N. M. Quoc, H. N. Quoc, T. T. Nhat, and M. Hauswirth, "The graph of things: A step towards the live knowledge graph of connected things," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 37, pp. 25–35, 2016.

[28] D. Corsar, P. Edwards, J. Nelson, C. Baillie, K. Papangelis, and N. Velaga, "Linking open data and the crowd for real-time passenger information," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 43, pp. 18–24, 2017.

[29] O. Hartig and J. Pérez, "LDQL: A query language for the web of linked data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 41, pp. 9–29, 2016.

[30] G. Montoya, H. Skaf-Molli, P. Molli, and M.-E. Vidal, "Decomposing federated queries in presence of replicated fragments," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 42, pp. 1–18, 2017.

[31] A. Stellato, "Dictionary, thesaurus or ontology? disentangling our choices in the semantic web jungle," *Journal of Integrative Agriculture*, vol. 11, no. 5, pp. 710–719, 2012.

[32] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, and E. Summers, "Key choices in the design of simple knowledge organization system (SKOS)," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 20, pp. 35–49, 2013.

[33] K. Patroumpas, N. Georgomanolis, T. Stratiotis, M. Alexakis, and S. Athanasiou, "Exposing INSPIRE on the semantic web," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 35, pp. 53–62, 2015.

[34] O. Zamazal and V. Svátek, "The ten-year OntoFarm and its fertilization within the onto-sphere," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 43, pp. 46–53, 2017.

[35] G. Bella, F. Giunchiglia, and F. McNeill, "Language and domain aware lightweight ontology matching," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 43, pp. 1–17, 2017.

[36] D. McGuinness, S. Sahoo, and T. Lebo, "PROV-O: The PROV ontology," W3C, W3C Recommendation, 2013, http://www.w3.org/TR/2013/REC-prov-o-20130430/.

[37] I. Altintas, O. Barney, and E. Jaeger-Frank, "Provenance collection support in the Kepler scientific workflow system," *Provenance and annotation of data*, pp. 118–132, 2006.

[38] C. S. Liew, M. P. Atkinson, M. Galea, T. F. Ang, P. Martin, and J. I. V. Hemert, "Scientific workflows: Moving across paradigms," *ACM Comput. Surv.*, vol. 49, no. 4, pp. 66:1–66:39, Dec. 2016. [Online]. Available: http://doi.acm.org/10.1145/3012429

[39] R. Mork, P. Martin, and Z. Zhao, "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, p. 4.

[40] T. Miksa and A. Rauber, "Using ontologies for verification and validation of workflow-based experiments," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 43, pp. 25–45, 2017.