Future Generation Computer Systems 101 (2019) 1–13

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Mapping heterogeneous research infrastructure metadata into a unified catalogue for use in a generic virtual research environment

Paul Martin[a], Laurent Remy[b], Maria Theodoridou[c], Keith Jeffery[d], Zhiming Zhao[a,*]

[a] Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands
[b] euroCRIS / IS4RI, France
[c] Institute of Computer Science, Foundation for Research and Technology—Hellas, Heraklion, Greece
[d] Keith G Jeffery Consultants, United Kingdom

HIGHLIGHTS

- Open science requires seamless integration of research infrastructure resources.
- Resource metadata from different sources can be mapped into a unified catalogue.
- Metadata mappings have been created for several common metadata schemes.
- A joint catalogue for research resources has been developed with these mappings.
- The VRE4EIC Metadata Portal helps researchers find datasets from many sources.

ARTICLE INFO

Article history:
Received 31 January 2019
Accepted 29 May 2019
Available online 31 May 2019

Keywords:
Virtual research environment
Science gateway
Research infrastructure
Metadata catalogue
Metadata mapping

ABSTRACT

Virtual Research Environments (VREs), also known as science gateways or virtual laboratories, assist researchers in data science by integrating tools for data discovery, data retrieval, workflow management and researcher collaboration, often coupled with a specific computing infrastructure. Recently, the push for better open data science has led to the creation of a variety of dedicated research infrastructures (RIs) that gather data and provide services to different research communities, all of which can be used independently of any specific VRE. There is therefore a need for generic VREs that can be coupled with the resources of many different RIs simultaneously, easily customised to the needs of specific communities. The resource metadata produced by these RIs rarely all adhere to any one standard or vocabulary however, making it difficult to search and discover resources independently of their providers without some translation into a common framework. Cross-RI search can be expedited by using mapping services that harvest RI-published metadata to build unified resource catalogues, but the development and operation of such services pose a number of challenges.

In this paper, we discuss some of these challenges and look specifically at the VRE4EIC Metadata Portal, which uses X3ML mappings to build a single catalogue for describing data products and other resources provided by multiple RIs. The Metadata Portal was built in accordance to the e-VRE Reference Architecture, a microservice-based architecture for generic modular VREs, and uses the CERIF standard to structure its catalogued metadata. We consider the extent to which it addresses the challenges of cross-RI search, particularly in the environmental and earth science domain, and how it can be further augmented, for example to take advantage of linked vocabularies to provide more intelligent semantic search across multiple domains of discourse.

1. Introduction

Virtual Research Environments (VREs) [1], also known as virtual laboratories or science gateways, provide integrated online environments for researchers engaged in data science, typically including tools for activities such as data discovery, data retrieval, researcher collaboration, process scheduling on remote computing resources (such as high performance compute clusters or the Cloud), and workflow management. VREs can be considered to be one of three types of science support environment developed to support researchers in data science [2], the other two being *research infrastructures* (RIs) and *e-infrastructure*. Where RIs focus on providing access to data and services based on those data to

[*] Corresponding author.
E-mail addresses: pwmartin.research@gmail.com (P. Martin), lremy@is4ri.com (L. Remy), maria@ics.forth.gr (M. Theodoridou), keith.jeffery@keithgjefferyconsultants.co.uk (K. Jeffery), z.zhao@uva.nl (Z. Zhao).

https://doi.org/10.1016/j.future.2019.05.076
0167-739X/© 2019 Elsevier B.V. All rights reserved.

particular research communities however, and e-infrastructure focuses on providing the fundamental compute, storage and networking facilities needed to support data science, VREs focus on supporting researchers in actually using the data, services and facilities made available by the other two kinds of infrastructure.

Many VREs are coupled with certain e-infrastructures to facilitate process scheduling and storage of user data, often making use of e-infrastructures provided specifically for the research community (via initiatives such as EGI[1] or EUDAT[2]) or public Cloud platforms. Data are brought into the dedicated infrastructure, and are then explored and manipulated via a particular data processing platform or scientific workflow management system [3]. A difficulty arises where research datasets and services are distributed across multiple e-infrastructures; the recurrent question of whether it is better to move data to where the computation will occur, or to move computation to where the data are (given the ever-increasing size of datasets) means that VREs need to be more flexible as to how and where they connect to different resources. In particular, overly restrictive couplings can be seen as contrary to the recent drive towards open science and open data which discourages solutions that force users to move data and services into a closed system rather than directly engage with openly-accessible data services hosted by RIs.

While moving data and computation onto a single controlled platform has advantages – primarily, that the utilisation of resources is simplified and the quality of service made easier to sustain – what we increasingly observe instead is the construction of dedicated RIs that aggregate and curate scientific data (including real-time observations) for a particular research community, and then provide access to these data via unified services [4] independently of any particular operational environment. These RIs often provide their own portals to retrieve data, and in some cases provide limited ability to access computational infrastructure for executing processes, but they also provide APIs to allow outside agents to retrieve data and access services, allowing for other VRE systems to potentially make use of their resource offerings. These APIs are not standardised across RIs however, nor are the metadata for the resources behind them, whether those resources be data, code, models, services or something else.

To help resolve this, 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] and the Research Data Alliance (RDA).[3] These initiatives focus on interoperable infrastructure and the adoption of best practices as embodied by the FAIR data principles for *findable*, *accessible*, *interoperable* and *reusable* data [6]. In addition, there have been a number of projects, such as VRE4EIC[4] and BlueBRIDGE,[5] to specify or develop generic VREs that can be easily coupled with different RIs and customised for specific communities, taking advantage of improved infrastructure and greater accessibility of data and resources. The lack of conformity of standards and vocabularies in some scientific domains, especially between traditional scientific specialities, makes it difficult however even to retrieve from providers the metadata that describes resources and their appropriate use. Instead, significant software engineering effort is often required on the behalf of data scientists and infrastructure engineers to build specific adapters for every potential coupling of RI and VRE.

Part of the reason it is so important to be able to retrieve resource metadata from different RIs into an integrated environment is to better support interdisciplinary research, which requires the ability to search across different RIs for similar or complementary datasets or services. This entails a complex interaction between a generic VRE and multiple RIs, distributing queries through multiple adapters and then aggregating the results, or else harvesting resource metadata in advance from all providers to allow preliminary queries to be conducted on a single catalogue before distributing follow-on queries to specific providers. Different approaches to managing this interaction balance competing concerns such as liveness, responsiveness, openness and scalability.

In this paper, after providing some foundational background (Section 2), we describe a framework for flexible metadata mapping and publication that can expedite the coupling of an enhanced VRE with resources (principally data, but also models, tools, services, *etc.*.) from different RIs, all using different metadata schemes, to provide cross-RI metadata search and discovery (Section 3). The ability to perform such search and discovery is the basis for any number of other integrative VRE services, including remote service invocation and workflow scheduling and orchestration. We further describe a metadata service that implements this framework, developed in the context of the VRE4EIC project (Section 4). We describe how mappings from standards such as ISO 19139 [7] and DCAT [8] to CERIF [9] using the X3ML framework [10] have been used to automatically ingest metadata published by different RIs to produce a joint catalogue. We describe the Portal that was developed to provide access to this catalogue, discuss its main virtues, and then describe the ongoing developments to further improve the Portal based on feedback received from the environmental and earth science RI community to which it was demonstrated (Section 5). We discuss other developments of relevance to our work and to VRE development in general (Section 6), and finally summarise our contributions (Section 7).

2. Background

Modern research depends on the collection, synthesis and analysis of large volumes of data gathered via sensors, human observations, simulations and experimentation in laboratories and other research settings. These data have to be stored, curated, and made available to those able to make good use of them. Indeed, researchers are now being called upon to collaborate 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 and access to data from a wide range of sources. The role of RIs in this context is therefore to support researchers with data, platforms and tools in order that they can carry out system-level science [11]; no single RI can hope to encompass the full research ecosystem however. Consequently, a researcher or research team engaged in interdisciplinary data science is unlikely to limit their investigations to only one RI, and so will need to gather data from multiple sources, potentially making use of many different tools and services. The challenge set for VREs then is to help researchers freely and effectively interact with the full range of research assets potentially available to them across the many RIs now available, allowing them to collaborate and conduct their research more effectively.

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

[2] https://eudat.eu/.

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

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

[5] http://www.bluebridge-vres.eu/.

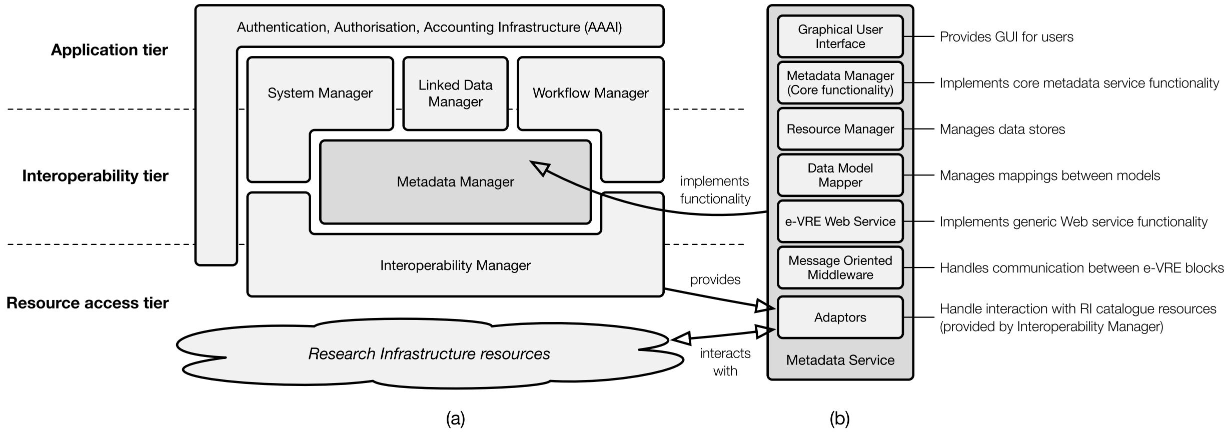

Application tier
Interoperability tier
Resource access tier
Authentication, Authorisation, Accounting Infrastructure (AAAI)
System Manager
Linked Data Manager
Workflow Manager
Metadata Manager
Interoperability Manager
Research Infrastructure resources
implements functionality
provides
interacts with
Graphical User Interface
Metadata Manager (Core functionality)
Resource Manager
Data Model Mapper
e-VRE Web Service
Message Oriented Middleware
Adaptors
Metadata Service
Provides GUI for users
Implements core metadata service functionality
Manages data stores
Manages mappings between models
Implements generic Web service functionality
Handles communication between e-VRE blocks
Handle interaction with RI catalogue resources (provided by Interoperability Manager)
(a)
(b)

Fig. 1. Overview of the e-VRE reference architecture: (a) six modular building blocks for generic VREs able to access the resources of different RIs, distributed across three tiers of operation; and (b) the recommended microservice stack to provide a metadata service implementing the metadata manager building block.

2.1. Metadata standards and technologies

Publishing metadata about resources online (indicating for example the type, coverage, provenance and access method for each resource) allows RIs to advertise their datasets and facilities, and allows researchers to browse and discover data and other resources useful to their research. While there exist standards for various kinds of metadata, such as ISO 19115 [12] and ISO 19139 [7] for describing geospatial metadata (particularly useful for environmental and earth science), the implementation of such standards by RIs can be somewhat idiosyncratic: certain metadata fields might only be loosely specified for example, leading to differing interpretations of how entries into such fields should be structured; other fields might be overloaded, used to carry extra information not part of the original standard, but useful in the particular context in which the standard is applied. These idiosyncrasies mean that often some degree of contextual knowledge is required to correctly interpret the metadata, and familiarity with one data corpus does not necessarily entail perfect comprehension of another corpus by default.

Resource catalogues themselves can be described using standards such as DCAT [8] and harvested via standard protocols such as CSW [13] or OAI-PMH [14]. Some RIs also use Semantic Web [15] technologies such as OWL [16] and SKOS [17] to describe their resources, adapting ontologies such as OBOE [18] (for observations) and vocabularies such as EnvThes [19] (for ecology) to meet their own community's needs, and providing access via SPARQL [20] endpoints. Such endpoints are interacted with very differently from CSW or OAI-PMH based endpoints, requiring alternative query technologies to correctly request and interpret results. Harmonisation of protocols, vocabulary and metadata between RIs thus remains a concern, with communities such as ENVRI[6] (for environmental science) and IVOA[7] (for astronomy) working to promote common models for their respective communities. Many of the interest groups in RDA also pertain to harmonisation of metadata or access, including the work on data type registries [21] and research data collections [22].

VREs benefit from the publication of resource metadata as the primary means to discover and access datasets and other resources provided by RIs. From the VRE perspective, the use of standard protocols, metadata schemes and vocabularies on the part of resource providers is clearly a positive, making it far easier to couple with a greater volume and variety of RIs, to their mutual benefit. VREs themselves however can also be diverse in functionality and operation, and so the use of standard architectural models and terminology is needed to support modular design and improve interoperability internally between VRE components and between VRE components and RI resources; common terminology also makes it easier to discuss methodologies for VRE and RI interaction without getting mired in the specifics of particular technology stacks.

2.2. The e-VRE reference architecture

Jeffery et al. [23] define a reference architecture for enhanced VREs ('e-VREs', as illustrated by Fig. 1) intended to be able to interact with resources provided by many different RIs. According to this architecture, microservices should be used to implement each of six key building blocks distributed across (and often straddling) three tiers of operation: *application*, *interoperability* and *resource access*. Each of the 'building blocks' defined by Jeffery et al. should be constructed via a microservice stack that can be deployed independently; this necessitates a certain degree of redundancy of functionality, but permits new e-VREs to be developed by adapting specific parts of other e-VREs, or indeed to use certain functionalities (such as workflow execution or metadata search and query) in isolation. This is illustrated in Fig. 1 for the stack of microservices needed to be integrated to produce a *metadata service* that implements the *metadata manager* building block. There are seven parts prescribed to such a metadata service by the architecture:

1. A front-end *graphical user interface* (generally a Web client interface, though other interfaces are possible).
2. The *metadata manager* service itself, providing the core functionality of the metadata manager building block distinct from the additional services needed to interact with RIs and other parts of the e-VRE.
3. A *resource manager* for coordinating back-end resources needed to support metadata management.
4. A *data model mapper* for converting ingested data into a common format for indexing and storage.
5. The *e-VRE Web Service*, which is the generic Web service platform upon which the specific functionality of the metadata manager (and other building blocks) is built.
6. *Message-oriented middleware* to communicate with other components in other building blocks.
7. *Adapters* for direct interaction with remote resources not part of the e-VRE.

[6] http://www.envri.eu/.
[7] http://ivoa.net/.

The other building blocks have similar compositions, especially with regard to the e-VRE Web Service and message-oriented middleware components of their respective service stacks. Note that in the e-VRE architecture, the metadata manager lies in the interoperability tier (providing data needed to broker various requests to discover and access resources needed for various applications), but a metadata service *implementing* the metadata manager functionality can stand alone, with its own front-end (which can be considered to belong to the application tier), and its own adapters (for resource access; though these components might be delegated to a dedicated interoperability manager if the service is indeed part of a larger VRE). In Section 4, we provide an example of such a standalone metadata service which is also be part of a larger VRE.

2.3. RI models and terminologies

Similar architectural models exist for RIs, for example the architecture defined by the International Virtual Observatory Alliance (IVOA) for a common Virtual Observatory for accessing astronomy data [24]. With similar goals in mind, Zhao et al. [25] describe the construction of a reference model (ENVRI RM) specifically for environmental science RIs, defining their archetypal elements in the context of the research data lifecycle (encompassing data acquisition, data curation, data publication, data processing and data use). Being based on the Reference Model for Open Distributed Processing (RM-ODP) [26], it models RIs from multiple viewpoints, each with its own concerns (*e.g.* information or computation). Notably, each view defines its own vocabulary and, instantiated for a specific RI, capture concepts that are also of interest at the interface between RIs and VREs. For example, the technology view can capture the software and standards used by services defined in the computational view with which a VRE might interact to discover, access or retrieve resources, while the information view can capture information about the kinds of information object (*e.g.* raw dataset, published dataset with persistent identifier or metadata record) provided by the RI. The most recent release of ENVRI RM is available online [27].

Aside from informing the architecture of research support environments, models such as the e-VRE reference architecture and ENVRI RM can provide controlled terminology for annotating information in databases and knowledge graphs. Open Information Linking for Environmental RIs (OIL-E) [28] provides an OWL specification based on ENVRI RM that acts a machine-actionable vocabulary and upper ontology for RI descriptions.[8] It can be used to contextualise different kinds of RI asset in architectural or operational terms, complementing general-purpose ontologies and terminologies for describing scientific phenomena such as BFO [29], which can also be used to classify RI assets in terms of their roles in scientific investigation. 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 [30]. These kinds of specifications can be used to enrich resource metadata with additional contextual information, or to provide additional relations to bridge linked data graphs [31] served online. More specifically, the terminologies provided by such models provide a way to better classify different kinds of resource as part of a faceted search environment, as we shall discuss in Section 5.

3. Methodology

In this section, we use the e-VRE reference architecture to explore how VREs can be constructed that support heterogeneity of RI resources and resource metadata, and how such a constructed VRE can facilitate cross-RI search and discovery by logically aggregating resource metadata from multiple sources.

3.1. Approaches to metadata retrieval

According to Jeffery et al. [23], VREs operating over multiple RIs can retrieve metadata describing RI resources via one of two means:

1. Via separate interfaces with each RI's own resource catalogues. Each request generated by the VRE is distributed to the relevant RI(s), or simply broadcast to all RIs regardless of relevance.
2. Via a joint resource catalogue that contains metadata about all of the RIs' resources. Metadata from each RI is harvested in advance of user requests, allowing queries to be dispatched to a central database, with only requests to download actual datasets forwarded on to RIs.

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. On the other hand, all information retrieved from the source RIs can usually be assumed to be accurate and up-to-date. Meanwhile, the latter approach simplifies search and discovery, improves query performance, and makes various kinds of data analytic easier to execute, but requires 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 addition, it is necessary to consider how changes to source catalogues are propagated to the joint VRE catalogue.

It may be feasible to meld the two approaches in practice. For example, only a critical juncture of common metadata might be put into the joint VRE catalogue, with the source RI catalogues queried for additional, more specific metadata. Either VRE users would be able to quickly identify which RIs might contain those resources of interest to them, then proceed to connect to those RIs directly, or the query service provided by the VRE would do this on behalf of the user while presenting a more seamless view of query results. Another approach is to have the joint catalogue function as a cache, whereby queries initially are forwarded to the source catalogues, but the results are retained in a central node to expedite future queries that require the same information. If the caching policy is only to retain recently or frequently requested information, a limited VRE catalogue of the 'most valuable' RI resource metadata will naturally emerge, with the most common queries returning results as swiftly as if there was only one central catalogue to search. With these possibilities in mind, we see value in the construction of joint catalogues for use by e-VREs, even in scenarios where only part of the source metadata is extracted.

3.2. Harvesting metadata from multiple RIs

Applying the terminology of the e-VRE reference architecture, Fig. 2 illustrates the main entities involved in the harvesting of resource metadata gathered from multiple RIs. The following steps are involved:

[8] http://www.oil-e.net/.

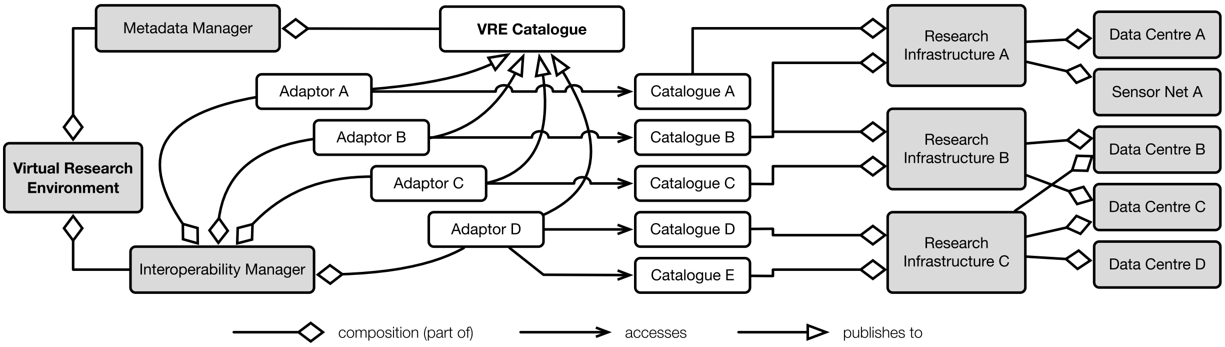


Fig. 2. An e-VRE produces adapters to harvest and convert metadata from different catalogues provided by RIs (often on behalf of multiple data centres or networks), building a common metadata catalogue for its users.

1. The RI must provide a resource catalogue from which to harvest resource metadata. Identification of this catalogue might be performed by a discovery service (assuming some standard publication framework and protocol), or be part of the manual configuration of a customised VRE metadata catalogue (*i.e.* handled by a human expert who knows which catalogue to use and how to access it).
2. The VRE's *interoperability manager* must provide an adapter 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), but also have a model for mapping metadata records retrieved from the source from its native scheme to the scheme used internally by the VRE. Ultimately, the VRE needs a single scheme to fuse the resource metadata from multiple sources into a single coherent joint catalogue.
3. The adapter 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. This process may be a one-off, but could also be repeated periodically to ensure the freshness of the harvested data. Depending on the number of records involved (and the number of data sources), this could be computationally-intensive process.
4. This ingested data is then made available to users of the VRE via its own search and query interface.

By providing the prerequisite adapters, the result is that metadata can now be harvested by the VRE's *metadata manager*. The use of standard APIs on the part of RIs may simplify construction of adapters, but it is unlikely that the blanket use of a single harvesting protocol (*e.g.* OAI-PMH) will be able to capture all details provided by the source RIs without some loss of precision in the resulting data due to differences in how certain common fields are used—for example, a field 'creator' might be assigned the individual who produced the data, the institution that uploaded the data to the local catalogue, or the organisation that published the metadata record. As such, even in cases where a standard protocol or metadata scheme is used, there is often still need to tailor the inter-operation between two separate systems (such as a VRE and a given RI resource catalogue) to account for the particulars of the (meta)data source.

3.3. Metadata aggregation within and between RIs

Although we have thus far referred to a joint catalogue combining catalogued metadata from multiple RIs as a 'VRE catalogue', it is quite possible for joint catalogues to be produced by the RIs themselves, either internally (in the case of federated RIs) or at a cluster level (often at domain level, *e.g.* for the marine or atmospheric research domains), on behalf of different clients or stakeholders. Many RIs contribute data to initiatives such as Copernicus[9] and GEOSS[10] which provide single points of access to certain classes of data. Some RIs also contribute metadata to services such as the EUDAT B2FIND service[11] for dataset discovery. Such catalogues might be used directly by VREs for search and discovery, or treated as another metadata source from which to harvest information for another downstream catalogue.

Internally, most RIs represent federations of existing data centres, many of which already have their own metadata catalogues. The RIs may internally consolidate these catalogues to produce a joint catalogue, or may simply focus on inter-operation between data centres and the specification of new or better standards for common adoption by those centres; regardless, almost all of them are interested in providing a single common data portal to their respective communities. Thus the RIs also face many of the same choices and challenges as faced by VREs.

This raises another question for developers of cross-RI VREs, which is whether they should couple with RIs as integrated entities or should instead directly couple with the data centres *within* RIs. Similarly, VREs could exploit the joint catalogues provided by aggregators such as Copernicus or B2FIND rather than build their own, or build on top of those joint catalogues to do further aggregation.

Clearly, the more degrees of separation between a VRE and the original data resources, the greater the risk of information loss or even error, as well as delays in propagating updates to resource metadata. On the other hand, directly connecting to every individual data centre requires the construction of more interfaces, and greater maintenance effort. Choosing the best approach requires comprehensive understanding of the resource and metadata landscape, but it can be observed that RIs supporting a specific community within a single domain will likely have less heterogeneity in their data, and be better equipped to standardise metadata schemes and their application across the data centres within their sphere of influence; it is therefore likely that any joint catalogue or data portal they produce will be of high quality and retain almost all useful metadata acquired from their respective data sources.

We now examine a system that implements and makes use of a joint catalogue for collecting RI metadata, and consider how well it addresses the needs of researchers.

[9] https://www.copernicus.eu/.

[10] https://www.earthobservations.org/geoss.php.

[11] https://eudat.eu/services/b2find.

4. Implementation

The VRE4EIC Metadata Portal was developed in accordance with the e-VRE reference architecture as part of a Common Reference Prototype[12] that implemented selected building blocks. This was done to demonstrate the viability of the e-VRE approach to the environmental and earth science RI community in particular. The Portal thus implements the necessary components to realise the *metadata manager* functionality of the architecture. All source code is available online,[13] released under the terms and conditions of the Apache 2.0 open source licence.

4.1. VRE4EIC Metadata Portal

The Metadata Portal provides faceted search over catalogue data harvested from multiple RI resources, all aggregated into a single unified catalogue structured according to the CERIF standard for research information systems. Search is therefore based principally on the *context* of research data, directed via associations between datasets, publications, projects, sites, instruments, people, *etc.* that allow related research assets to be retrieved based on exploration of particular facets. Fig. 3 shows an example of such a search, looking specifically for publications produced by a specific individual and related to a specific facility. Similar searches can be made for any basic CERIF entity, relating to datasets, equipment or services for example. This represents a 'typical' search using the Portal, which permits the conjunction (or disjunction) of multiple facets in order to permit more precise queries. Queries constructed using the portal can be saved for later reuse; results can also be exported in various formats.

As well as faceted search based on specific entities, the portal supports geospatial search, which is critical for environmental and earth science applications. Fig. 4 shows an example of search filtering based on setting geographical bounding boxes, which can also be used to filter results in other compound queries such as illustrated by Fig. 3.

The Portal has been designed as a modern service-oriented Web platform, with an implementation based on the combination of Spring Boot[14] at the back-end, and the MVC AngularJS framework[15] for the front-end. It uses the Material Design[16] and Bootstrap[17] user-interface component frameworks to create a clean, modern-looking user interface. Session state and other data management separate from the actual metadata catalogue itself is managed using the H2[18] relational database management system.

The dominant factor in the performance of the Portal is that of the underlying joint catalogue. The CERIF joint metadata catalogue has been implemented in RDF (based on an OWL 2 ontology) hosted within an instance of the open source version of Virtuoso Universal Server[19] behind a RESTful API. Due to the modular architecture of the Portal, any data store that can ingest RDF data and supports SPARQL querying via REST can be used—for example an earlier version of the portal operated on an instance of the Blazegraph triple store.[20] For the current version of the Portal however, Virtuoso was selected due to its scalability, cross-platform flexibility, and the fact that it is capable of combining relational, graph, and document data management with Web application server and Web services platform functionality. Virtuoso has also fared well in prior comparative performance analyses for similar data corpora, for example for biomedical graph data [32] and geospatial Smart City graph data [33].

The whole platform can be packaged in a single Java archive, executed from the command line as a standalone Maven application; this is achieved by embedding a server container (Eclipse Jetty[21] by default).

4.2. CERIF joint catalogue

Metadata harvested from external RI sources are converted into CERIF RDF using the X3ML mapping framework [10], a system for mapping XML-based documents that use a given source scheme into RDF documents that can then be ingested into any graph-based data store that can read RDF (*e.g.* Virtuoso Universal Server). The mapping process itself is as illustrated in Fig. 5, with the major stages as follows:

1. Sample metadata records, along with their corresponding metadata schemes, are retrieved for analysis from RI resource catalogues.
2. Mappings are defined in X3ML that dictate the transformation of records structured according to the selected XML-based schemes into CERIF-compliant RDF documents.
3. Metadata records are then harvested in quantity from RI resource catalogues (typically served by systems such as GeoNetwork[22] or CKAN[23]) in their native format, *e.g.* as ISO 19139 XML or DCAT-AP data.
4. The X3ML mappings are used to transform the harvested metadata records into CERIF RDF format.
5. The transformed RDF data are ingested into the unified CERIF metadata catalogue.

Once ingested, these data then become available to users of the metadata portal, who can query and browse data upon authentication by the front-end authentication service (implementing the *AAAI* component of the e-VRE reference architecture, as described below). As the underlying data model for the unified catalogue is RDF-based, queries constructed using the Portal are submitted to the underlying database using SPARQL 1.1. It is possible for users to directly construct SPARQL queries via the Portal, or to edit as SPARQL queries constructed via the graphical Web interface; it is a principle of the Portal's design however that most users should never need to.

X3ML mappings are described using the 3M Mapping Memory Manager.[24] 3M is a Web application, that can be run in a servlet container environment such as Apache Tomcat,[25] which allows for mappings to be viewed, shared and edited as part of a community via any standard Web browser. 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 Fig. 6. Besides the actual specification of mapping rules, 3M supports the specification of generators to produce logical identifiers for new concepts constructed during translation of terms, and it provides test and analytic facilities to determine the functionality and coverage of mappings. 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 [34].

[12] http://v4e-lab.isti.cnr.it/.

[13] https://github.com/vre4eic.

[14] https://spring.io/projects/spring-boot.

[15] https://angularjs.org/.

[16] https://material.io/.

[17] https://getbootstrap.com/.

[18] https://www.h2database.com/.

[19] https://virtuoso.openlinksw.com/.

[20] https://www.blazegraph.com/.

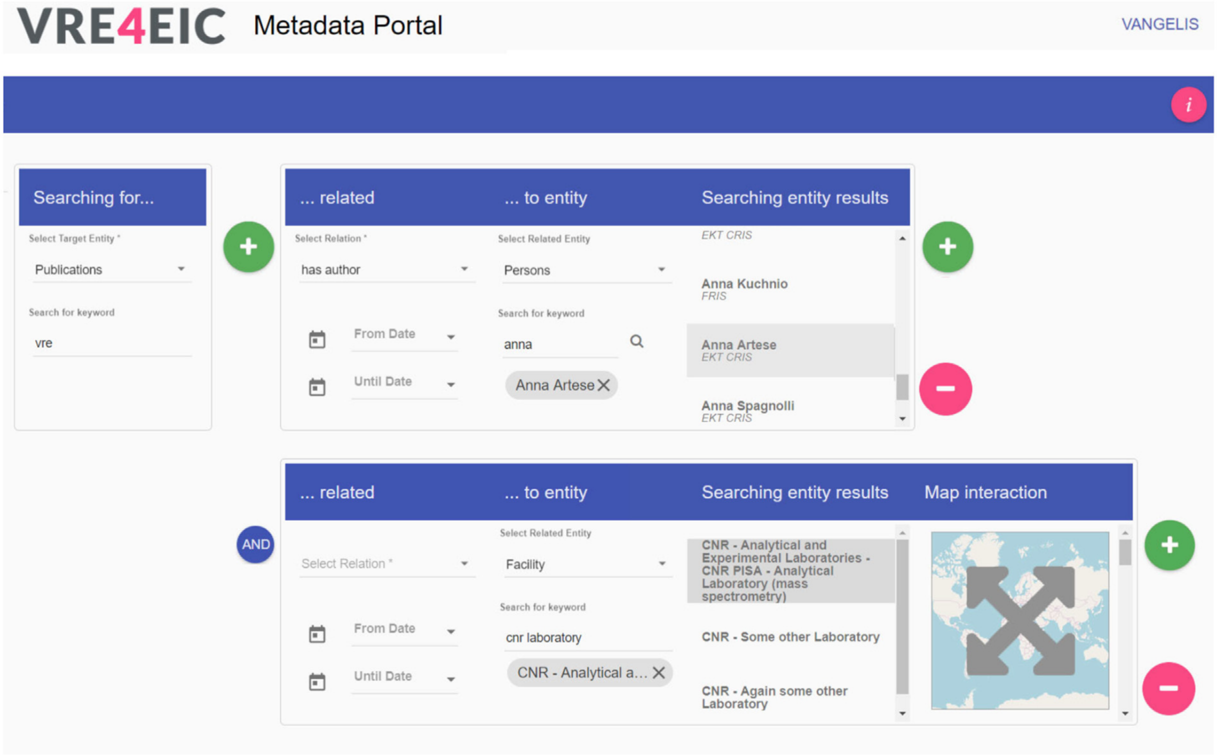
[21] https://www.eclipse.org/jetty/.

[22] https://geonetwork-opensource.org/.

[23] https://ckan.org/.

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

[25] http://tomcat.apache.org/.



Select Target Entity *: Publications

Search for keyword: vre

Select Relation *: has author

Select Related Entity: Persons

From Date

Until Date

Search for keyword: anna

Anna Artese ✕

EKT CRIS

Anna Kuchnio
FRIS

Anna Artese
EKT CRIS

Anna Spagnolli
EKT CRIS

Select Relation *

Select Related Entity: Facility

From Date

Until Date

Search for keyword: cnr laboratory

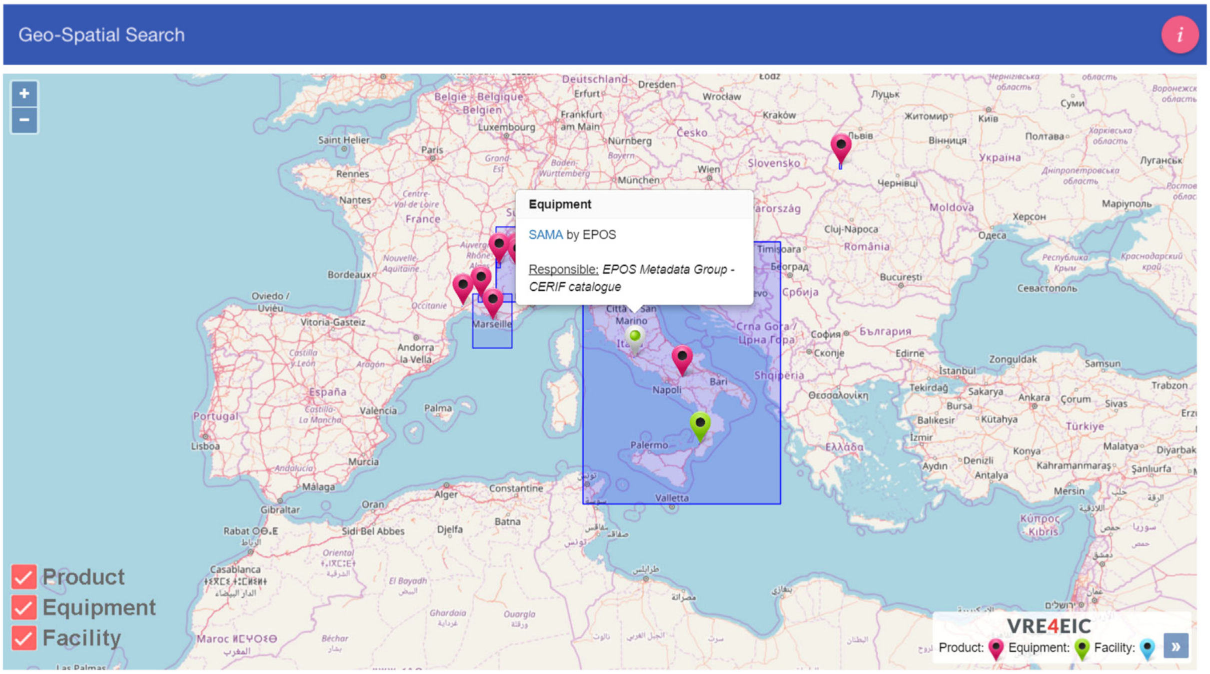
CNR - Analytical a... ✕

CNR - Analytical and Experimental Laboratories - CNR PISA - Analytical Laboratory (mass spectrometry)

CNR - Some other Laboratory

CNR - Again some other Laboratory

Fig. 3. The VRE4EIC Metadata Portal: searching for data publications authored by Anna Artese relating to CNR Pisa's mass spectrometry analytical laboratory.



Equipment

SAMA by EPOS

Responsible: EPOS Metadata Group - CERIF catalogue

☑ Product
☑ Equipment
☑ Facility

Product: ● Equipment: ● Facility: ●

Fig. 4. The VRE4EIC metadata portal: identifying equipment and facilities provided by the EPOS RI in Italy.

4.3. Identification and security

The VRE4EIC Metadata Portal has been made available to developers as part of the VRE4EIC Common Reference Prototype (CRP). As of writing, the CRP implements three key services: the Metadata Portal (implementing *metadata manager* functionality, a Node Service (implementing *AAAI* functionality), and a simple workflow service for invoking online Web Services (implementing *workflow manager* functionality). Of interest here is the Node Service, which implements all functionalities related to user profile management and e-VRE system administration. Apache Zookeeper[26] acts as a start-up broker for secure communication using SSL, and is launched as an embedded server

[26] http://zookeeper.apache.org/.

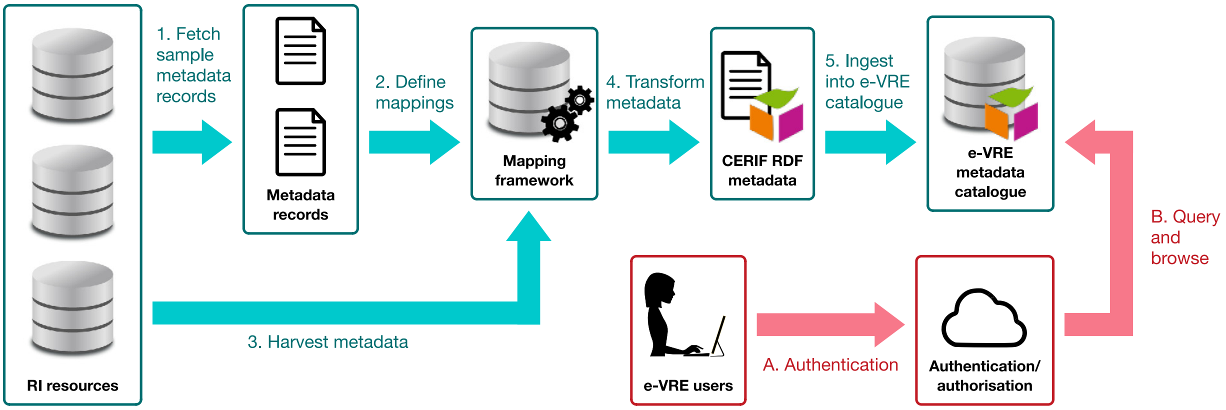
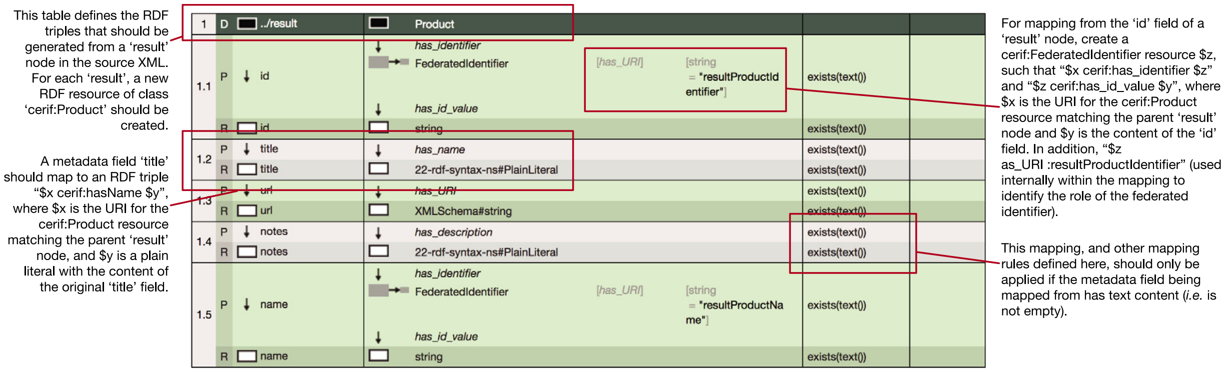


Fig. 5. e-VRE metadata acquisition and retrieval workflow: metadata records are acquired from multiple sources, mapped to CERIF RDF and stored in the VRE catalogue; authenticated VRE users query data via the e-VRE.



This table defines the RDF triples that should be generated from a ‘result’ node in the source XML. For each ‘result’, a new RDF resource of class ‘cerif:Product’ should be created.

A metadata field ‘title’ should map to an RDF triple “$x cerif:hasName $y”, where $x is the URI for the cerif:Product resource matching the parent ‘result’ node, and $y is a plain literal with the content of the original ‘title’ field.

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())

For mapping from the ‘id’ field of a ‘result’ node, create a cerif:FederatedIdentifier resource $z, such that “$x cerif:has_identifier $z” and “$z cerif:has_id_value $y”, where $x is the URI for the cerif:Product resource matching the parent ‘result’ node and $y is the content of the ‘id’ field. In addition, “$z as_URI :resultProductIdentifier” (used internally within the mapping to identify the role of the federated identifier).

This mapping, and other mapping rules defined here, should only be applied if the metadata field being mapped from has text content (*i.e.* is not empty).

Fig. 6. Example of mapping rules generated in 3M: *result* metadata taken from a CKAN repository is mapped to a CERIF *product* with data and object properties corresponding to each possible attribute in the source XML scheme.

by the Node Service. It provides user authentication for the VRE and connected e-RIs, authorisation and accounting services, and data encryption layers for components that are accessible over potentially insecure networks. For users of the Metadata Portal, multi-factor authentication is provided for granting access to users. This mechanism requires users to present two pieces of evidence in order to log in; their regular credentials and a code sent to a Telegram account.[27] User access to the joint VRE catalogue can be further regulated using role-based access control.

4.4. Comparisons with other VREs and portals

Frameworks dealing with the construction of VREs for different communities often focus on the provisioning of e-infrastructure for data processing (*e.g.* Globus Galaxies [35] or the CIPRES workbench [36]). The VRE4EIC Metadata Portal is focused on the discovery and retrieval of scientific data from RIs, but is conceived as a constituent element of a larger VRE, which may include other components such as a workflow manager for scheduling processes. These components might themselves be built using existing technologies already used by other VREs, such as Galaxy [37] or Taverna [38]. Consequently, the Portal is not in direct competition with these other frameworks, but rather represents another component for constructing generic VREs tailored to specific communities.

As a catalogue front-end, the Portal can be compared to community aggregators such as EUDAT's B2FIND service, which uses CKAN for content management. Our portal uses the CERIF standard to structure its data, which provides a flexibility not found in more rigid schemes such as used for B2FIND, but this flexibility comes at the cost of additional complexity in the underlying data model. This is offset however by the use of a robust metadata mapping pipeline and by a simple-to-use user interface for constructing queries. We can also compare with RI-specific data portals such as the ICOS Carbon Portal[28] for greenhouse gas data. The ICOS portal uses Semantic Web technologies just as our portal, but only serves data for a specific RI, and uses a more specific set of facets for locating specific datasets. Properly configured, the VRE4EIC Metadata Portal can encompass all these facets within its own metadata model and act as intermediary, directing queries to the RI-specific portal as needed.

Sister projects to VRE4EIC such as BlueBRIDGE make use of the D4Science platform for VREs [39]. D4Science provides a ready-made host environment for community VREs which is suitable for wide range of use-cases, but also encloses computation and data within a single environment which, though accessible from the outside, is contrary to the open approach we have taken in which existing services distributed across RIs and e-infrastructures are loosely coupled together through standard protocols and APIs. Both approaches have merit however, and depend on the needs of different research communities.

5. Further development

The VRE4EIC Metadata Portal was demonstrated to the ENVRI community cluster of environmental science RIs in Europe as

[27] https://telegram.org/.

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

well as directly to the European Plate Observing System (EPOS),[29] with sample data harvested from a cross-section of RIs across the ecosystem and solid earth science domains, further augmented by synthetic data for a total data corpus of approximately 53 million RDF triples as of October 2018.

Feedback was broadly positive, but indicated a number of specific improvements that would make this kind of metadata portal more useful to the RI community. We now examine these key improvement areas, and discuss any ongoing developments which address them.

5.1. Better handling of under-defined metadata

CERIF was originally designed based on a relational database model and so consequently defines a number of strictly disjoint entity classes without any kind of default hierarchy as would typically be found (for instance) in an ontology. This grants CERIF a certain flexibility (since it allows arbitrary relations to be defined between any two entities), but can cause difficulties where metadata elements are under-defined. As an example, CERIF strictly distinguishes between *people*, *organisation units* and *facilities*, but does not formally define a more generic 'agent' concept. In a number of cases however, we found that source metadata records would define certain agents (such as 'publisher', 'creator' or 'owner') that could be people, organisations or institutes without any definitive way to distinguish between them. Without the means to map to a precise entity class, these concepts would be lost upon translation into the strict CERIF standard, thus raising the question of how to manage 'graceful degradation' where important but semantically-ambiguous entities are to be found in harvested metadata. One possibility is to make use of CERIF's support for probabilistic relations between entities, or to make use of additional information sources to disambiguate entities. The use of additional external information sources, such as registries of entities (people, organisations, institutions, *etc.*) used in metadata records from a given source (*e.g.* the RIDE database [40] for EPOS) is being investigated to help with disambiguation of under-defined metadata. Text-based analysis could also help (given that most ambiguous fields typically use free text), but essentially trades away precision for greater recall, which may not be acceptable for catalogues made available to scientists as 'production-ready'.

5.2. Greater exploitation of common terminology

A notable feature of CERIF is how it separates its semantic layer from its primary entity-relationship model. Most CERIF relations between two entities are semantically agnostic, lacking any particular interpretation beyond identifying a link. Almost every entity and relation can be assigned though a classification drawn from a classification scheme 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). Though different vocabularies (*e.g.* ISO 19115 codelists) were investigated in the VRE4EIC project as possible classification schemes for CERIF entities and relations in the context of environmental science, a harmonised set of schemes would be needed for any particular instance of the portal to fully exploit CERIF's semantic layer.

The vocabulary provided by OIL-E has been identified within VRE4EIC as a means to further classify entities and relations between entities 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.

An overview of OIL-E concepts that can be used to classify various CERIF entities and relations was published by the VRE4EIC project [41]; Table 1 shows some examples of such classifications. Classifying CERIF entity classes such as *person*, *facility*, *result product* 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 product* 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.

5.3. Integration of semantic search facilities

The identification of synonymous, subsuming and intersecting terms 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. Currently, the VRE4EIC portal principally supports faceted search based explicitly on entity classes with keyword filters, but making use of linked vocabularies would allow for more 'free text' searches alongside structured search, and would also 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.

Regarding linked vocabulary for semantic search, RIs such as AnaEE[30] and LTER-Europe[31] are actively developing better vocabularies for describing ecosystem and biodiversity research data, building upon existing SKOS vocabularies. Both the AnaEE data vocabulary [42] and LTER's environmental thesaurus EnvThes [19] have mappings to other established domain vocabularies such as Agrovoc[32] and GEMET.[33] These RIs are now collaborating with other RIs in the ENVRI community to harmonise their vocabularies in order to provide semantic linking between terms used in their respective sub-domains; this work will be performed in the context of the ENVRI-FAIR project,[34] which focuses on implementing FAIR principles in data and services across the environmental sciences.

6. Discussion

Any sustainable VRE cataloguing solution will need to address certain challenges, including how to integrate new RI resources, handle updates to standards, scale with ever-greater data volumes, and ensure proper attribution of credit for data made available to researchers. All of these challenges require both technical and governance solutions broadly supported by research communities, requiring continued collaboration between various interest groups. In this section we make observations on topics that relate to VRE development in general, and indicate where our own contributions intersect with them.

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

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

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

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

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

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

Table 1
Examples of OIL-E classifications of CERIF entities: the OIL-E concept that acts as the classification scheme is identified along with examples of sub-concepts that act as classification instances. For readability, the concepts' RDFS labels rather than URIs are used.

CERIF entity	OIL-E base class	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 Product'	'persistent data'	'QA-assessed data', 'annotated data'
Service	'computational object'	'catalogue service', 'data broker'

6.1. Integrating new resources

More data, more data resources and more research infrastructure all place additional pressure on 'seamless' integrative environments. Standardisation in protocols, schemes and vocabularies remain the best mechanisms for dealing with greater heterogeneity in distributed data volumes, but there is always need for data mapping, especially across disciplinary boundaries. As joint catalogues are mainly concerned with *meta*data rather than the actual datasets themselves, aggregated catalogue data do not tend to fit the profile of 'big data' in terms of raw volume, but the act of synthesis and integration itself is still challenging. The use of frameworks such as X3ML and flexible target schemes such as CERIF or GeoDCAT-AP [43], can make this integration viable, while the use of a standard metadata mapping framework with tool support (*e.g.* X3ML with the 3M editor) allows for a fairly rapid adaptation of mappings between a schemes in response to changes at source or destination. Automation using machine learning can help to accelerate the construction of new mappings, but rarely without issue. Ochieng and Kyanda [44] survey automated ontology matching tools and highlight the role of interactive matching tools [45], whereby experts repair weaknesses in matches generated by automated matchers, noting diminishing returns on improvements to both precision and recall in recent years in unsupervised approaches.

6.2. Maintaining the catalogue

It will be necessary to periodically refresh the content of any joint catalogue as datasets are cleaned, extended and updated at their source. Datasets hosted by different RIs have varying update regimes, meaning that a single policy (*e.g.* update every 24 h) is not practical. In practice, updates can be pulled by the VRE (via periodic polling of RI resources) or pushed by the RI (by broadcasting updates to subscribers). The latter approach is desirable, but requires RI communities to support some kind of subscription mechanism for VREs. The Euro-Argo RI[35] has been developing a data subscription service for researchers [46]; a VRE subscription service running on similar principles may be feasible.

Whether a push or pull model is used to acquire metadata updates, a joint catalogue should maintain a history of changes to metadata, as an aid both to search and to general reproducibility. Such data provenance can be structured according to established standards such as PROV [47]), which can be integrated or linked to entities in the joint catalogue—CERIF, for example, is able to represent time-bounded semantic relationships that can provide historical context. One issue is that metadata currently provided by RIs still often lacks this kind of provenance information; the adoption of standardised 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 joint catalogues.

6.3. Linking with semantic web

Semantic Web technologies represent one approach to resource metadata publication. The use of such technologies is plagued by the recurrent problems of knowledge representation in general such as computability, inconsistency and incompleteness, but with added further problems of data redundancy and unreliability. Considerable attention has thus been given to the openness, extensibility and computability of Semantic Web standards, weighing different options (*e.g.* the use of SKOS over OWL [48,49] for terminology specifications). Nevertheless, the use of linked data [31] for describing resources (of all kinds) is well-established, with research now focusing on different approaches for generating linked data from various sources both static and dynamic, as well as with how to navigate and query distributed information once made available. Recent examples of such research include the generation of a navigable Graph of Things from live Internet of Things data sources [50] and the use of crowdsourcing to provide real-time transport data in rural areas [51], both topics with parallels 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 [52] and LILAC [53], which can make linked data based search over distributed catalogues more practical than is currently the case by better distributing queries across catalogue nodes with less redundancy and then joining the results efficiently. Such developments reduce the need to aggregate as much metadata in a joint catalogue, however the demands of search (particularly with regard to perceived responsiveness to queries by end-users) make it still generally necessary to cache key metadata in a central store.

In the geospatial domain prominently occupied by current environmental science RIs, most standards have been developed independently of the Semantic Web, with recommendations such as INSPIRE[36] being all but disjoint from it, though technologies such as GeoSPARQL[37] do attempt to address this by bridging the capability gap. This current separation poses a barrier for integration of geospatial catalogues published via CSW or OAI-PMH into the Semantic Web, and adapters are still needed to query such data sources and present responses in RDF format (*e.g.* [54]); generic standards such as R2RML [55] make it easier now however to construct RDF-based views on relational data in databases, which helps create more seamless interoperability between the Semantic Web and other data frameworks.

6.4. Data objects and workflows

Most scientific investigations follow a workflow, and there have been a number of workflow management systems developed with different characteristics and target applications [56], several of which have been applied to science [57] and to VRE

[35] https://www.euro-argo.eu/.

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

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

applications [58]. The use of ontologies for verification and validation of workflows has already been explored (*e.g.* [59]), 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, but to any resource that researchers may want to discover and access.

Integration between workflow systems and provenance recording is essential to the reproducibility of research results. To this end, VREs must be able contribute as well as use provenance data in their integrated workflow systems (*e.g.* the Kepler workflow management system [60]).

Capturing all the various relationships between different entities involved in research requires some consideration of how to package these relationships into a single unit. The joint catalogue used by the VRE4EIC Metadata Portal can capture many of the necessary facets of research objects [61] via the use of the CERIF standard, but more work could be done to support data object collections, taking in the recommendations of the Research Data Collections working group of the RDA [22]. Many of the entities provided by RIs are collections, or are part of collections, but there is still general uncertainty in the data science community as to how best to serve such collections to researchers and to support the internal search and discovery of collection content.

Related to this, there remains a broader question in data science regarding the accessibility of 'dark data' [62]—datasets produced by individual researchers and small research teams not represented by any of the large RIs and therefore not discoverable via their catalogues, but perhaps only by smaller institutional repositories. In this paper, we considered only the coupling of VREs with the curated assets of formally amalgamated RIs with mature data management systems and policies, but it may also be worth giving more consideration as to how data produced in the long tail of science can be made visible via a VRE; for example, making use of the catalogues of open repositories such as Zenodo,[38] or increasing the visibility of small institutional repositories.

6.5. Governance

The construction, deployment and maintenance of joint catalogues for use by VREs is as much a governance problem as a technical one. The use of standard protocols and terminology that make the production of joint catalogues much easier requires consensus across the research community, and this requires effective forums in which stakeholders can hold dialogues and agree best practices—hence community initiatives such as ENVRI for the environmental and earth science RI community, and RDA for data science in general, which then can make recommendations.

These recommendations influence the work of VRE developers. There is strong correlation between the metadata element set recommended by the RDA Metadata interest group and the primary entities defined by CERIF; likewise, the e-VRE reference architecture was conceived to directly address the need for standard VRE architectures espoused by the RDA VRE interest group. The use of standard terminology for classifying entities is also heavily influenced by communities such as ENVRI. This influence is natural, and demonstrates the importance of such bodies to this kind of work.

In general, governance of evolving standards for data science belong to those initiatives that have the support of the scientific community. For VRE developers, this means that close collaboration is essential, as is the agility to adapt to the continued evolution of the RI resource landscape. This adaptability is key, because while the fundamental requirements of data science as embodied (for example) by the FAIR principles are becoming increasingly well-understood, the specific technologies and methods used to address these principles will continue to change; hence there also need to be equivalent principles for generic, flexible VRE design, at the core of which is how we gather resource metadata into searchable, expressive catalogues as the basis for nearly all key VRE services.

7. Conclusion

In this paper we linked the development of VREs (alternatively referred to as science gateways or virtual laboratories) to the outgrowth of dedicated RIs providing curated data services to research communities, and argued the need for new VREs that can be freely coupled with different RI resources based on the evolving requirements of researchers and of data science. In particular, we argued that to provide researchers seamless discovery and access of RI resources, it is necessary to build VREs that can interface with as wide a range of resources as possible, brokered by catalogue services either provided by the RIs themselves or created at the VRE side by harvesting metadata from RI resource catalogues.

In order to realise such a network of linked infrastructure and catalogue services however, we asserted that some degree of metadata mapping is essential to facilitate cross-RI search and discovery, mostly due to the fundamental diversity of metadata schemes, vocabularies and protocols used to access resource catalogue data published by different RIs, but also due to idiosyncracies in how such schemes, vocabularies and protocols are used in practice. We examined how metadata might be aggregated into a single logical catalogue, comparing the approach of actually harvesting metadata to make a single physical catalogue versus simply brokering requests redirected towards multiple separate catalogues. We also looked at how such aggregation is currently being performed within RIs and RI clusters. Focusing on scenario of the creation of a joint catalogue drawing metadata from heterogeneous sources, we outlined a methodology for building such a catalogue based on the e-VRE reference architecture. We also examined the steps required for building a robust mapping pipeline for handling heterogeneous metadata, which is essential for building such catalogues.

We provided an example in the VRE4EIC Metadata Portal of how our methodology is applied in practice; in this instance, the VRE4EIC project took the approach of building a single centralised catalogue using CERIF, a European research information standard, as a framework for aggregating resource metadata from different metadata catalogues provided by members of the ENVRI cluster of environmental and earth science RIs. We described the use of the X3ML framework to produce effective mappings from XML-based metadata records to RDF data suitable for building a unified knowledge graph. We used 3M, an X3ML editor and transformation platform, to translate ISO 19139 XML, CKAN, Dublin Core, DCAT-AP and OIL-E data into CERIF RDF for ingestion into a CERIF RDF knowledge graph hosted within a Virtuoso data store. Based on the feedback given by RIs involved in VRE4EIC and in the ENVRI community, we identified key areas where more work was necessary, and described the ongoing development that is being done or is planned to address these areas for future iterations of the Portal and other e-VRE services. Finally, we discussed more broadly some of the issues that may bear impact on VRE and VRE catalogue development in general, such as the refreshing of metadata records, the coupling of VREs with other types of service provided by RIs, and the need to closely follow the activities and recommendations of community initiatives for establishing standards for data science in specific domains and in general.

[38] https://zenodo.org/.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreements 654182 (ENVRIplus), 676247 (VRE4EIC), 676564 (EPOS IP), and 824068 (ENVRI-FAIR).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Candela, D. Castelli, P. Pagano, Virtual research environments: an overview and a research agenda, Data Sci. J. 12 (2013) 75–81.

[2] Z. Zhao, P. Martin, C. de Laat, K. Jeffery, A. Jones, I. Taylor, A. Hardisty, M. Atkinson, A. Zuiderwijk, Y. Yin, 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 2016), 2016.

[3] E. Deelman, D. Gannon, M. Shields, I. Taylor, Workflows and e-Science: An overview of workflow system features and capabilities, Future Gener. Comput. Syst. 25 (5) (2009) 528–540.

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

[5] The European Commission, Realising the European Open Science Cloud, The European Commission, 2016, URL https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf.

[6] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, et al., The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3 (2016).

[7] ISO 19139:2007, Geographic Information—Metadata—XML Schema Implementation, ISO/TS Standard, 2007.

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

[9] B. Jörg, CERIF: The common european research information format model, Data Sci. J. 9 (2010) 24–31.

[10] Y. Marketakis, N. Minadakis, H. Kondylakis, K. Konsolaki, G. Samaritakis, M. Theodoridou, G. Flouris, M. Doerr, X3ML mapping framework for information integration in cultural heritage and beyond, Int. J. Digit. Libr. (2016) 1–19.

[11] I. Foster, C. Kesselman, Scaling system-level science: Scientific exploration and it implications, Computer 39 (11) (2006) 31–39.

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

[13] D. Nebert, U. Voges, L. Bigagli, OGC Catalogue Services 3.0—General Model, OGC implementation standard, Open Geospatial Consortium, 2016, URL http://docs.opengeospatial.org/is/12-168r6/12-168r6.html.

[14] C. Lagoze, H. Van de Sompel, The making of the open archives initiative protocol for metadata harvesting, Libr. Hi Tech 21 (2) (2003) 118–128.

[15] T. Berners-Lee, J. Hendler, O. Lassila, et al., The semantic web, Sci. Am. 284 (5) (2001) 28–37.

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

[17] S. Bechhofer, A. Miles, SKOS simple knowledge organization system reference, W3C recommendation, W3C, 2009, URL http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

[18] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, F. Villa, An ontology for describing and synthesizing ecological observation data, Ecol. Inf. 2 (3) (2007) 279–296.

[19] H. Schentz, J. Peterseil, N. Bertrand, EnvThes—interlinked thesaurus for long term ecological research, monitoring, and experiments, in: EnviroInfo, 2013, pp. 824–832.

[20] W3C SPARQL Working Group, SPARQL 1.1 overview, W3C recommendation, W3C, 2013, URL http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

[21] L. Lannom, D. Broeder, G. Manepalli, Data Type Registries Working Group Output, RDA, 2015, http://dx.doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458.

[22] T. Weigel, B. Almas, F. Baumgardt, T. Zastrov, U. Schwardmann, M. Hellström, J. Quinteros, D. Fleischer, Recommendation on Research Data Collections, RDA, 2017, http://dx.doi.org/10.15497/RDA00022.

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

[24] C. Arviset, S. Gaudet, et al., IVOA architecture, 2011, arXiv preprint arXiv:1106.0291.

[25] Z. Zhao, P. Martin, P. Grosso, W. Los, C.d. Laat, K. Jeffrey, A. Hardisty, A. Vermeulen, D. Castelli, Y. Legre, W. Kutch, Reference model guided system design and implementation for interoperable environmental research infrastructures, in: 2015 IEEE 11th International Conference on E-Science (E-Science), IEEE, 2015, pp. 551–556.

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

[27] A. Nieva de la Hidalga, B. Magagna, M. Stocker, A. Hardisty, P. Martin, Z. Zhao, M. Atkinson, K. Jeffery, The ENVRI Reference Model (ENVRI RM) version 2.2, 2017, URL https://doi.org/10.5281/zenodo.1050349.

[28] P. Martin, P. Grosso, B. Magagna, H. Schentz, Y. Chen, A. Hardisty, W. Los, K. Jeffery, C. de Laat, 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.

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

[30] D. Bailo, D. Ulbricht, M.L. Nayembil, L. Trani, A. Spinuso, K.G. Jeffery, Mapping solid earth data and research infrastructures to CERIF, Procedia Comput. Sci. 106 (2017) 112–121.

[31] T. Berners-Lee, Linked data, w3c design issues, 2006, URL https://www.w3.org/DesignIssues/LinkedData.html (Accessed 26 February 2018).

[32] H. Wu, T. Fujiwara, Y. Yamamoto, J. Bolleman, A. Yamaguchi, BioBenchmark Toyama 2012: an evaluation of the performance of triple stores on biological data, J. Biomed. Semant. 5 (1) (2014) 32.

[33] P. Bellini, P. Nesi, Performance assessment of RDF graph databases for smart city services, J. Vis. Lang. Comput. 45 (2018) 24–38.

[34] M. Theodoridou, D. Ivanovic, P. Martin, L. Remy, M. Muckensturm, X3ML mappings from common metadata schemes to CERIF RDF, 2019, URL https://doi.org/10.5281/zenodo.2548732.

[35] R. Madduri, K. Chard, R. Chard, L. Lacinski, A. Rodriguez, D. Sulakhe, D. Kelly, U. Dave, I. Foster, The Globus Galaxies platform: delivering science gateways as a service, Concurr. Comput.: Pract. Exper. 27 (16) (2015) 4344–4360.

[36] M.A. Miller, T. Schwartz, P. Hoover, K. Yoshimoto, S. Sivagnanam, A. Majumdar, The CIPRES workbench: a flexible framework for creating science gateways, in: Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled By Enhanced Cyberinfrastructure, ACM, 2015, p. 39.

[37] J. Goecks, A. Nekrutenko, J. Taylor, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol. 11 (8) (2010) R86.

[38] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, et al., The Taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud, Nucleic acids Res. 41 (W1) (2013) W557–W561.

[39] L. Candela, P. Pagano, D. Castelli, A. Manzi, Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience, in: International Symposium on Grids and Clouds, ISGC, vol. 1, 2014, p. 22.

[40] D. Bailo, A. Bartoloni, K.G. Jeffery, A. Clemenceau, T.L. Hoffmann, RIDE: the Research Infrastructure Database for EPOS, in: EGU General Assembly Conference Abstracts, vol. 15, 2013.

[41] L. Remy, D. Ivanovic, J. van Ossenbruggen, T. Patkos, A. Kritsotaki, M. Sbarra, P. Martin, D4.2 matching and mapping VRE elements to CERIF, H2020 VRE4EIC Project, 2017, http://www.vre4eic.eu.

[42] Anaee-France semantic group, Anaee thesaurus, 2016, URL http://dx.doi.org/10.15454/1.4894016754286177E12.

[43] A. Perego, A. Friis-Christensen, M. Lutz, Geodcat-AP: Use cases and open issues, smart descriptions & smarter vocabularies (SDSVoc), Amsterdam, Netherlands, 2016.

[44] P. Ochieng, S. Kyanda, Large-scale ontology matching: State-of-the-art analysis, ACM Comput. Surv. 51 (4) (2018) 75.

[45] Z. Dragisic, V. Ivanova, P. Lambrix, D. Faria, E. Jiménez-Ruiz, C. Pesquita, User validation in ontology alignment, in: International Semantic Web Conference, Springer, 2016, pp. 200–217.

[46] S. Koulouzis, P. Martin, H. Zhou, Y. Hu, J. Wang, T. Carval, B. Grenier, J. Heikkinen, C. de Laat, Z. Zhao, Time-critical data management in clouds: Challenges and a dynamic real-time infrastructure planner (DRIP) solution, Concurr. Comput.: Pract. Exper. (2019) e5269.

[47] P. Groth, L. Moreau, PROV-overview, W3C note, W3C, 2013, URL http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

[48] A. Stellato, Dictionary, thesaurus or ontology? Disentangling our choices in the semantic web jungle, J. Integr. Agricult. 11 (5) (2012) 710–719.
[49] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, E. Summers, Key choices in the design of simple knowledge organization system (SKOS), Web Semant. Sci. Serv. Agents World Wide Web 20 (2013) 35–49.
[50] D. Le-Phuoc, H.N.M. Quoc, H.N. Quoc, T.T. Nhat, M. Hauswirth, The graph of things: A step towards the live knowledge graph of connected things, Web Semant. Sci. Serv. Agents World Wide Web 37 (2016) 25–35.
[51] D. Corsar, P. Edwards, J. Nelson, C. Baillie, K. Papangelis, N. Velaga, Linking open data and the crowd for real-time passenger information, Web Semant. Sci. Serv. Agents World Wide Web 43 (2017) 18–24.
[52] O. Hartig, J. Pérez, LDQL: A query language for the web of linked data, Web Semant. Sci. Serv. Agents World Wide Web 41 (2016) 9–29.
[53] G. Montoya, H. Skaf-Molli, P. Molli, M.-E. Vidal, Decomposing federated queries in presence of replicated fragments, Web Semant. Sci. Serv. Agents World Wide Web 42 (2017) 1–18.
[54] K. Patroumpas, N. Georgomanolis, T. Stratiotis, M. Alexakis, S. Athanasiou, Exposing INSPIRE on the semantic web, Web Semant. Sci. Serv. Agents World Wide Web 35 (2015) 53–62.
[55] S. Sundara, S. Das, R. Cyganiak, R2RML: RDB to RDF mapping language, W3C recommendation, W3C, 2012, URL http://www.w3.org/TR/2012/REC-r2rml-20120927/.
[56] C.S. Liew, M.P. Atkinson, M. Galea, T.F. Ang, P. Martin, J.I.V. Hemert, Scientific workflows: Moving across paradigms, ACM Comput. Surv. 49 (4) (2016) 66:1–66:39, http://dx.doi.org/10.1145/3012429, URL http://doi.acm.org/10.1145/3012429.
[57] R. Mork, P. Martin, 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.
[58] Z. Zhao, A. Belloum, C. de Laat, P. Adriaans, B. Hertzberger, Distributed execution of aggregated multi domain workflows using an agent framework, in: Services, 2007 IEEE Congress on, IEEE, 2007, pp. 183–190.
[59] T. Miksa, A. Rauber, Using ontologies for verification and validation of workflow-based experiments, Web Semant. Sci. Serv. Agents World Wide Web 43 (2017) 25–45.
[60] I. Altintas, O. Barney, E. Jaeger-Frank, Provenance collection support in the Kepler scientific workflow system, Proven. Annot. Data (2006) 118–132.
[61] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, I. Buchan, Research objects: Towards exchange and reuse of digital knowledge, in: The Future of the Web for Collaborative Science (FWCS 2010), 2010.
[62] P.B. Heidorn, Shedding light on the dark data in the long tail of science, Libr. Trends 57 (2) (2008) 280–299.

Dr. Paul Martin obtained his Ph.D. in Informatics in 2011 from the University of Edinburgh with an interest in semantic modelling, distributed artificial intelligence and argumentation. After working within the Data Intensive Research group at the University of Edinburgh for four years, he joined the System and Network Engineering group at the University of Amsterdam in 2015. Since 2011 he has worked in a number of software and infrastructure-related EU projects: the FP7 projects ADMIRE, ENVRI and VERCE, and the H2020 projects SWITCH, ENVRIplus and VRE4EIC.

Laurent Remy is a senior analyst-developer with a large amount of experience in IT development and training, and has been working with CERIF-based systems for many years. He has experience on IT systems integration and has been involved in several EC funded projects for euroCRIS. His specialities include: Data Analysis, Research Information Systems, Systems Integration and Programming.

Maria Theodoridou is an R&D engineer in the Information Systems Laboratory and the Centre for Cultural Informatics of the Institute of Computer Science, FORTH. She holds a Master of Applied Science in Electrical & Computer Engineering from the University of Toronto and a Diploma in Electrical Engineering from the Aristotle University of Thessaloniki. Maria has been actively involved in several national and international cultural information systems' projects. Her research interests include conceptual modelling, cultural information systems, source material management systems and digital libraries. Projects she has been involved include: ARIADNE, ITN-DCH, PARTHENOS, VRE4EIC and the X3ML mapping framework.

Prof. Keith G Jeffery is a past President of ERCIM, past President of euroCRIS and past Director of IT at STFC Rutherford Appleton Lab. He has extensive experience in research information management and VREs. He also has extensive experience of the relevant technologies and in particular metadata. He has participated in many EC-funded R&D projects and has worked with industry as well as academia and is now a consultant. He has many peer-reviewed publications. He holds 3 honorary/visiting professorships, is a Chartered Engineer and Chartered IT Professional and is a fellow of both the British Computer Society and the Geological Society. He chaired the EC Expert Groups on GRIDs and CLOUDs.

Dr. Zhiming Zhao obtained his Ph.D. in computer science in 2004 from the University of Amsterdam, and is currently a senior researcher leading the team of Quality-Critical Computing on Programmable Infrastructure in the context of the System and Network Engineering research lab. His research focuses on novel programming and control models for quality-critical systems on programmable infrastructures applying optimisation, Semantic linking, and artificial intelligence. He coordinated the H2020 SWITCH project and is leader of the 'Data for Science' theme within the H2020 ENVRIplus project. He leads the UvA effort in other EU projects including VRE4EIC, ARTICONF and ENVRI-FAIR.