Building an Integrated Enhanced Virtual Research Environment Metadata Catalogue

Remy, L.; Ivanović, D.; Theodoridou, M.; Kritsotaki, A.; Martin, P.; Bailo, D.; Sbarra, M.; Zhao, Z.; Jeffery, K.

DOI
10.5281/zenodo.3497055
10.1108/EL-09-2018-0183

Publication date
2019

Document Version
Author accepted manuscript

Published in
The electronic library

Building an Integrated Enhanced Virtual Research Environment Metadata Catalogue

Purpose – The building of an Integrated Catalogue of Research Assets Metadata should boost multi-disciplinary research. Such an integrated catalogue should enable researchers to solve problems or analyze phenomena that require a view across several scientific domains.

Design/methodology/approach – There are two main approaches for integrating metadata catalogues provided by different e-RIs: centralized and the distributed. We decided to implement a central metadata catalogue that describes, provides access to, and records actions on the assets of a number of e-RIs participating in the system. We chose the CERIF data model for description of assets available via the Integrated Catalogue. Analysis of popular metadata formats used in e-RIs has been conducted, and mappings between popular formats and the CERIF data model have been defined using an XML-based tool for description and automatic execution of mappings.

Findings – An Integrated Catalogue of Research Assets Metadata has been created. Metadata from e-RIs supporting Dublin Core, ISO 19139, DCAT-AP, EPOS-DCAT-AP, OIL-E, and CKAN formats can be integrated into the Catalogue. Metadata are stored in CERIF RDF in the Integrated Catalogue. A web portal for searching this catalogue has been implemented.

Research limitations/implications – Only five formats are supported at this moment. However, description of mappings between other source formats and the target CERIF format can be defined in the future using the 3M tool, an XML-based tool for describing X3ML mappings that can then be automatically executed on XML metadata records. The approach and best practices described in this paper can thus be applied in future mappings between other metadata formats.

Practical implications – The Integrated Catalogue is a part of the eVRE prototype, which is a result of the VRE4EIC H2020 project.

Social implications – The Integrated Catalogue should boost the performance of multi-disciplinary research, thus it has the potential to enhance the practice of data science and so contribute to an increasingly knowledge-based society.

Originality/value – A novel approach for creation of the Integrated Catalogue has been defined and implemented. The approach includes definition of mappings between various formats. Defined mappings are effective and shareable.

Keywords: virtual research environments, metadata catalogue, X3ML tool, CERIF**,** vocabulary homogenisation

1. Introduction

Nowadays, e-science Research Infrastructures (e-RIs) are community or domain specific, thus not allowing researchers to solve problems or analyse phenomena that require a view across several scientific domains. This limitation, together with the increasing need for multidisciplinary research, is addressed by Virtual Research Environments (VREs). JISC (the Joint Information Systems Committee) describes a VRE as comprising of "*a set of online tools and other network resources and technologies interoperating with each other to facilitate or enhance the processes of research practitioners within and across institutional boundaries*"[1]. VRE4EIC was a European Horizon 2020 project charged with the development of a Europe-wide interoperable VRE to empower research communities to perform multidisciplinary research more easily and effectively, and so accelerate innovation and collaboration in the European research community. This project aims to bridge across existing e-RIs such as EPOS, ICOS and SeaDataNet by taking VREs one step further towards the enhanced-VRE (eVRE) model with a standard reference architecture, generic reusable building blocks for VRE development and explicit cross-e-RI interoperability support. A cross-e-RI metadata catalogue essentially enables scientists to discover and utilize data and services from different communities and domains; however, the development of such a catalogue requires a context-rich metadata schema which can glue together different metadata schemas from different individual e-RI catalogues.

As mentioned, VRE4EIC has developed a reference architecture and software components for VREs based on the concept of the eVRE, a modular VRE based on standard building blocks that can easily be adapted for use by different research communities. Because the eVRE is intended to hide the IT complexity of the underlying implementation layers from its users, this implies that one of the main concerns that an eVRE has to deal with is heterogeneity in terms of protocols, data and metadata formats, techniques for accessing data, etc. The Reference Architecture (Meghini et al. 2016) defines six main conceptual components that interoperate with well-defined interfaces. In this paper, we focus on the Metadata Manager, responsible for the building, maintaining, and querying of the eVRE's internal metadata catalogue.

The motivation for creation of the cross-e-RI metadata catalogue described in this paper is to allow researchers to solve problems or analyse phenomena that require a view across several scientific domains. The Integrated Catalogue should boost the performance of multi-disciplinary research, and thus enhance the practice of data science and so contribute to an increasingly knowledge-based society. The problem which is addressed in this paper is how such an integrated catalogue can be created in the most efficient way possible. The research questions addressed by this paper are the following:

1. RQ1: Which approach for creation of catalogues for the needs of integration of e-RIs is more suitable: the centralized approach or the distributed approach?

[1] https://www.jisc.ac.uk/guides/implementing-a-virtual-research-environment-vre

2. RQ2: Which metadata schemas can be used to integrate the different metadata schemas used for different individual e-RI catalogues?

3. RQ3: Can some mapping tool enhance, and make effective and sharable the mappings of different metadata schemas from different individual e-RI catalogues to a single target schema?

2. Literature review

In the last two decades, there has been continued exponential growth in the number of digital projects providing online access to a range of information resources (Woodley, 2008). These projects generally should provide one single point from which to search all information resources preserved in digital libraries, Web resources, etc (Gibson et al., 2009). The implementation of these aggregator systems can be based on system interoperability and data integration techniques. Data Integration is a long-standing issue that entails a process of schema matching, the goal of which is to identify semantic correspondences between elements of two schemas, and of schema mapping, i.e. establishing specific relations between elements or attributes of the two schemas. Such an issue has been discussed and analysed since the early 80s (Batini et al., 1986). Since then, much work has been carried on this subject (Rahm & Bernstein, 2001) progressing even in the related field of Web semantics where data integration entails data extraction (Shvaiko & Euzenat, 2013).

There are various definitions of system interoperability in literature. Haslhofer and Klas (2010) define interoperability in the context of information systems as the ability of a system to work with or use parts of other systems. Roberts (2017) defines interoperability at data level as real-time data exchange between systems without middleware. Other authors define interoperability as exchanging metadata between two or more systems without or with minimal loss of information (NISO, 2004; ALCTS CC:DA, 2000). Hunter and Lagoze (2001) define interoperability as the ability to apply single query syntax over descriptions expressed in multiple descriptive formats. The development of VRE has to consider practices of different research communities and the technologies provided by underlying data, computing, sensor and network infrastructures. VREs have been developed within several initiatives since EU FP6 and 7 (Table 1). However, it is already clear from existing VRE initiatives that researchers who use existing VREs to conduct multidisciplinary research in environmental, earth, social and other sciences often face data heterogeneity problem. There is a need for creation of Integrated Enhanced Virtual Research Metadata Catalogue in order to allow researchers to solve problems or analyse phenomena that require a view across several scientific domains. Our final goal wasto enable users (researchers) to search e-RI platforms from one single point using a single query syntax. We created a few hypothetical or visionary use cases which show how integrated metadata catalogue can be used for multidisciplinary research (Muckensturm et al., 2018, page 14).

Table 1 VRE related initiatives

Name	Area	Domain
EPOS	EU	Earth/geophysical sciences
ENVRI / ENVRI+	EU	Environmental science
ACGT	EU	Healthcare
ANFAS	EU	Environment-water
ARIADNE	EU	General VRE
BigDataEurope	EU	General VRE
CESSDA	EU	Social sciences
CoreGRID	EU	General load-balancing architecture
CRUCID	EU	Environment water
D4Science	EU	General VRE
DARIAH	EU	Social sciences and humanities
DASHISH	EU	Arts and humanities
DECAIR	EU	Environment - air
DILIGENT	EU	General VRE
DRIVER	EU	VRE for publications
eCloud	EU	Cultural heritage sciences

ELIXIR	EU	Biological sciences
ENGAGE	EU	Social sciences and humanities
ESIMEAU	EU	Environmental sciences
ESTEEM	EU	Materials sciences and physics
Europeana v1.0-v.3.0	EU	Cultural heritage sciences
Fish4Knowledge	EU	Environmental sciences
iMarine	EU	Environment - marine
InGeoCloudS	EU	Geosciences
MICROKELVIN	EU	Environmental sciences
MyExperiment	UK	Bioscience
MyGRID	UK	General loadbalancing but applied to bioscience
NMI3	EU	Materials sciences and physics
Open Academic Environment	EU	Librarian sciences
OpenAIRE	EU	Scholarly publications
OpenAIREPlus	EU	Scholarly publications, datasets and research evaluation
PaaSage	EU	Cloud middleware using CERIF
PaNData	EU	Proton and neutron sciences

RadioNet	EU	Astronomy
Share-PSI 2.0	EU	General open data
SIMES	EU	Multimedia systems
Smart Open Data	EU	Environment – open data
Smart Tea	UK	Experimental chemistry workflow
TAMBIS	UK	Data interoperability for bioscience
TAVERNA	UK	Research workflow for bioscience
THETIS	EU	Environment – coastal zone
VPH	EU	Healthcare
WADI	EU	Environment-water

The main interoperability issue here is the heterogeneity of systems and their data models. Based on the perspective of heterogeneities in information systems, Tolk (2006) proposes six main levels of interoperability concern including syntactic interoperability (having a common structure to exchange information) and semantic interoperability (havng common information model). Haslhofer and Klas (2010) meanwhile differentiate two classes of heterogeneity: structural heterogeneity and semantic heterogeneity. Structural heterogeneities on the model level occur because of model incompatibilities. Domain and element representation conflicts produce structural heterogeneities. An example of an element representation conflict would be a ‘naming conflict’ which occurs because model elements representing the same real-world entity are given different names. An example of a domain representation conflict would be an ‘abstraction level incompatibility’ which turns up when the same real-world entities are arranged in different generalization hierarchies or aggregated differently into model elements. Domain conflicts and terminology mismatches are examples of semantic heterogeneities. Domain conflicts represent domains overlapping or domains incompatibility, while terminology mismatches include synonym and homonym conflicts. These conflicts could be resolved by homogenization of vocabularies (Naudet et al., 2010).

There are the following three techniques for resolving heterogeneities of systems and achieving metadata interoperability: model agreement, meta-model agreement, and model reconciliation. We used the last technique for the implementation of the integrated catalogue presented in this paper. If there is no central authority that can impose a metadata standard in some domain, reconciling heterogeneities among models is necessary. The reason why we selected the last approach for resolving heterogeneities of systems and achieving metadata interoperability is the fact we can't impose model or meta-model agreement in eRIs domain. Moreover, there is no initiative for the creation of some central authority for this purpose at the moment for our best knowledge. Model reconciliation includes the following techniques: language mapping (Bernauer et al., 2004; Lethi and Frankhauser, 2004), schema mapping (Pierre and LaPlant, 1998), and instance transformation (Doan et al., 2001).

The process of mapping schemata implies a lot of time and effort from experts on the source and target schemas. Many difficulties can be encountered, including misunderstandings and how to manage different versions of the mapping definitions; there is also a need for an exhaustive knowledge of the schemas being mapped to and from. There have been some attempts to automate part of the task (Rahm and Bernstein, 2001). To simplify and accelerate different parts of the task, mapping tools can help (Choi et al., 2006, Xu et al., 2006). We used the X3ML toolkit (Marketakis et al. 2016) for implementation of mapping tasks in our project. The X3ML toolkit describes schema mappings in such a way that they can be collaboratively created and discussed by experts. We decided to use this tool in order to accelerate mapping process, as well as to enable collaboration of metadata formats (CERIF, Dublin Core, etc) experts and mapping experts which we had in our research team. The X3ML toolkit is a visual tool with great usability.

3. Methodology

The methodology for implementation of the integrated eVRE metadata catalogue includes the following steps:

1. **Selection of an approach to the cooperation between e-RIs and a VRE** - Pros and cons of the centralized and the distributed approaches for integrating metadata catalogues coming from different e-RIs were analyzed. The projected usage of the eVRE system was taken into account in the process of selection of the final approach.

2. **Definition of a metadata model (schema) for the integrated catalogue -** To identify the metadata model to be used for the eVRE catalogue, an analysis of various e-RIs was carried out. The goal was to identify the most widely used metadata models. To support interoperability, as well as interdisciplinarity, across various metadata models, we have to select one model to service as the data model of integrated catalogue, based on the following criteria: 1) the model must be an accepted standard; 2) it must include all basic research entities, and in particular be able to cover non-community-specific entities; and

3) the model should also be able to flexibly deal with the various semantics used by the different communities.

3. **Creation of the integrated catalogue -** In the context of the central eVRE catalogue, after selection of e-RIs which could be data providers and definition of an integrated catalogue schema (see the previous bullet), there is a need to match and map information existing in the individual e-RI catalogues to the central eVRE metadata catalogue. The process of matching, harvesting, mapping and transforming data coming from different e-RIs is shown in Figure 1. The first step is to define the semantic matching between the source schema and the eVRE catalogue schema (steps A1 & A2). Mapping should be defined for all metadata formats dominant in the VRE related initiatives listed in Table 1. The lost of information in the process of metadata transformation (step A4) should be as less as possible. Metadata which can be used as metadata criteria for searching and filtering of catalogue are especially important to be mapped and transformed well. A transformation engine applies this matching to the set of data that have been harvested from the participating e-RI, following the source schema (steps A3 & A4). The output of this mapping populates the eVRE metadata catalogue (step A5). The final step A5 is implementation of information retrieval features over the integrated catalogue. Search of the integrated catalogue should enable researchers to solve problems or analyze phenomena that require a view across several scientific domains. Thus, metadata criteria for searching and filtering data in an integrated catalogue should be created in accordance with the previously stated main purpose of integrated catalogue. The step A4 is the execution of the result of the step A2 (which defines the required mapping). We also identified the different vocabularies used in the source schemata and homogenized them. Homogenized vocabularies improve user experience when querying the integrated catalogue by increasing the coverage of specific terms used in queries.

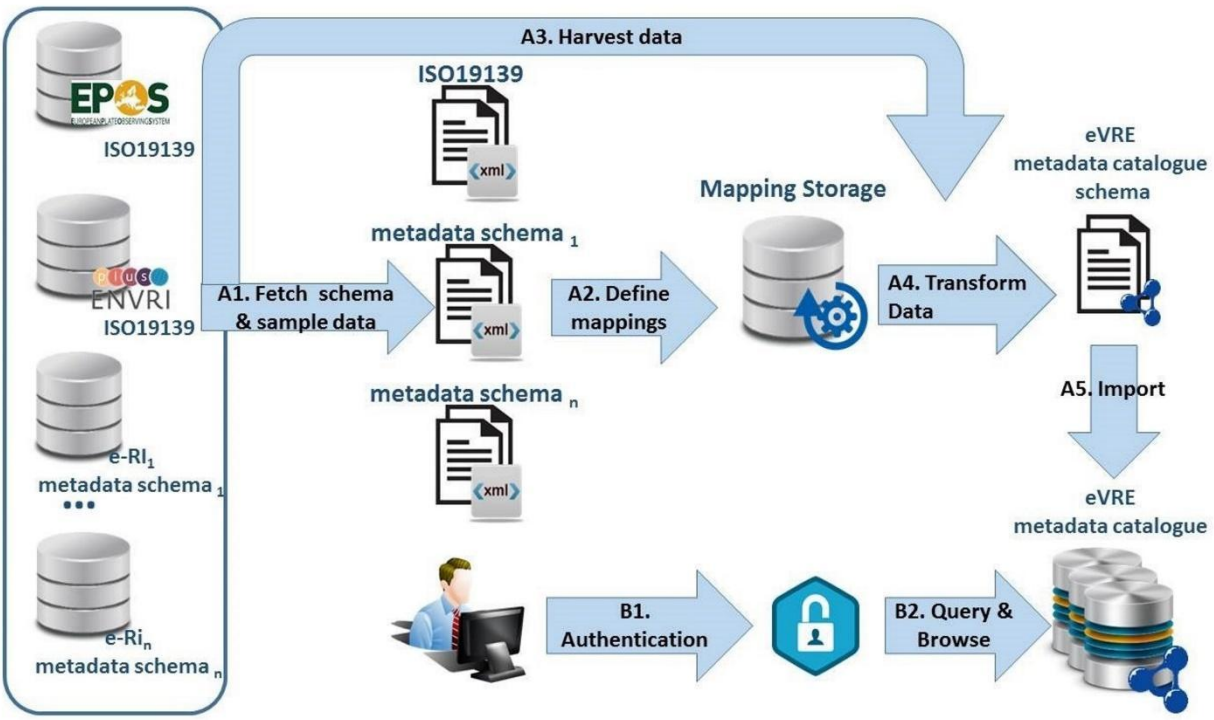

Figure 1 Creation of the eVRE metadata catalogue

We used the following data (assets) for the creation of the integrated catalogue prototype:

- EKT – Synthetic CERIF data from the National Documentation Centre of Greece
    - http://cc-refim.ekt.gr/cerif-rest-refim/
    - 26776 Organisations, 20205 Projects, 506688 Researchers, 162625 Publications
- RCUK – Gateway to Research, Research Councils UK
    - http://gtr.rcuk.ac.uk/cerif/
    - 90860 Organisations, 42092 Projects, 191421 Researchers, 169774 Publications
- FRIS – Flanders Research Information Space Research Portal
    - http://www.researchportal.be/en/
    - 2157 Organisations, 26998 Projects, 28712 Researchers, 3296 Publications
- Synthetic Locations for 117636 Organisations

- EPOS – EUROPEAN RESEARCH INFRASTRUCTURE ON SOLID EARTH
- https://www.epos-ip.org/
- 26 Datasets, 2 Facilities, 1 Equipment, 114 WebServices, 6 WADLs, 2 Software

- ENVRIplus –Environmental and Earth System Research Infrastructures
  - http://www.envriplus.eu/
  - 6 Datasets

4. Results

a. Selection of an approach to the cooperation between e-RIs and a VRE

This section presents the result of the methodology step listed as the first bullet in the Methodology section. There are two main approaches for integrating metadata catalogues coming from different e-RIs: centralized and distributed (Figure 2). Each approach has its own pros and cons, well known in distributed system design. The availability of a VRE Catalogue facilitates all VRE operations that rely exclusively on resource descriptions, such as resource discovery. For operations that require data access, such as data discovery, the centralized approach can alleviate the problem of querying multiple sites, by having a more complete overview of the data available when executing operations. On the other hand, the distributed approach makes it easier to have complete information in real-time, since it does not require propagation of updates to the central catalogue. In our case, it was decided to implement a central metadata catalogue (answering **Research Question 1**) that describes, provides access to and records actions on the assets of the e-RIs participating in the eVRE. It facilitates one important service, namely the construction of workflows across one or more RIs. The construction of workflows requires numerous accesses to resource descriptions, followed by optimisation for parallel/distributed operations; the centralized approach makes it possible to implement this access in the most efficient way possible.

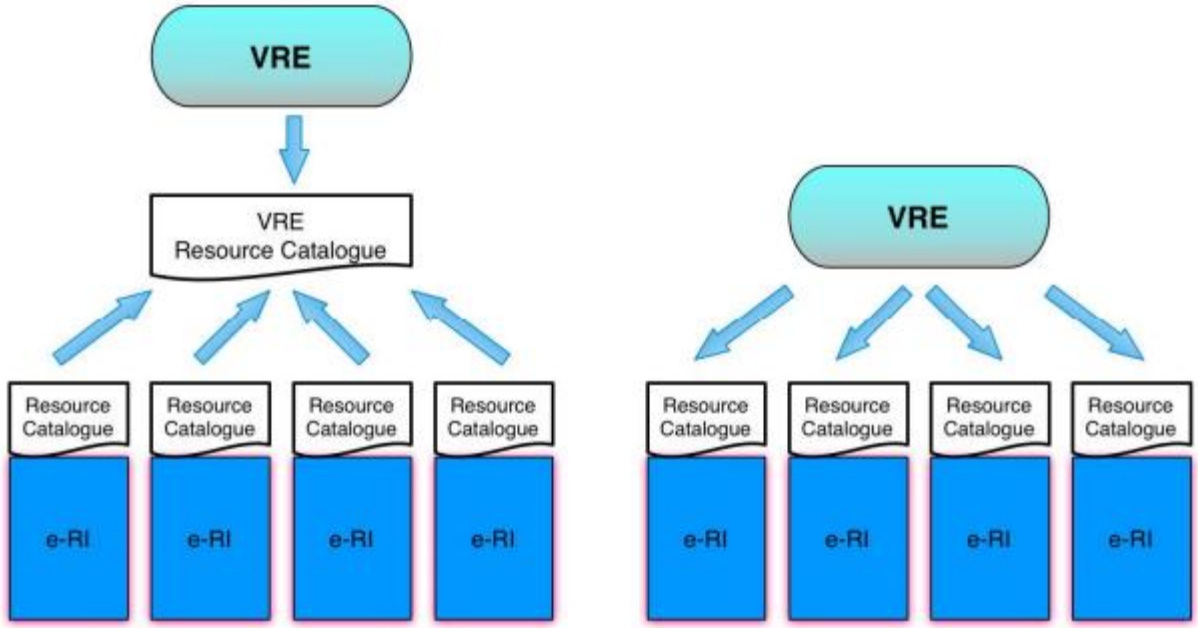

Figure 2 Centralized (left) and distributed (right) approaches to the cooperation between e-RIs and a VRE

b. Definition of a metadata model (schema) for the integrated catalogue

This section presents the result of the methodology step listed as the second bullet in the Methodology section. An analysis of various e-RIs and their data models was carried out (van Ossenbruggen et al, 2017). Here we overview the most useful standards we identified within the results we obtained from the analysis. After that, we will describe the process of selecting the target data model used for representation of research entities in the central catalogue.

i. Dublin Core

The Dublin Core schema is a small set of elements that can be used to describe Web resources (videos, images, web pages, etc.), as well as physical resources (books, CDs, artworks, etc.). Dublin Core Metadata may be used for multiple purposes, from representing simple resource descriptions, to combining metadata vocabularies of different metadata standards, to providing interoperability for metadata vocabularies in Linked Data Semantic Web implementations.

ii. ISO19115/19139

The ISO19115 standard defines the ISO schema for describing geographic information[2]. The schema provides information about the identification, extent and quality of spatial and temporal data, spatial references, and distribution of digital geographic data. The standard is applicable to cataloguing of datasets, especially geographic datasets, dataset series, individual geographic features and feature properties. The ISO19115 standard can be used to provide the metadata about geographic datasets required by the European INSPIRE directive[3]. ISO19139 defines the XML profile for ISO19115.

iii. DCAT-AP

DCAT stands for Data CATalogue vocabulary. It is a recommendation by W3C[4] that is designed to facilitate interoperability between data catalogues published on the Web. The main entities managed by the DCAT vocabulary are Catalog, Dataset and Distribution. It relies on the FoaF[5] vocabulary to describe persons and organisations. A specification of DCAT for data portals in Europe has been designed: DCAT-AP[6]. This specification introduces several additional mandatory and recommended classes like Agent, Category, Category Scheme and License document. It also introduces a status for the properties: DCAT recommends usage of some properties for each class, whereas DCAT-AP defines sets of mandatory, recommended and optional properties for each class.

iv. EPOS-DCAT-AP

EPOS-DCAT-AP is an extension of the DCAT-AP for Research Infrastructures in the environmental domain, with a specific focus on the EPOS Research Infrastructure in the subdomain of earth science. It extends the description of datasets, dataset series, equipment and services. It largely leverages existing models and vocabularies, like schema.org. EPOS-DCAT-AP was implemented by defining an RDF (Resource Description Framework) syntax that can be used for the exchange of descriptions of spatial datasets, dataset series, and services among communities.

---

[2] https://www.iso.org/standard/53798.html

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

[4] https://www.w3.org/TR/vocab-dcat/

[5] http://www.foaf-project.org/

[6] https://joinup.ec.europa.eu/asset/dcat_application_profile/description

v. OIL-E

Open Information Linking for Environmental science research infrastructures (OIL-E)[7] is a framework for addressing the semantic linking requirements of environmental science e-RIs (Martin et al. 2015). It aims to provide a machine-readable bridge between the ENVRI Reference Model (ENVRI RM)[8] used by e-RIs within the ENVRI cluster of European environmental science research infrastructures to model their architecture and design (Nieva et al. 2017), and other concept models related to research infrastructure, architecture and scientific (meta-)data. The ENVRI RM ontology within OIL-E captures all the archetypes defined across the three views for science, information and computation, providing a standard vocabulary for many of the actors, resources, information objects and computational services used in environmental science e-RIs. OIL-E is intended for linking concepts used in a variety of different standards and specifications as a means to map out and harmonise technical developments in e-RIs from an infrastructural and operational perspective.

vi. CKAN

CKAN is an open source and open architecture software platform for data management[9]. The CKAN data management platform is in use by numerous governments, organizations and communities around the world. It can be easily installed and customized for the specific needs of some organisation. Also, taking into account its open architecture it can be extended with some plugins. The CKAN platform can preserve various data types – datasets, source codes, documentations, etc. Note that CKAN itself is a platform, not a metadata standard, but it has a widely used internal metadata model, also used by some e-RIs, so we decided to consider that model (the 'CKAN model') as one of the source schemas to map from.

vii. CERIF

CERIF is a conceptual metadata model which allows a representation of research entities, their activities and their output. It has high flexibility with formal (semantic) relationships, enables quality maintenance, archiving, access, and interchange of research information covering all parts of the research life-cycle (Figure 3). CERIF supports knowledge transfer to decision makers, research managers, strategists, researchers, editors and the general public.

---

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

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

[9] https://ckan.org/

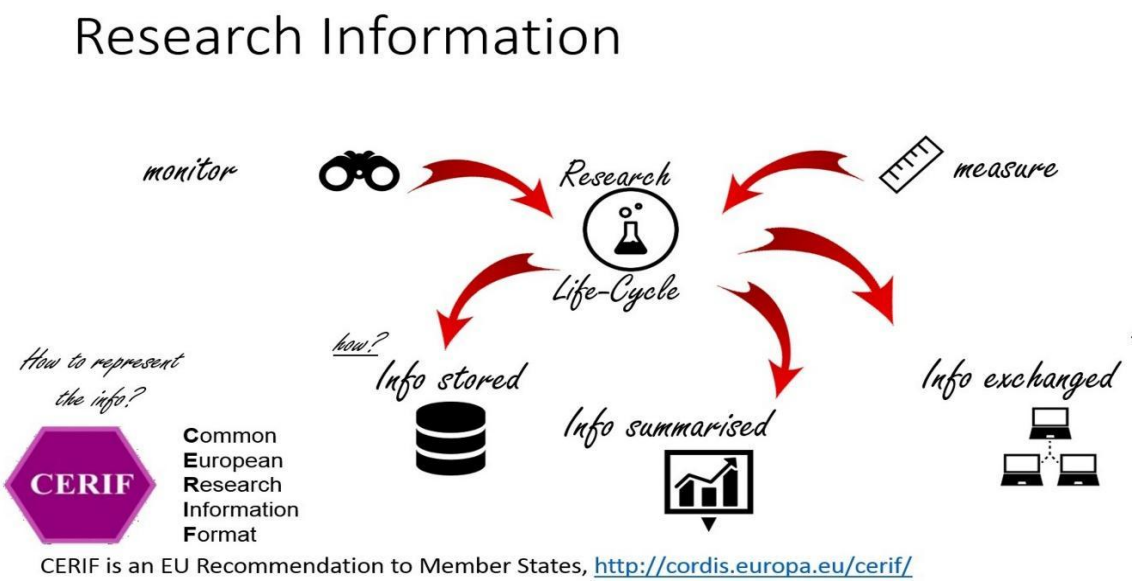

Figure 3 The Research Life-Cycle

CERIF describes base entities in the Research domain, such as person, organisation, project, publication, patent, data, facility, equipment, service, funding, measurement, indicators, identifiers and – also as entities - their relationships (Asserson et al 2002; Dvořák 2015). The relationships consist of a n-tuple with the two base entities in the relationship, the role relationship between them and date/time start and end of the relationship (which automatically provides provenance and versioning). Further attributes such as probability may be added, allowing some measure of the certainty of the relationship. Figure 4 presents an overall view over the Research domain and its related entities, where the colours indicate possible contexts, such as results (orange), quantitative outcomes (red), actors (green) and infrastructure (purple).

Since the publication of the first version of the CERIF model, it has grown in the quantity and quality of the concepts represented in it. The model became a recommendation to Member States by the European Commission in 1991[10]. It is targeted to providers of research information systems, seeking to benefit from published information and data exchange.

[10] http://cordis.europa.eu/pub/cerif/docs/cerif1991.htm

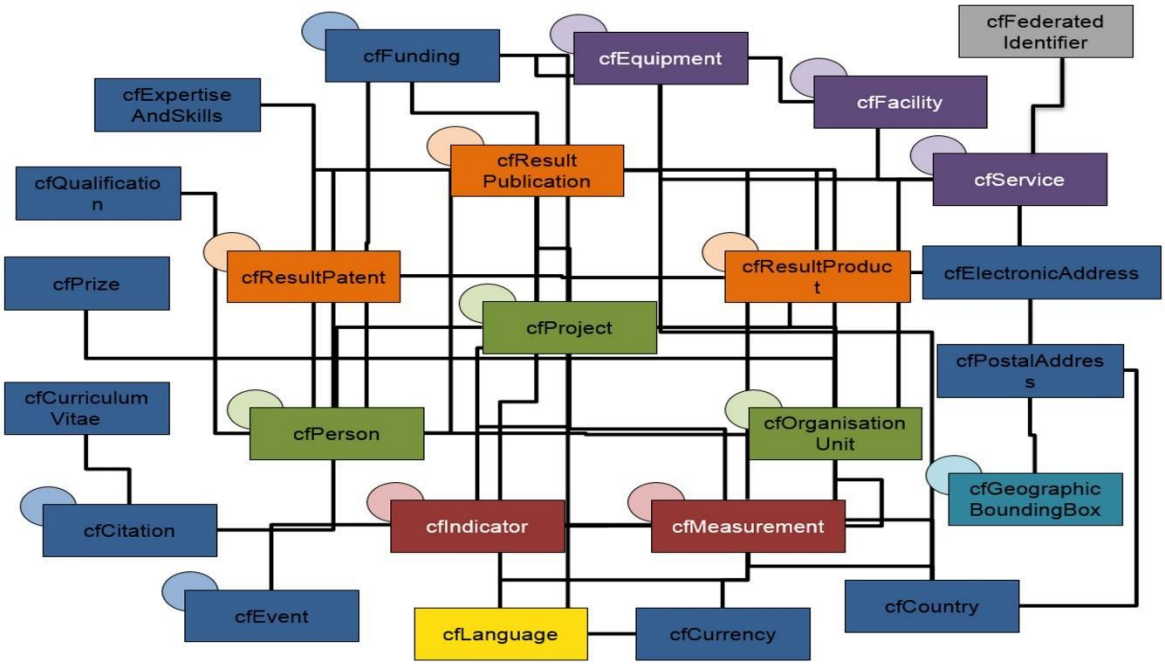

Figure 4 A snapshot of CERIF high level entities

viii. Selection of the integrated catalogue schema

Table 2 shows the most-used standard within the results we obtained from the analysis. We take as a basis for our analysis the research presented in the paper by the EPOS project group (Bailo et al., 2017). To support interoperability and interdisciplinarity across these various metadata models, we selected the data model for the integrated catalogue that meets the criteria described in the Methodology section: 1) the model must be an accepted standard; 2) it must include all basic research entities, and in particular be able to cover non-community-specific entities; and 3) the model should also be able to flexibly deal with the various semantics used by the different communities. Basically, we selected a catalogue data model that is generic enough to handle most of the divergences in different domain semantics. This means that the target model is context-rich and able to glue together different metadata schemas from different individual e-RI catalogues without significant loss of information. Upon completing our analysis, **CERIF** was the metadata model deemed to fit all our criteria (answering **Research Question 2**).

Table 2 The summary of the mapping of entities in analyzed standards

Entity	Dublin Core	ISO19115	DCAT-AP	EPOS-DCAT-AP	OIL-E	CKAN	CERIF
Research Infrastructure	--	--	--	eposap:Facility	resource	--	cfFacility
Equipment	--	--	--	eposap:Equipment	resource	--	cfEquipment
Software	dct:type=Software (DCMI Type)	--	--	eposap:SoftwareSourceCode	--	Resource (resource_type=code)	cfResultProduct (with specific semantic)
Dataset	dct:type=Dataset (DCMI Type)	scope code=Dataset	dcat:Dataset	eposap:Dataset	--	Dataset (Package)	cfResultProduct (with specific semantic)
Person	dct:Agent	(responsible party)	foaf:Person	eposap:Person	actor	creator	cfPerson
Organisation	dct:Agent	(responsible organisation)	foaf:Organization	eposap:Organisation	actor	organization (owner_org)	cfOrganisationUnit
Webservice	dct:type=Service (DCMI Type)	scope code=Service	dcat:Distribution	eposap:WebService	--	Resource	cfService (with specific semantic)
Service	dct:type=Service (DCMI Type)	scope code=Service	--	eposap:Service	computational object	--	cfService
Publications	dct:type=Text (DCMI Type)	--	foaf:Document	eposap:Publication	--	Resource (resource_type=documentation)	cfResultPublication

In the context of our eVRE, CERIF fit all the criteria we defined: it is a metadata model standard, recommended by the European Commission to EU Member States; it covers the whole scope of non-community-specific entities used by e-RIs; finally, CERIF is semantically agnostic by the use of a semantic layer allowing integration of any vocabulary used by e-RIs. Using this semantic layer, it is also possible to describe relations between vocabularies and terms to add some homogeneity between (multilingual) vocabularies.

The CERIF metadata model has been expressed in various formats, namely as an RDBMS schema and an XML encoding for interoperability among CERIF applications/installations (Jörg et al. 2012). For interoperability among CRIS (Current Research Information Systems) with different supporting metadata schemas, other approaches were also discussed (Pinto et al. 2014). In the last decade, the core technology for a widespread, distributed and structured service for research information has been the Semantic Web technology and its standard models such as RDF (Lassila & Swick, 1999). The need to define a CERIF RDF[11] encoding became apparent (Berners-Lee, 1998) and was carried out in the context of the VRE4EIC project for the most recent version of CERIF. The definition of the CERIF RDF encoding resulted from a bottom-up approach of the transformation of the existing relational structure into an ontological structure (Remy et al. 2017).

c. Creation of the integrated catalogue

This section presents the result of the methodology step listed as the third bullet in the Methodology section. In order to implement those steps, we used **a mapping tool**. The process of matching and mapping implies a lot of time and effort from experts on the source and target schemas. To simplify and accelerate the process, a tool needs to be adopted for automation. Besides enhancement of mapping development, such a tool should make the implementation of mappings more effective and sharable. We used the X3ML toolkit (Marketakis et al. 2016) presented in the next section. Moreover, we identified the different vocabularies used in the source schemas and homogenized those vocabularies. Matchings of entities and attributes between source schemas and the catalogue target schema have been performed and expressed in the X3ML language.

i. The mapping tool

The X3ML toolkit with the 3M editor[12], an open-source application suite, was chosen as the appropriate mapping technology. This toolkit allows several steps and tasks of the process of harvesting, matching, mapping and integrating the data from the sources to the eVRE metadata catalogue to be efficiently performed. Within this toolkit, 3M guides the user to specify the schema matching and the instance generators, i.e. the functions that will create the appropriate

[11] https://github.com/EuroCRIS/CERIF-RDF

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

CERIF URIs for example, or to format the dates homogeneously (step A2 – Figure 1 in the Methodology section). Another component is the X3ML engine that automatically transforms the source data into CERIF data instances applying the X3ML mapping definition file issued by 3M (step A4 – Figure 1 in the Methodology section).

The components of the X3ML toolkit addresses several of the issues identified during the process of matching and mapping (answering **Research Question 3**). 3M is the component of the X3ML toolkit which eases the process of matching by parsing and analyzing the source and target schemas, thus allowing auto-completion when selecting the entities and properties to be matched. This mechanism speeds the matching process and allows non-expert users (users that do not have an extended knowledge of the whole schema) to define a matching. The description of the matching is homogenized, which reduces the misunderstandings between experts. 3M also includes a versioning mechanism that allows storage of different versions of the matchings. And finally, the X3ML engine can be used exhaustively to test any version of the matching at any time just by providing a sample of data and applying the transformation. The result is immediately available and can be analysed to check for defaults or implemented corrections.

ii. Vocabularies homogenisation

In order to facilitate data retrieval from various sources, the e-VRE has to understand both the format and the semantics of the metadata describing the resources from those sources. The format issues are taken into account by the matching and mapping operations. The semantics define the meaning and definitions of the resources and the links among them. As CERIF allows use of any semantics, one step towards data integration must concern the integration of the semantics of the different sources.

Various information used in a dataset can be expressed using different terms in different standards: author or creator, creation dates and type of resources, etc. Various terms can be used to indicate that some resource is a dataset. For instance, in a Dublin Core record, one can define a dataset by using the property dcterms:type (http://purl.org/dc/terms/type) with the value Dataset (http://purl.org/dc/dcmitype/Dataset) coming from the Dublin Core Type Vocabulary (http://purl.org/dc/terms/DCMIType), while in a CKAN record, the free text “dataset” term can be used. All these terms and vocabularies should be harmonised so that e-VRE users can retrieve both results by asking the search engine for “datasets”.

A specific task was conducted to achieve harmonisation among metadata mapped into CERIF. We used a bottom-up approach by first listing in the matchings all the attributes that match the cerif:Classification entity, with the corresponding values identified in the source, or in the vocabularies identified or recommended by the documentation of the standard used by the source. We then categorised the attributes regarding their values in the following categories:

- Terms introduced by the matcher - the expert that does the matching.
- Terms coming from a controlled vocabulary - a list of terms officially maintained.

- Out of control terms - these are free text terms that appear in the source files and that are not part of a controlled vocabulary.

We ended up with 259 different attributes in our mappings. 60% were terms introduced by matchers. 11% use controlled vocabularies. 29% use out of control terms, which means that a third of the attributes have no recommended vocabulary to use.

The terms introduced by the matchers have been homogenised and taken from controlled vocabularies whenever possible. Most of these terms come from the CERIF vocabulary. Controlled vocabularies also need to be harmonized when different vocabularies deal with the same concept. For example, the *ADMS*[13] *status vocabulary* and the *MD_ProgressCode* share some common concepts, like the term 'completed' (http://purl.org/adms/status/Completed & MD_ProgressCode_completed). This also needed to be integrated into the catalogue. CERIF provides a way to describe that similitude by using the cerif:ReflexiveLink_Classification entity. This entity semantically links two cerif:Classification instances. The semantic link is provided by a link to a cerif:Classification instance that provides the meaning of this link. In this case, we use the SKOS[14] vocabulary to describe these relations between terms.

For the out of control terms, several solutions are possible to integrate them in the CERIF schema, depending on the attribute and the rest of the matching. The first solution is to add the value of the attribute as a keyword for the entity. In this case, the context of the attribute is lost, as we only keep the value in the destination, and not the name of the attribute. This solution is to be considered if the concept of the attribute is close to the concept of a keyword. The second solution is to use the cerif:Measurement entity. This solution is to be considered only if the value of the attribute represents a measure linked to the entity. The last solution is to create an uncontrolled vocabulary. This solution works for any kind of attribute. It consists of defining a cerif:ClassificationScheme specifically for the attribute. The various values of the attribute will be added to that vocabulary using cerif:Classification individuals. This is not a perfect solution either as the quality of this cerif:ClassificationScheme is highly dependent on the sources: if the sources allow any kind of value, the catalogue may end up with redundant classifications due to typos, etc. The following table illustrates some statistics about the solutions chosen for the out of control terms.

Table 3 Statistics about the solutions applied to identified metadata schemata in VRE4EIC

Solutions	Out of control terms concerned	Example of term
Out of control vocabulary	78%	[CKAN] resources – resource_type
Keywords	7%	[Dublin Core] Text – subject
Measurement entity	15%	[ISO 19139] resourceFormat – version

[13] https://www.w3.org/TR/vocab-adms/

[14] https://www.w3.org/2004/02/skos/

iii. Implementation of mappings

Figure 1 (the Methodology section) presents a high-level view of eVRE metadata catalogue creation. However, steps A1, A2 and A4 are more complex when considered in detail. The matching and mapping process requires expert knowledge of both the source and target schemata implying a close cooperation between domain and target schema (CERIF) experts. Typically, the process is iterative and starts by creating a first matching and mapping of a representative set of metadata elements which is tested and gradually improved and augmented. During the first phase, when the first matching and mapping is created, the experts need to explore the source and target schemata, ensure a common understanding and select a set of representative metadata elements to start with. The selection of elements depends on the desired result which is usually the answer to a typical research question. The matching and mapping progresses by building step-by-step the first matching rules and running test mappings and transformations.

Matching of all identified schemata Dublin Core, ISO19139, DCAT-AP, EPOS-DCAT-AP, OIL-E, and CKAN to CERIF RDF has been provided in the context of the VRE4EIC project (Remy et al. 2017). The full matching between those models to CERIF RDF is expressed in the 3M tool. Those mappings may be found at the following Zenodo link: https://zenodo.org/record/2548732 or by using 3M tool instance. To view the full mapping in 3M:

1) open the link http://www.ics.forth.gr/isl/3M-VRE4EIC

2) log in using username *vre4eicGuest* and password *vre4eic*

3) find the certain mapping project in the list and select it

4) click on the icon  .

There are a few basic options provided via the 3M menu for each selected mapping project. In the Info section of the mapping, users can find source schema and sample data (Dublin Core, ISO19139, DCAT-AP, EPOS-DCAT-AP, OIL-E, or CKAN), as well as target schema – CERIF RDF. In RDF, the resources are identified using URIs. A set of rules have been defined for URI generation of CERIF RDF resources. These rules have been implemented as instance generators for 3M which XML description can be found also in the Info section.

In the Matching Table section, mapping of input source to the target output is presented. In the Generators section, users can find a generator assigned for each entity which is a result of a mapping rule defined in the section Matching Table.

Thus, the X3ML Toolkit supports and encourages cooperation among experts by supporting mapping definition sharing, annotating with comments and versioning. Moreover it provides MAZE (Anyfantis, 2016), a component that helps analyze the matching, monitor the coverage of the source and target schemata and compare matchings. Figure 5 presents an example of the source schema coverage for a specific mapping definition. In this example, all parent elements have been matched while there have been a few uncovered child elements. The user can review this information easily and decide when the desired completeness has

been achieved. For instance, the 3M tool mapping #61 represents mapping of ISO19139 XML metadata to CERIF RDF with 117 parent elements and 136 child elements. 84% of the parent elements and 82% of the child elements were matched. We do not have a 100% coverage since there exist wrapper elements with no values (e.g. gmd:metadataExtensionInfo) and leaf elements (e.g. gco:CharacterString) that are used only in the instance generation process. To analyze a defined mapping in 3M tool:

1) open the link http://www.ics.forth.gr/isl/3M-VRE4EIC

2) log in using username *vre4eicGuest* and password *vre4eic*

3) find the certain mapping project in the list and select it

4) click on the option More->Analysis in the main menu

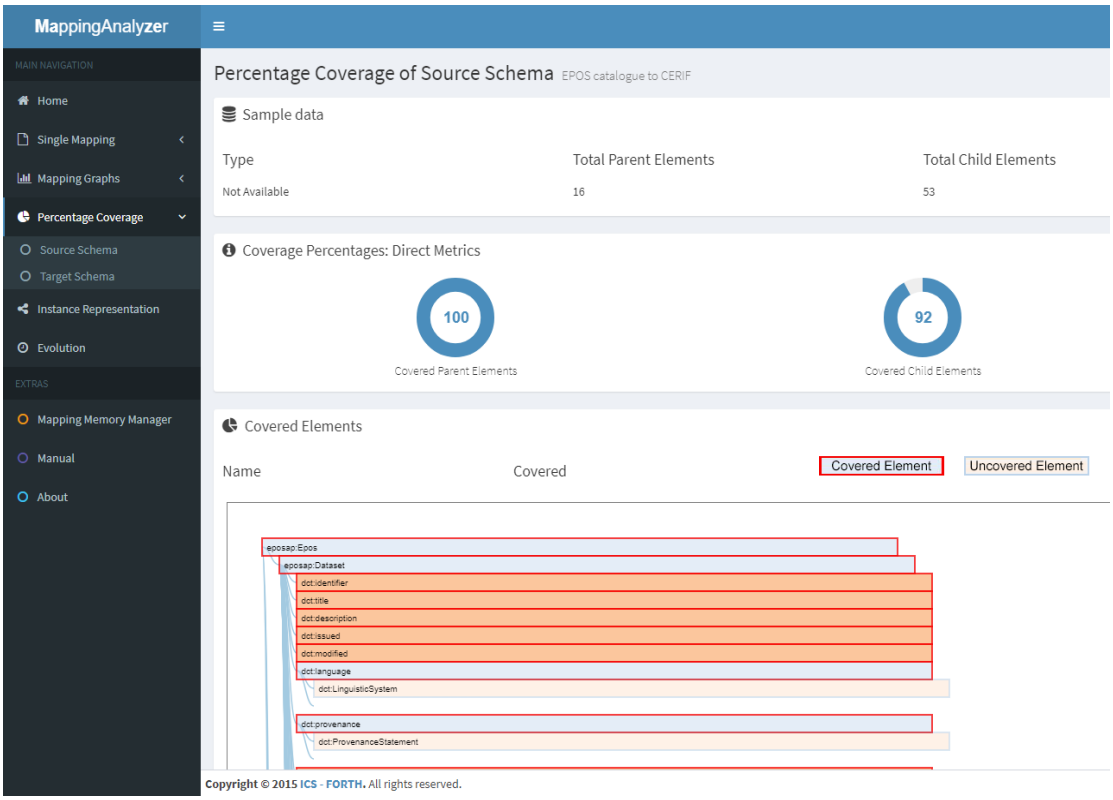

Figure 5 Source schema coverage

The X3ML toolkit provides also the RDFvisualizer component (Figure 6) which presents a transformed data instance in an intended list form easily understood by a user that is not familiar with RDF and LOD (Linked Open Data). This component helps the user understand if the original data transformed properly, as he or she had planned. The RDFvisualizer is invoked directly through the 3M Editor and so it plays the role of a debugger, providing a fast, easy way to check the mapping correctness.

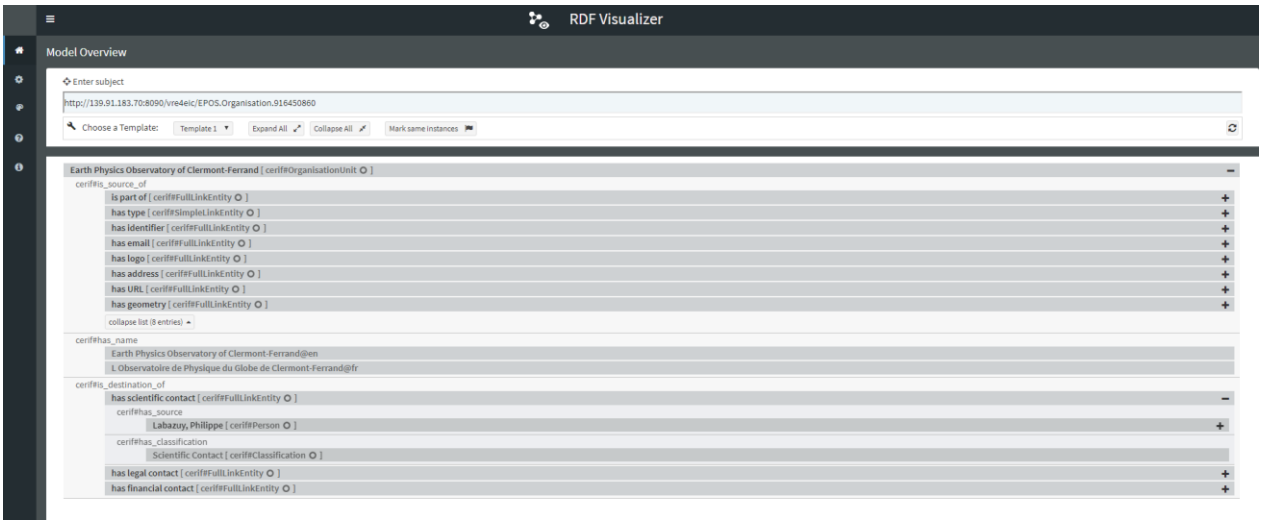

Figure 6 RDFvisualizer

For the remainder of this section, we look at two mappings in more detail: the mappings of ISO19139 and CKAN. Listing 1 presents a simplified example of a source file expressed in ISO19139 format. The result of the mapping into CERIF RDF is displayed in Figure 7.

```
<gmd:MD_Metadata …>
  <gmd:hierarchyLevel>
    <gmd:MD_ScopeCode
       codeList="…MD_ScopeCode" codeListValue="dataset" codeSpace="005">
             dataset
    </gmd:MD_ScopeCode>
  </gmd:hierarchyLevel>
  <gmd:identificationInfo>
    …
    <gmd:title>Counts of Zerynthia rumina in Doñana</gmd:title>
    …
    <gmd:CI_ResponsibleParty>
       <gmd:individualName>Jacinto Roman</gmd:individualName>
             …
        <gmd:electronicMailAddress>jroman@ebd.csic.es</gmd:electronicMailAddress>
        …
        <gmd:role>
            <gmd:CI_RoleCode codeList="…CI_RoleCode" codeListValue="author">
                  author
             </gmd:CI_RoleCode>
        </gmd:role>
     </gmd:CI_ResponsibleParty>
     …
  </gmd:identificationInfo>
</gmd:MD_Metadata>
```

Listing 1: Simplified example of source file expressed in ISO19139 format

Listing 2 presents a simplified example of CKAN JSON. The result of the mapping in CERIF RDF is displayed in Figure 8.

Ingesting both results in the same RDF triple-store allows users to retrieve both datasets metadata using only one query on the metadata catalogue, with homogeneous results.

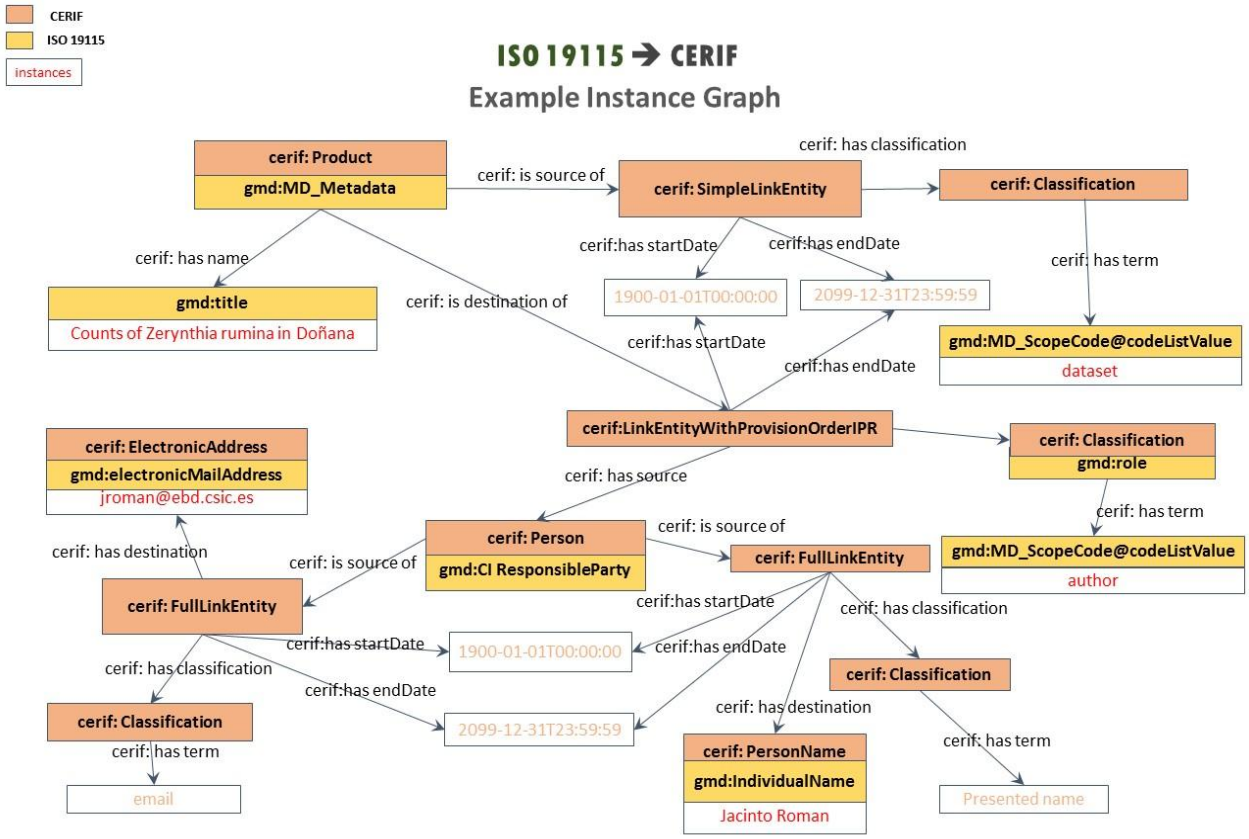

Figure 7 ISO19139 data mapped to CERIF

```
{
        …
        "result":
        {
                "type": "dataset",
                "title": "UK: Adur District Council Spending Data",
                "author": "Lucy Chambers",
                "author_email": "lucy.chambers@gmail.com"
        }
}
```

Listing 2 Simplified example of source file expressed in the CKAN JSON format

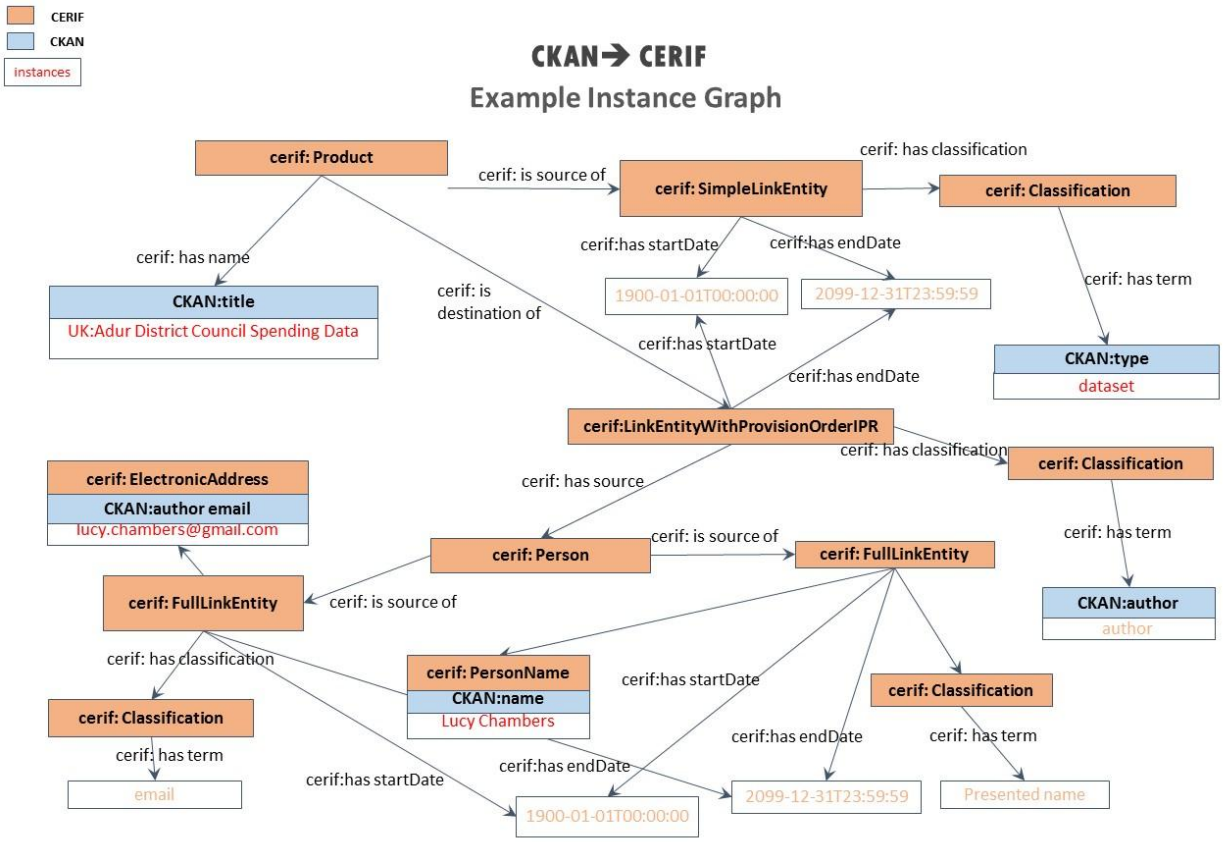

Figure 8 CKAN data mapped to CERIF

iv. Prototyping

A prototype e-VRE Metadata Catalogue has been implemented using CERIF RDF resource descriptions that were produced via the mappings described above. Both real and synthetic data were harvested (step A3 in Figure 1 in the Methodology section), transformed (step A4) and ingested (step A5) in the e-VRE Metadata Catalogue (https://www.vre4eic.eu/evre/software). Assets used in the prototype have been described in the last part of the Methodology section. The experimental e-VRE Metadata Catalogue (instance accessed on 26/9/2018) comprises approximately 53M triples. The harmonized vocabularies together with the semantic layer of CERIF were used to populate the catalogue. We also implemented a prototype of web GUI for authorized searching of ingested assets (steps B1 and B2 in Figure 1). Source code is available at the github repository: https://github.com/vre4eic.

5. Discussion

The results of the analysis of various e-RIs and their data models revealed significant heterogeneity in the usage of metadata models. Some e-RIs use standard models, but there are also a lot of custom-made metadata models that fill the needs of specific communities. A set of

standards and schemata have been identified as important for the e-VRE community and have been matched and mapped to CERIF. All the necessary elements from the matched schemata have been covered and in cases where we had a significant amount of source data (such as for the EPOS DCAT-AP and CKAN metadata) we tested them by loading them into the central metadata catalogue and performing various test queries. Moreover, vocabularies (controlled or free) used or recommended by the standards have been collected and harmonized. They form a core set of vocabularies that inform the semantic layer of CERIF, allowing for consistent classification of entities and relationships in the integrated catalogue. We ended up with 259 different attributes in our mappings which should be mapped to the semantic layer of CERIF: 60% were terms introduced by matchers; 11% use controlled vocabularies; 29% use out of control terms. Unfortunately, the scientific landscape includes a wide range of standards and schemas making integration a difficult task. The process of matching and mapping implies a lot of time and effort from experts on the source and target schemas. To simplify and accelerate the process, we adopted X3ML tool for automation. Besides enhancement of mapping development, the tool made the implementation of mappings more effective and sharable. We didn't achieve 100% completeness for implemented metadata model mappings for various reasons; for instance, there exist wrapper elements with no values.

A prototype e-VRE Metadata Catalogue has been implemented as it is described in the previous section. Since we followed a centralized approach with one common eVRE metadata catalogue, the need to enhance the catalogue with data from new (not matched yet) schemas in the future is obvious. However, the mapping technology used supports collaboration and the template matchings that we have implemented provide a solid basis for creating new matchings and mappings quickly and efficiently in the future. The VRE4EIC project has set up a systematic methodology for evaluating the project results and for assessing their impact. We organized three workshops to elicit feedback from the scientific community regarding the eVRE metadata catalog prototype. Those evaluation workshops have been co-located with ICIST 2018, CRIS 2018 and IWSG 2018 conferences, respectively. Collected feedback has been used to improve the prototype. Also, a set of competency queries was created to test and validate the contents of the eVRE Metadata Catalogue.

A metadata catalogue covering the datasets provided by a group of e-RIs can be created using defined mappings similar as prototype implemented for the needs of the VRE4EIC project. Also, a similar approach can be applied in order to build an integrated metadata catalogue for some other domain.

An integrated catalogue, whether the prototype developed for the VRE4EIC project or a new catalogue developed in the future based on the methodology and mappings defined above, should provide a boost for multi-disciplinary research, thus enhancing the further development of data science as well as contributing to knowledge-based society.

Conclusion

Allowing researchers to discover, access and use resources from various domains is not trivial. The heterogeneity in terms of technologies and standards used by different research

communities does not allow for direct interoperation between e-RIs. The concept of a VRE was defined to help solve this problem.

VRE4EIC has built a reference architecture for an eVRE (enhanced VRE) that helps provide researchers with data science capabilities that cross e-RI and disciplinary boundaries. The architecture defined relies on a metadata catalogue that stores the metadata describing the resources provided by multiple e-RIs. A prototype eVRE Metadata Catalogue has been implemented using CERIF RDF resource descriptions that have been produced via mappings to a single common schema from several source schemas in use today. To feed this catalogue, metadata was harvested from the e-RIs to be stored within the eVRE. In order to achieve this, an analysis was done within the VRE4EIC project to determine some of the most used standards in the environmental and earth science research community. The work presented in this paper summarizes the matching and mapping work performed during the VRE4EIC project. A set of standards and schemata were identified as important for the eVRE community and were matched and mapped to CERIF, a standard for research information endorsed by the European Commission. The CERIF format was selected as the target research domain format because it provides a conceptual metadata model which allows for representation of research entities, their activities and their output. It has high flexibility with formal (semantic) relationships, and enables quality maintenance, archiving, access and interchange of research information covering all of the research life-cycle. Moreover, vocabularies (controlled or free) used or recommended by the standards have been collected and harmonized. They form a core set of vocabularies part of the semantic layer of CERIF.

The X3ML toolkit has been used to construct automatic mappings and a common matching language. The methodology used to build the matching can be reused for other standards that may also need to be integrated within any VRE built according to the eVRE reference architecture. By being expressed using the same standard, metadata are easily findable for researchers, no matter the original format of the metadata. The homogenisation of the metadata also allows homogenisation of the vocabularies used to categorise the metadata.

The prototype implementation of the e-VRE Metadata Catalogue proved the feasibility of building an integrated, context aware catalogue that supports interoperability and reuse across RIs. Further work can be done towards improvement of such a catalogue, but it became apparent that matching and mapping schemata to a common metadata schema such as CERIF is an effective, efficient way to achieve interoperability and contextual awareness among RIs.

References

ALCTS CC:DA. 2000. Task Force on Metadata: Final Report. Association for Library Collections & Technical Services (ALCTS). Available at: http://www.libraries.psu.edu/tas/jca/ccda/ tf-meta6.html

Anyfantis, N. (2016), *Mappings Management for Ontology-based Integration*, Master thesis, University of Crete, available at: https://www.ics.forth.gr/_publications/anyfantis_thesis.pdf (accessed 18 September 2018)

Asserson, A, Jeffery, K.G, Lopatenko, A (2002), "CERIF: Past, Present and Future", *in Adamczak,W & Nase,A (Eds): Proceedings CRIS2002 6th International Conference on Current Research Information Systems; Kassel University Press ISBN 3-0331146-844,* pp 33-40, available at: https://dspacecris.eurocris.org/handle/11366/131 (accessed 18 September 2018)

Bailo, Daniele, Damian Ulbricht, Martin L. Nayembil, Luca Trani, Alessandro Spinuso, and Keith G. Jeffery. "Mapping solid earth Data and Research Infrastructures to CERIF." *Procedia Computer Science* 106 (2017): 112-121.

Batini, C, Lenzerini, M and Navathe, S B (1986), "A comparative analysis of methodologies for database schema integration", *ACM Computing Surveys*, Vol. 18, No. 4, 323–364. DOI: 10.1145/27633.27634

Bernauer, M., Kappel, G., and Kramler, G. (2004), "Representing XML schema in UML - a comparison of approaches". In ICWE, N. Koch, P. Fraternali, and M. Wirsing, Eds. Lecture Notes in Computer Science, vol. 3140. Springer, pp. 440–444.

Berners-Lee, T (1998), *Why RDF model is different from the XML model*, available at: https://www.w3.org/DesignIssues/RDF-XML.html (accessed 18 September 2018)

Choi, N., Song, I.Y. and Han, H., (2006), A survey on ontology mapping. *ACM Sigmod Record*, Vol. *35,* No. 3, pp.34-41.

Doan, A., Domingos, P. and Halevy, A.Y. (2001), "Reconciling schemas of disparate data sources: A machine-learning approach". In *ACM Sigmod Record*, Vol. 30, No. 2, pp. 509-520. ACM.

Dvořák, J (2015), *CERIF 1.6 Tutorial [Paris]. euroCRIS Membership Meeting 2015 – Spring (AMUE, Paris, May 11-12, 2015)*, available at: http://hdl.handle.net/11366/380 (accessed 18 September 2018)

Gibson, I., Goddard, L. and Gordon, S. (2009) One box to search them all: Implementing federated search at an academic library. *Library Hi Tech*, Vol. *27*, No. 1, pp.118-133.

Haslhofer, B. and Klas, W. (2010), "A survey of techniques for achieving metadata interoperability". *ACM Computing Surveys (CSUR)*, Vol. *42, No.* 2, p.7

Jörg, B, Dvořák, J. and Vestdam, T (2012), "Streamlining the CERIF XML Data Exchange Format", *In 11th International Conference on Current Research Information Systems (CRIS2012): "e-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production", Prague, Czech Republic, June 6-9,* 221–230, available at http://hdl.handle.net/10760/17355 (accessed 18 September 2018)

Lassila, O and Swick, R (1999), “Resource Description Framework (RDF) Model and Syntax Specification,” *W3C Recommendation*, available at http://www.w3.org/TR/REC-rdf-syntax/ (accessed 18 September 2018)

Lethi, P. and Frankhauser, P. (2004), “XML data integration with OWL: Experiences and Challenges“. In 2004 Symposium on Applications and the Internet (SAINT 2004). IEEE Computer Society, Tokyo, Japan, pp. 160–170.

Marketakis, Y, Minadakis, N, Kondylakis, H, Konsolaki, K, Samaritakis, G, Theodoridou, M, Flouris, G and Doerr, M (2016), "X3ML Mapping Framework for Information Integration in Cultural Heritage and Beyond", *International Journal on Digital Libraries (IJDL), Special Issue on* "*Extending, Mapping and Focusing the CIDOC CRM*", Vol. 18, No. 4, 1-19. DOI: 10.1007/s00799-016-0179-1

Martin, P, Grosso, P, Magagna, B, Schentz, H, Chen, Y, Hardisty, A, Los, W, Jeffery, K, de Laat, C and Zhao, Z (2015), "Open information linking for environmental research infrastructures", *In 2015 IEEE 11th International Conference on e-Science (e-Science)*, pp. 513–520. IEEE. DOI: 10.1109/eScience.2015.66

Meghini, C, Jeffery, K, Concordia, C, Marchetti, E, Patkos, T, Minadakis, N, Marketakis, Y, Chrysakis, I, van Ossenbruggen, J and Wielemaker, J (2016), *Deliverable D3.1 Architecture Design*, available at https://www.vre4eic.eu/images/Public_deliverables/D3.1_Architecture_Design.pdf (accessed 18 September 2018)

Muckensturm, M, Ivanovic, D, Remy, L, Zhao, Z, Bailo, D and Zuiderwijk, A (2018), *Deliverable D2.6 Use-case report*, available at https://www.vre4eic.eu/images/Public_deliverables/D2.6-Use_case_report_second_version_PU.pdf (accessed 18 September 2018)

Naudet, Y., Latour, T., Guedria, W. and Chen, D. (2010) "Towards a systemic formalisation of interoperability". *Computers in Industry*, Vol. *61*, No. 2, pp.176-185.

Nieva de la Hidalga, A, Magagna, B, Stocker, M, Hardisty, A, Martin, P, Zhao, Z, Atkinson, M and Jeffery, K (2017) *The ENVRI Reference Model (ENVRI RM) version 2.2*, DOI: 10.5281/zenodo.1050349

NISO. 2004. Understanding Metadata. National Information Standards Organization (NISO). Available at: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.

Rahm, E and Bernstein, P A (2001), "A survey of approaches to automatic schema matching", *VLDB Journal*, Vol. 10, No. 4, 334–350, DOI: 10.1007/s007780100057

Pierre, M. S. and LaPlant, W. P. (1998), "Issues in crosswalking content metadata standards". Tech. rep., National Information Standards Organization (NISO). October. Available at: http: //www.niso.org/press/whitepapers/crsswalk.html.

Pinto, C S, Simões, C and Amaral, L (2014), "CERIF - Is the standard helping to improve CRIS?", Procedia Computer Science, Vol. 33, 80–85, DOI: 10.1016/j.procs.2014.06.013

Remy, L, Ivanovic, D, Ossenbruggen, J, Patkos, T, Kritsotaki, A, Sbarra, M and Martin, P (2017), *Deliverable D4.2 Matching and mapping VRE Elements to CERIF*, available at

https://www.vre4eic.eu/images/Public_deliverables/D4.2_Matching_and_mapping_VRE_elements_to_CERIF.pdf (accessed 18 September 2018)

Roberts, Bobby (2017), "Integration vs Interoperability: What's the Difference? ", Surgical Information Systems, available at: https://blog.sisfirst.com/integration-v-interoperability-what-is-the-difference

Shvaiko, P and Euzenat, J (2013), "Ontology Matching: State of the Art and Future Challenges", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25(X), 158–176, DOI: 10.1109/TKDE.2011.253

van Ossenbruggen, J, Bogaard, T, Patkos, T, Martin, P, Brasse, V, and Bailo, D (2017), *Deliverable D4.1 Review of existing VRE Metadata*, available at https://www.vre4eic.eu/images/Public_deliverables/D4.1_Review_of_existing_VRE_Metadata.pdf (accessed 18 September 2018)

Yin, Y, Zuiderwijk, A, van Ossenbruggen, J, Concordia, C, Theodoridou, M, Remy, L (2018), *Deliverable D5.4 Strategies for the VRE end-users to handle security, privacy and trust issues – second version*, available at https://www.vre4eic.eu/images/Public_deliverables/D5.4_Strategies_for_the_VRE_end_users_to_handle_security_privacy_and_trust_issues-second_version.pdf (accessed 18 September 2018)

Xu, Z., Zhang, S. and Dong, Y. (2006), Mapping between relational database schema and OWL ontology for deep annotation. In *Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence* (pp. 548-552). IEEE Computer Society.

Woodley, M.S. (2008.) Crosswalks, metadata harvesting, federated searching, metasearching: Using metadata to connect users and information. Getty Research Institute.