VRE4EIC

A Europe-wide Interoperable Virtual Research Environment to Empower Multidisciplinary Research Communities and Accelerate Innovation and Collaboration

Deliverable D4.1

Review of existing VRE Metadata

Document version: 1.0

VRE4EIC DELIVERABLE

Name, title and organisation of the scientific representative of the project's coordinator:
Mr Philippe Rohou t: +33 4 97 15 53 06 f: +33 4 92 38 78 22 e: philippe.rohou@ercim.eu
GEIE ERCIM, 2004, route des Lucioles, Sophia Antipolis, F-06410 Biot, France
Project website address: http://www.vre4eic.eu/

Project	
Grant Agreement number	676247
Project acronym:	VRE4EIC
Project title:	A Europe-wide Interoperable Virtual Research Environment to Empower Multidisciplinary Research Communities and Accelerate Innovation and Collaboration
Funding Scheme:	Research & Innovation Action (RIA)
Date of latest version of DoW against which the assessment will be made:	14.01.2015
Document	
Period covered:	M1-M18
Deliverable number:	4.1
Deliverable title	Review of existing VRE Metadata
Contractual Date of Delivery:	31/03/2017
Actual Date of Delivery:	31/03/2017
Editor (s):	Jacco van Ossenbruggen (CWI)
Author (s):	Tessel Bogaard (CWI), Theodore Patkos (FORTH), Paul Martin (UvA), Valérie Brasse (euroCRIS), Daniele Bailo (INGV)
Reviewer (s):	Maria Theodoridou (FORTH), Carlo Meghini (CNR)
Participant(s):	**CWI**, FORTH, euroCRIS, INGV, UvA
Work package no.:	4
Work package title:	Interoperability, metadata and research contextualisation
Work package leader:	euroCRIS
Distribution:	Public
Version/Revision:	1.0
Draft/Final:	Final
Total number of pages (including cover):	22

What is VRE4EIC?

VRE4EIC develops a reference architecture and software components for VREs (Virtual Research Environments). This e-VRE bridges across existing e-RIs (e-Research Infrastructures) such as EPOS and ENVRI+, both represented in the project, themselves supported by e-Is (e-Infrastructures) such as GEANT, EUDAT, PRACE, EGI, OpenAIRE. The e-VRE provides a comfortable homogeneous interface for users by virtualizing access to the heterogeneous datasets, software services, resources of the e-RIs and also provides collaboration/communication facilities for users to improve research communication. Finally it provides access to research management /administrative facilities so that the end-user has a complete research environment.

Disclaimer

This document contains description of the VRE4EIC project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the VRE4EIC consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

VRE4EIC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676247.

Table of Contents

1 Introduction **6**

2 Methodology **7**

3 Overview of metadata usage in current VREs and RIs **8**

3.1 Dublin Core 8

Original Design 8

Typical Usage 9

Potential Usage within a VRE 9

3.2 ISO 19115 10

Original Design 10

Typical Usage 10

Potential Usage within a VRE 10

3.3 DCAT 10

Original Design 10

Typical Usage 11

Potential Usage within a VRE 11

3.4 CKAN 12

Original Design 12

Typical Usage 12

Potential Usage within a VRE 13

3.5 OIL-E 13

Original Design 13

Typical Usage 14

Potential Usage within a VRE 15

3.6 CERIF 15

Original Design 15

Typical Usage 16

Potential Usage within a VRE 17

3.7 Research data contextualisation: quality and validation metadata 18

4 Metadata interoperability analysis **18**

4.1 Common and cross-domain metadata 19

Direct versus mapping of common elements 19

Expressivity versus simplicity trade-off 19

Spatial and temporal coverage 20

4.2 Domain-specific metadata 20

Event-centric metadata models 20

Classification metadata 20

Tabular and statistical datasets 21

Workflow-oriented metadata 21

Provenance-oriented metadata 21

5 Summary and conclusions **22**

5.1 Summary 22

5.2 Impact on metadata matching and mapping 22

1 Introduction

This document provides an overview of metadata usage in existing VREs. It provides an analysis of the metadata elements they have in common and the differences in terms of semantics and syntax. We identify which elements are generically applicable, and which are specific for the research domain in which they have originally been developed. We provide extra details for metadata that is targeted at cross-domain interoperability, and metadata that is designed to contextualize research data in terms of scope, implicit or explicit measures of data quality, and validity estimates obtained from automatically analyzed data.

The results described here will be one of the inputs to Task 4.2 on VRE metadata matching and mapping.

After briefly discussing the methodology used in Chapter 2, Chapter 3 will discuss metadata standards and models, including Dublin Core, ISO 19115, DCAT, CKAN, OIL-e and CERIF and other metadata standards used for research contextualization. Chapter 4 discusses interoperability opportunities and challenges, both in terms of common, cross-domain metadata elements and in terms of more domain-specific elements. Finally, we will summarize the discussion and sketch the impact on metadata matching and mapping in Chapter 5.

2 Methodology

A requirements elicitation process under users and developers of existing (e-)RIs and VREs and the characterisations used for the deliverable “D3.2 Gap analysis” has provided input for this deliverable. The (e-)RIs and VREs investigated include EPOS, ENVRIplus, West-Life, MuG, READ, BlueBRIDGE, OpenDreamKit, PhenoMeNal, VI-SEEM, ARIADNE, SoBigData, and PARTHENOS. These systems have been selected as key systems in the current VRE landscape, and each was characterised in terms of the following functional areas (see D3.2 for details):

- Identification and citation
- Curation
- Cataloguing
- Data processing
- Optimization
- Provenance
- Collaboration, training & support

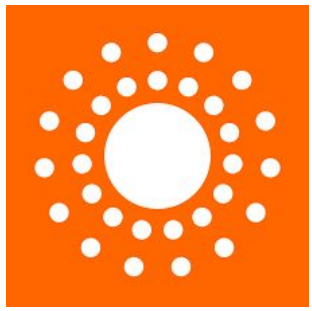
Based on the information gathered in 3.2 on characterisations in terms of metadata usage for the above functionalities, we provide an overview of the most commonly used metadata standards, and how these are used for what purpose. We continue with a discussion of how metadata standards impacts interoperability across VREs and (e-)RIs and what is needed to improve interoperability across e-RIs and VREs.

3 Overview of metadata usage in current VREs and RIs

In this section we give a brief overview of the key metadata models used in the current e-RI and VRE landscape: Dublin Core, ISO 19115, DCAT, CKAN, OIL-e and CERIF. For each model we give a brief overview of its background and original design, how it is typically used and what its potentials are for usage within the context of an e-VRE system.

3.1 Dublin Core

Original Design



The original Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. The Dublin Core standard was later extended to include two levels: Simple and Qualified. Simple Dublin Core comprises the original fifteen elements; Qualified Dublin Core includes three additional elements (Audience, Provenance and RightsHolder), as well as a group of element refinements (also called qualifiers) that refine the semantics of the elements in ways that may be useful in resource discovery. In 2012, the larger vocabulary known as The Dublin Core Metadata Initiative (DCMI) Metadata Terms became the new standard, incorporating both the simple and qualified versions of the earlier standards.

Dublin Core states as its goals: Simplicity of creation and maintenance; Commonly understood semantics; International scope; Extensibility. More information on Dublin Core can be found on the main site[1].

Typical Usage

Originating from the library community, DC is used in many domains to describe books, articles and other published creative works. In addition, it is used in the cultural heritage community to describe museum and other cultural artifacts. Its generality and simplicity typically helps to enable high-level interoperability among potentially very heterogeneous organisations without requiring complex IT knowledge, skills or infrastructure. For interoperability on more domain-specific levels, DC offers a wide range of mechanisms to specialize, extend or profile the more generic solutions. These mechanisms, however, tend to require more skills and expertise, and their uptake is not as extensive as the generic solutions.

[1] http://dublincore.org/

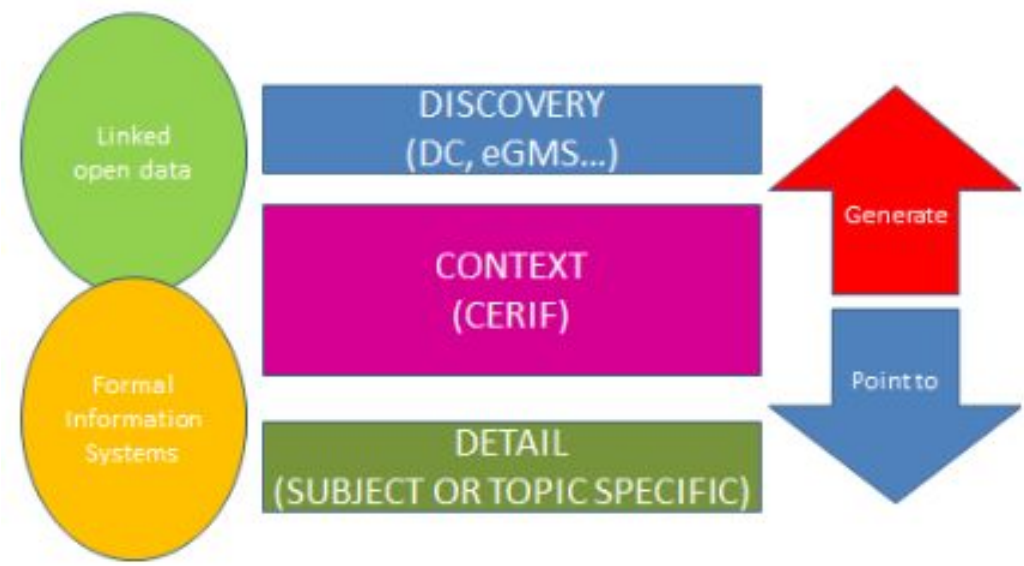


Figure: 3-Layer Model for Metadata

Potential Usage within a VRE

Referring to the three layer architecture proposed by Jeffery et al.[2], and also used in the context of VRE4EIC and the EPOS Research Infrastructure, Dublin Core can be used as a discovery layer.

The three layer metadata is structured as follows (see also figure above):

1. The discovery layer, using Dublin Core as metadata system extended to include the capability to generate from the underlying contextual layer – in addition to Dublin Core – DCAT, INSPIRE and both CKAN and eGMS to allow integration with government open data (data.gov) sources;

2. The contextual layer, using CERIF (Common European Research Information Format, recommended to the EU Member States as a tool to harmonize databases on research projects);

3. The detailed layer, which includes detailed metadata standards by domain or even individual database for each kind of data (or software, computer resources or detectors/instruments) to be (co)-processed.

The first version of DC was machine readable but not machine understandable (Jeffery 1999) and a formalized DC was developed. The current thinking is that DC provides a high-level associative-descriptive metadata and that more detailed, formal and domain specific associative-descriptive metadata sets are required for real interoperability or advanced information processing including query improvement and results explanation (Lagoze 2000).

In the experience of some of the Research Infrastructures involved in this project, Dublin Core can be used successfully for domain independent generic discovery. In this sense it is very useful.

However for more domain-specific search, where the usage of contextual metadata is needed, Dublin Core can't respond to the complex requirements raised from the scientists who need detailed data and metadata. DC is, in theory, extensible and promotes domain-specific application profiles. But in practice, many communities do not seem to use that, with the result that they provide generic access

[2] Jeffery, K., Asserson, A., Houssos, N., & Jörg, B. (2013). A 3-Layer Model for Metadata. Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013, 3–5

to their metadata. This doesn't enable, of course, scientists to get the data they seek for, thus becoming not really usable for the research data lifecycle[3].

3.2 ISO 19115

Original Design

The ISO 19115:2003 "Geographic Information - Metadata" standard defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data[4]. It defines:

- mandatory and conditional metadata sections, metadata entities, and metadata elements;
- the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data);
- optional metadata elements - to allow for a more extensive standard description of geographic data, if required;
- a method for extending metadata to fit specialized needs.

For a concise overview of the often used "core" of ISO 19115:2003, we refer to Annex III of "GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe, v1.0.1".

Typical Usage

While the specification was updated in 2014, many applications are still based on the "core" profile of the 2003 version. Specifically, a profile of ISO 19115:2003 was adopted in 2007 as the common metadata standard for the Infrastructure for Spatial Information in the European Community (INSPIRE). The other profiles of ISO 19115 in use in European Member States have been made compliant with INSPIRE. The INSPIRE Directive aims to create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment. This European Spatial Data Infrastructure (SDI) should enable the sharing of environmental spatial information among public sector organisations, facilitate public access to spatial information across Europe and assist in policy-making across boundaries[5].

Potential Usage within a VRE

Many research datasets have a key geospatial aspect that is used in discovery and other important activities. Finding, reusing, comparing and integrating such datasets from different sources without interoperability of the (often complex) geospatial metadata is not possible on a large scale. The wide adoption of ISO 19115 within and outside Europe, along with the available mappings from and to other geospatial data formats, makes it a pivotal standard for geospatial metadata interoperability.

[3] http://www.data-archive.ac.uk/create-manage/life-cycle
[4] https://www.iso.org/standard/26020.html
[5] http://inspire.ec.europa.eu

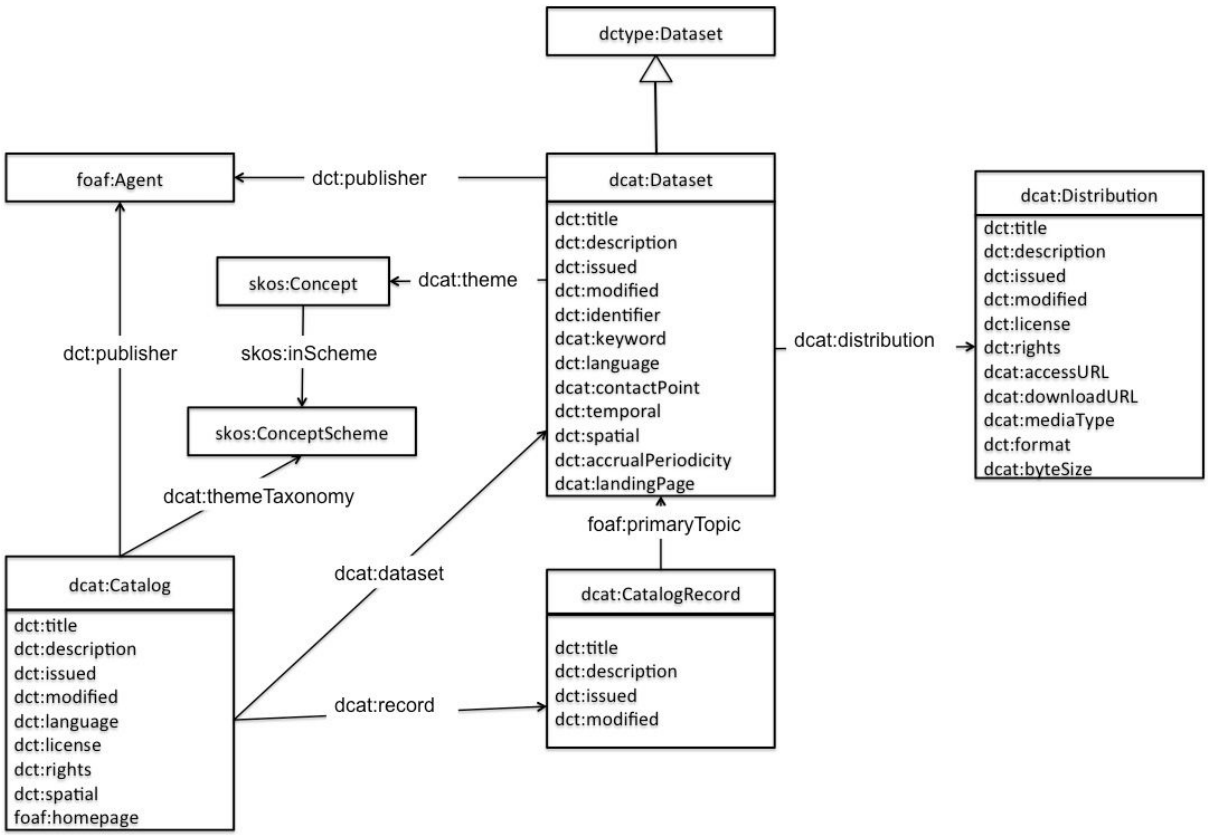
3.3 DCAT

Original Design

The W3C Recommendation DCAT[6] is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. DCAT defines three main classes:

- dcat:Catalog represents the catalog
- dcat:Dataset represents a dataset in a catalog.
- dcat:Distribution represents an accessible form of a dataset

An overview of the vocabulary is provided in the figure[7] below:



Typical Usage

DCAT originates from the need to describe open government data on the web in an interoperable way. By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation.

[6] https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
[7] https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/dcat-model.jpg

Potential Usage within a VRE

DCAT could be used as a metadata format to describe and exchange descriptions of research metadata available on the web within an e-RI or VRE context. Compared to CERIF and Dublin Core, it is much smaller in scope, focusing on metadata for data catalogues. It is also a relatively young format, with limited uptake outside the community of intended use. We see the potential of direct use of DCAT within a VRE mainly as a lightweight, that is, low overhead, vocabulary that could be used as a target for mapping more comprehensive formats into. This would allow exposing catalogues on the web in simplified form, and provide some generic, high level of interoperability across dataset catalogues.

In addition, specific application profiles for DCAT are being developed that are relevant for the use of DCAT in this domain. DCAT-AP is a profile based on DCAT for describing public sector datasets in Europe. Its basic use case is to enable cross-data portal search for data sets and make public sector data better searchable across borders and sectors[8]. GeoDCAT-AP is an extension of DCAT-AP for describing geospatial datasets, dataset series, and services. It provides an RDF syntax binding for the union of metadata elements of the core profile of ISO 19115:2003 and those defined in the framework of the INSPIRE Directive. Note that the V1.0.1 specification of GeoDCAT-AP[9] contains a number of Annexes that help in relating the various standards related to geospatial metadata interoperability. For example, Annex III compares the ISO 19115:2003 core requirements with the INSPIRE metadata requirements and the discovery metadata defined in the more recent ISO 19115-1:2014.

3.4 CKAN

Original Design

The Comprehensive Knowledge Archive Network (CKAN) is a web-based open source management system for the storage and distribution of open data. Being initially inspired by the package management capabilities of Linux, CKAN has developed into a powerful data catalogue system that is mainly used by public institutions seeking to share their data with the general public [10]. CKAN[11] is a fully-featured, mature, open source data management solution. CKAN provides a streamlined way to make data discoverable and presentable. Each dataset is given its own page with a rich collection of metadata, making it a valuable and easily searchable resource. CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available.

The central entity in CKAN is the Dataset which is associated to Resources (actual data: files, APIs etc.) and to metadata. CKAN supports a variety of metadata which are revisioned – i.e. all changes are recorded. Datasets have a set of “core” metadata attributes but they can also have an unlimited amount of arbitrary additional metadata in the form of “extra” key/value associations or by adding tags belonging to Vocabularies[12]. Datasets are linked to each other via relationships between datasets (depends on, child of, derived from etc) and they can be grouped.

[8] https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11
[9] https://joinup.ec.europa.eu/catalogue/distribution/geodcat-ap-v101-pdf
[10] https://en.wikipedia.org/wiki/CKAN
[11] https://ckan.org/
[12] http://docs.ckan.org/en/latest/maintaining/tag-vocabularies.html

Typical Usage

CKAN's codebase is maintained by Open Knowledge International[13]. The CKAN data management platform is in use by numerous governments, organisations and communities around the world. Being open source, it is difficult to know all the instances around the world. An extensive list of instances is presented on the official site of CKAN[14]. In Europe, there are 71 instances including government data catalogues for Austria, Greece, Italy, the Netherlands, Romania, Slovakia, UK and many more. Additionally, several cities (Amsterdam, Berlin, Copenhagen, Florence, Thessaloniki, etc.) use CKAN catalogues for their data.

Apart from governmental bodies, CKAN has been used by projects such as:

- DART: Detection of Archaeological Residues using Remote-sensing techniques
- NGDS: a catalog of documents and datasets that provide information about geothermal resources
- NOAA: National Oceanic and Atmospheric Administration (United States)

Potential Usage within a VRE

Because of CKANs popularity, having the potential to discover relevant research data sets published on CKAN systems, to import such data sets and to publish new data sets on CKAN could be desirable in specific research domains. This would necessitate a schema-level metadata mapping into and from CKANs data model along with the tools to convert instance data, not unlike the DCAT case sketched above. Note that direct mappings between DCAT and CKAN metadata have already been investigated my Neumaier et al[15].

3.5 OIL-E

Original Design

Open Information Linking for Environmental science research infrastructures (OIL-E)[16] is a developing framework for addressing the semantic linking requirements of environmental science e-RIs. Specifically, it aims at providing a machine-readable bridge between the ENVRI Reference Model[17] and other concept models related to research infrastructure, architecture and scientific (meta)data.

[13] https://okfn.org/
[14] https://ckan.org/instances/
[15] https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_16
[16] http://oil-e.net
[17] http://envri.eu/rm

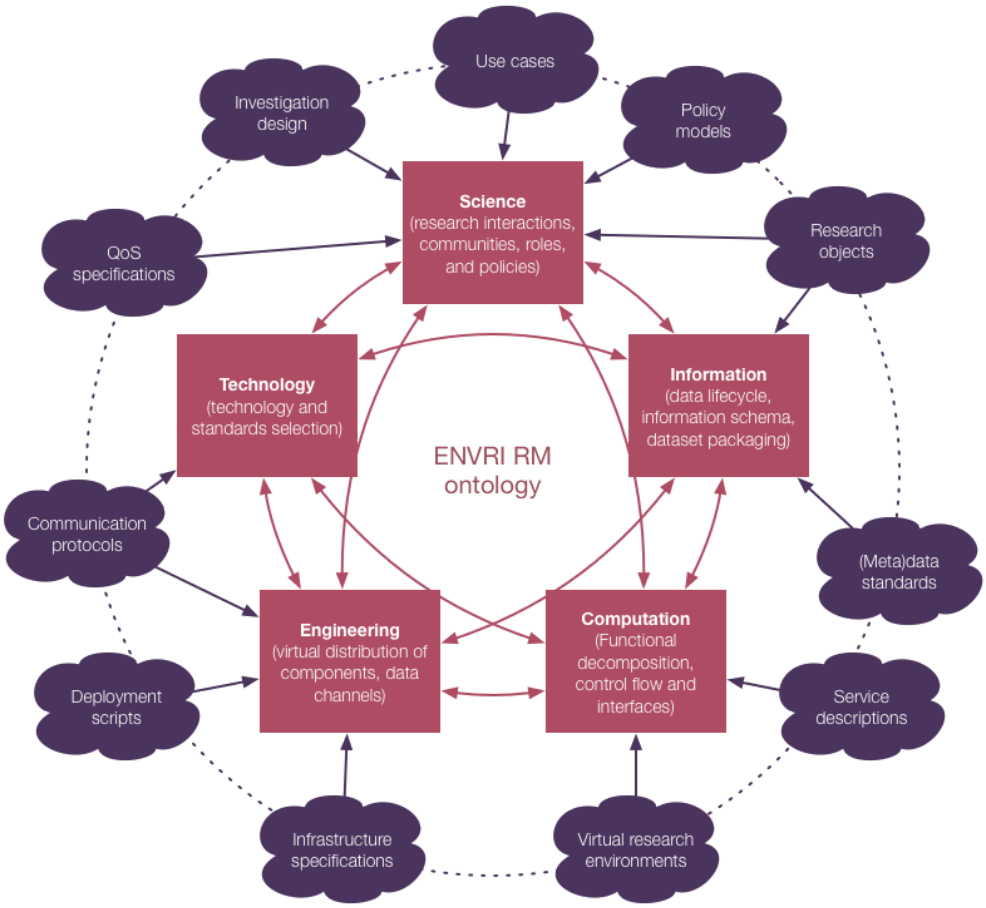


Figure: OIL-E is intended to act as a hub for establishing associations between different e-RI-related concept models.

The ENVRI Reference Model (E-RM) is constructed using the Open Distributed Process (ODP) for modelling complex distributed systems; ODP requires the modelling of a system from five different viewpoints (enterprise, information, computation, engineering and technology) with the correspondences between the five resulting views ensuring their mutual validity. This viewpoint-based approach provides clarity to each 'facet' of the end model by reducing the number of competing elements to only those that match a particular set of concerns (such as the flow of information through the system), while still retaining the aggregate complexity needed to model any substantive distributed system. At present, E-RM models three of the five views prescribed by ODP: *enterprise* (renamed *science* in respect to the subject area), *information* and *computation*. These three viewpoints best capture the generic aspects across all e-RIs, with the engineering and technology viewpoints being more e-RI-specific (though there is a plan now within ENVRIplus to generate these views). The E-RM ontology within OIL-E defines all the objects defined in the three existing views and their relations. It is intended that OIL-E will link concepts used in a variety of different standards and specifications to E-RM as a means to harmonise technical developments in RIs; a number of small pilots were carried out in the original ENVRI project to gauge the feasibility of this process.

Typical Usage

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

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

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

Although the core of OIL-E (the ENVRI reference model ontology) has been published and is subject to iterative review based on updates to the reference model it encodes, the full framework is still immature, making its full potential impact on the research data landscape unclear.

In terms of future development, OIL-E is being updated within the context of the ENVRIplus project, which will continue until early 2019. Within the ENVRIplus project, OIL-E is expected to be used to help with:

- The formal specification of a number of active e-RIs in using E-RM.
- The construction of a landscape of e-RI architecture and metadata models (and their linkage with cross-disciplinary initiatives such as GEOSS).
- The linking of data models used to translate research investigation requirements into constraints on e-infrastructure provisioning and configuration.

Potential Usage within a VRE

The final bullet point in the previous section may have some impact on the integration of an e-VRE with e-RI services, as the e-VRE essentially provides an interface for forwarding computational research investigation requirements to services hosted on (virtual) e-infrastructures.

The E-RM ontology itself is primarily useful for modelling e-RI architecture in terms of primary actors, information objects and computational processes. Thus in connection to a research relation meta-schema such as used by CERIF, its primary immediate use is as a taxonomy of classifiers for different kinds of relationships, particularly as they pertain to resources and facilities. Conversely, when viewed in relation to a resource schema such as Dublin Core, E-RM identifies a wide range of objects that can be specified in terms of such schemas in order to generate the basic metadata needed for cataloguing.

3.6 CERIF

Original Design

CERIF is the Common European Research Information Format, an international standard data model for research information. CERIF is an official EU recommendation to Member States[18].

It is developed and curated by euroCRIS[19], a not-for-profit association that brings together experts on research information in general and research information systems (CRIS) in particular. The organisation has 200+ members, mainly coming from Europe, but also from some

[18] http://cordis.europa.eu/cerif/
[19] http://eurocris.org/

countries outside of Europe. The various upgrades and extensions of the model are led by the CERIF Task Group. Current Version is CERIF 1.6.

CERIF deals with the different concepts involved in research information:

- Basic concepts: persons (researchers, etc.), projects, organisation units (teams, etc.)
- Concepts linked to results: products, patents, publications
- Concepts linked to research infrastructure: services, facilities, equipments
- Indicators and measurements
- And several other additional concepts: funding, addresses, geographic bindings, languages, etc.

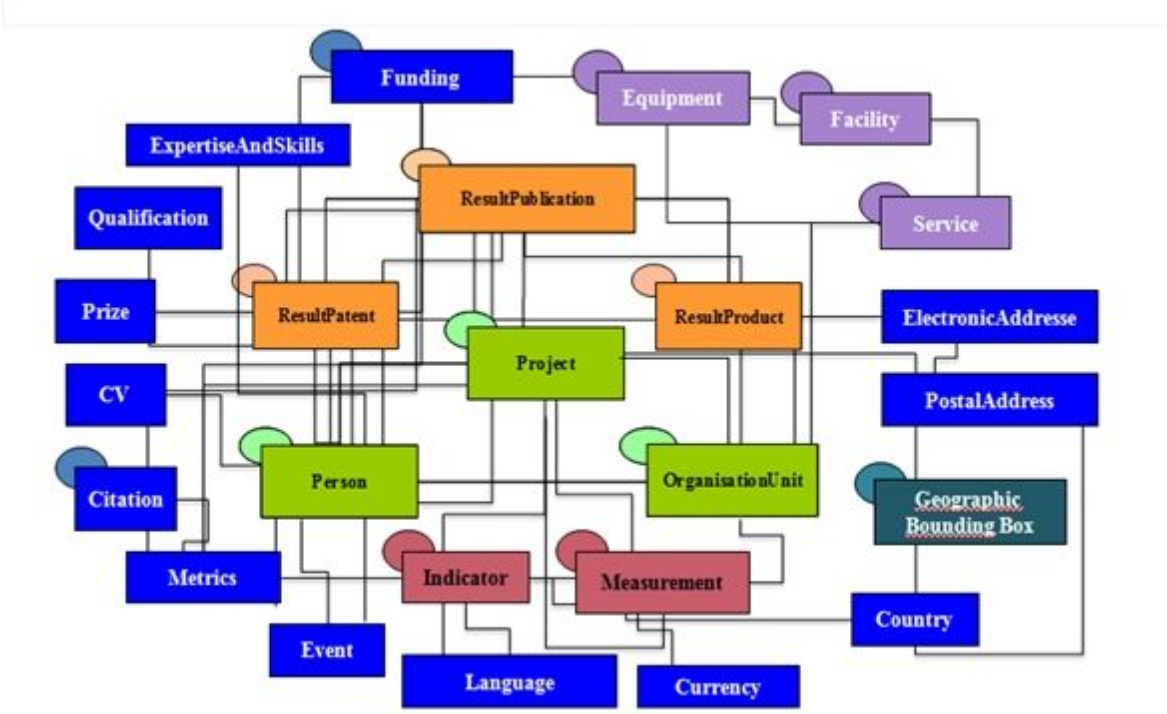


Figure: Key concepts in the CERIF 1.6 model

CERIF integrates several mechanisms that enrich the model: multilingual capacity, time range validity for relations, federated identifiers and a semantic layer. The multilingual capacity allows to deal with translations for most of the textual attributes of the entities, like names, titles, descriptions, etc. The time range validity is a mechanism that allows to deal with changes over time in the relationships that link the entities. CERIF is a model that also integrates a federated identifiers management for all entities. And last but not least, CERIF integrates a semantic layer that allows to manage classifications for entities and roles for relationships between entities. This last mechanism also allows to deal with synonyms and whatever kind of links between terms of the semantic layer.

In addition to the data model, CERIF also defines an XML exchange format and specification for a CERIF API[20]. The objective of this API is to facilitate the interoperability of systems and their integration with other information systems.

Typical Usage

CERIF is typically used as a data model to promote interoperability among Current Research Information Systems (CRIS). CERIF is a data-model that is freely available, thus it is hard to know at

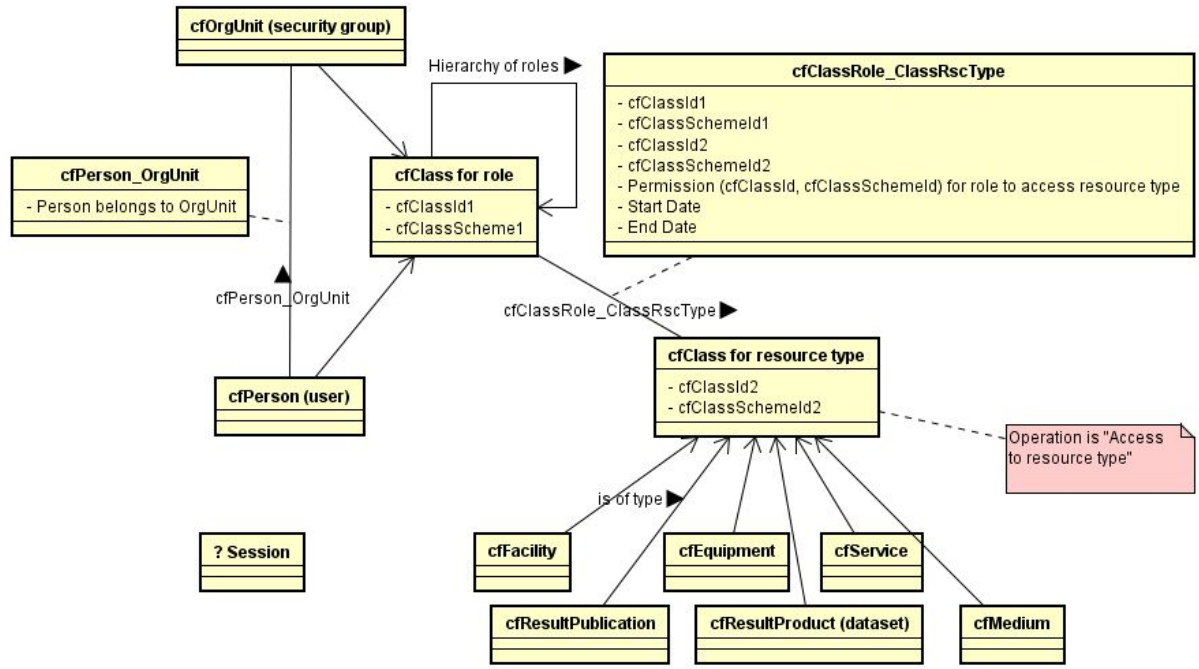
[20] http://eurocris.org/cerif-api-v10

which point it is used. The main stakeholders on the CRIS market use CERIF or are CERIF compliant. Products include Pure, Converis, Symplectic for commercial products, and DSpace-CRIS as an open-source product.

Potential Usage within a VRE

Several projects use CERIF as a data-model to store research data. MERIL[21], HOLA CLOUD[22] or ENGAGE[23] are examples of such projects. Another one is EPOS[24], which uses CERIF as its metadata catalogue. This last one is an example of the use of CERIF in a VRE as a metadata standard for an integration platform. In such a case, the strength of CERIF is that it has been designed to handle all facets of research metadata. In this way, it is an expressive model that is able to handle most of the existing metadata standard or homemade metadata models. We see the main use of CERIF in a VRE as an highly expressive "lingua franca" to convert metadata in other models into.

In addition, CERIF mechanisms can also be used to express concepts that are not directly related to research. For example, the figure below shows how, by using the CERIF semantic layer, it is possible to represent an RBAC (Role Based Access Control) strategy to manage access to the different types of resources. More details on this can be found in chapter 7 of Deliverable D5.1 "A strategy for the VRE4EIC project to handle security, privacy and trust issues".



[21] https://portal.meril.eu/meril/
[22] https://www.holaportal.com/
[23] http://www.engagedata.eu/
[24] https://www.epos-ip.org/

3.7 Research data contextualisation: quality and validation metadata

In many VRE use cases, researchers using the VRE would reuse research data from other researchers, potentially with the goal to answer new research questions using other methodologies than those used in the context in which the research data sets have originally been constructed. In those cases, the researchers need to be able to assess if, and to what extent, the data can indeed be applied in this new context (See the sections on "Trust" in D5.1 for more detail).

Many metadata elements discussed in the models above are partly or fully designed to help in such assessments. These include metadata elements related to licensing, tracking provenance and versioning, etc.

In addition, data reuse may require the data to be persistent, available and citable for periods that are longer than the original research context can provide. For a VRE, this means it is sometimes necessary to import and export data to external trusted data repositories created especially with this longevity in mind. This might result in extra interoperability requirements to a wide range of external systems, including national and university repositories, international systems such as Zenodo, figshare, EUDAT CDI, etc.

In parallel we see a growing trend to improving data citation, which comes with a need to support a wide variety of persistent identifier schemes, not only for data and data sets, but also for authors, organisations, etc (DOI, ePIC, URI, social networking identifiers, ORCID). Reusing large numbers of third party data sets also requires automation and interoperability in the terms of use and licensing.

While most of the models discussed above have elements that help tracing the lineage or provenance of data, most of these are either informal and only human-interpretable, or limited in scope. More formal and interoperable tracking requires that all steps that modify data in a complex, multi-system pipeline are logged, along with information about which changes are made and by whom. Key here is that even if different steps use different metadata formats relevant to their own domain, provenance info needs to be available in an interoperable manner across all steps in the pipeline. Dedicated metadata formats, designed only for this task, such as PROV-O, are therefore gaining in popularity.

4 Metadata interoperability analysis

The metadata models discussed in the previous section differ across many dimensions, which makes interoperability between systems using these models non-trivial. The assumption that one day these problems will disappear because systems will evolve into the use of a single overarching and unified model is generally regarded as unrealistic, so some form of metadata brokering will be necessary.

Often, the problems that the models are designed to solve are different, which may lead to differences in scope between the models that are sufficiently large that full compatibility, even when translating one model into another, is not feasible. However, in practical cases there is often sufficient common ground in a set of shared assumptions that partial compatibility is possible and very useful. In section 4.1 we sketch the metadata elements that are shared and, while different models use different techniques to represent them, can often be mapped. In section 4.2 we sketch

the other side of the coin: metadata elements that are intentionally domain-specific, to the extent that interoperability with systems from other domains is of lesser priority.

4.1 Common and cross-domain metadata

Direct usage versus mapping of common elements

It is common practice in the RDF-based metadata models to use URIs to identify metadata elements, also on the schema level. This provides the opportunity for a model to define its own URI namespace with its own schema terms and definitions, but also to reuse terms from other models directly. This is the case, for example, with many of the general metadata elements in DCAT, which directly use terms from the Dublin Core vocabulary. This provides a direct level of interoperability for these DCAT elements when used by systems supporting Dublin Core, and it also simplifies mapping DCAT metadata to other standards for which mappings to Dublin Core have already been defined. A similar approach of direct reuse of metadata elements through shared URIs is taken by the CKAN schema. An alternative to this direct reuse of URIs for common elements is to mint new URIs for the common terms and provide a mapping to other standards. The downside here is that this does not provide the direct interoperability of shared URIs, but it provides more flexibility, as mappings can be changed over time without touching the instance metadata and mapping to multiple standards can be defined in a more symmetric fashion. If the mapping can be defined by standardized mechanisms such as `rdfs:subPropertyOf` statements, this allows standard-conforming systems to support querying of the metadata using terminology from both standards transparently, providing almost the same level of interoperability as the direct reuse of shared URIs.

Expressivity versus simplicity trade-off

Every data model that aims at improving interoperability by defining a shared set of common elements needs to trade off expressivity against simplicity. Different choices on this dimension often lead to incompatibilities when modeling “standard” solutions for similar problems, and this may lead to a deliberate decision to not reuse terms from other models. A pragmatic way to restore such incompatibilities is to allow user communities to benefit from the simplicity of a number of less expressive formats that can be later translated into a common, expressive, but often more complex, model. Or, alternatively, to allow such translations in multiple steps. The EPOS project, for example, has proposed such an approach to simplify the mapping to CERIF for users that are not familiar with the full complexity of CERIF. EPOS defines an intermediate data model that can be used by user communities associated with research infrastructures of different types and different levels of maturity. The model is used as a target model to map a variety of metadata formats into, done in such a way that the result can be relatively easily mapped into CERIF in a second step.

This intermediate model is referred to as the “EPOS Baseline” by Bailo et al[25] (see the EPOS project internal document[26] for details). We use it here as an indication of which metadata elements are both present in a wide variety of standardization efforts and found to be of practical importance by various user communities.

EPOS takes the minimum set of metadata elements required by ISO 19115 and INSPIRE, and extends these with some key elements that are required by the EPOS community (e.g. elements to describe Research Infrastructures and Equipment, in addition to Software, Model and Datasets as specific Research Product entities). Other common entities in the EPOS baseline that need metadata descriptions and are covered in most models include Persons, Organisations and Publications. Note

[25] http://hdl.handle.net/11366/537
[26] http://people.na.infn.it/~festa/EPOS/EPOSmetadatamodelreference.pdf

that many of the metadata elements describing persons and organisations are targeted on how to find and identify them, and then how to contact them. The elements describing infrastructure, services, datasets, etc. also target finding and identification, but then focus on elements describing how to use the entity, under what conditions and who is responsible in case of problems.

Spatial and temporal coverage

In addition to the common organisational and administrative metadata elements discussed above, there is also growing consensus on more technical metadata elements describing datasets, services, software and models. For many datasets, their geo-spatial and temporal extent is of key importance for discovery and identification purposes. There is only a limited number of schema-level elements associated with this requirement, which simplifies mappings and other interoperability issues. Unfortunately, there is still a wide variety of conventions in use to fill the associated slots. For example, GeoDCAT-AP lists several methods to specify the value of a geospatial extent. One of these methods is by specifying a bounding box geometry, for which there are again multiple alternatives to choose from (quoting from the GeoDCAT-AP specification):

- *a URI - e.g., by using the geo URI scheme [IETF-RFC-5870], or a geohash URI [GEOHASH, GEOHASH-36]*
- *a syntax encoding scheme - e.g., geohashes [GEOHASH, GEOHASH-36], WKT [ISO-19125], GML [GML], KML [KML], GeoJSON [GEOJSON]*
- *a semantic representation - using vocabularies like W3C Lat/long [LAT-LONG] or schema.org [SCHEMA]*

Clearly, providing full interoperability across research infrastructures will need additional mappings services for instance data in case infrastructures do not use the same encoding schemes for spatial extent values, as schema-level mappings alone will then prove to be insufficient. For temporal extents the situation is less complex, with multiple models using begin/end data stamps using standardized date encodings (e.g. gml:TimePeriod in OGC's Geography Markup Language).

4.2 Domain-specific metadata

In addition to the common elements discussed above, many VREs deploy metadata elements that are more domain-specific, or elements of which the usage in practice is currently limited to a few domains.

Event-centric metadata models

For example, in D3.2 "Gap Analysis" both ARIADNE[27] and Parthenos[28] report using their own domain-specific metadata formats based on CIDOC-CRM[29]. One of the key features of CIDOC-CRM is that it is promoting interoperability through an event-centric modeling point of view, that is, a view where the significance of common entities such as the Persons, Organisations, Projects, Products is primarily defined by the roles they play in the real-world events described in the data. While this modeling stance is currently primarily used in a variety of museum documentation systems, and in VREs on archeology (ARIADNE) and heritage (Parthenos), it has the potential to be applied to many other domains.

[27] http://portal.ariadne-infrastructure.eu
[28] http://www.parthenos-project.eu/
[29] http://cidoc-crm.org/

Classification metadata

Many metadata standards aim at standardizing the use of classifications and classification schemes, without committing to a specific domain-specific terminology or classification scheme (e.g. see the description of the "semantic layer" in the section on CERIF above). W3C's SKOS Recommendation[30] is often used for this purposes, for example by DCAT and ARIADNE.

Tabular and statistical datasets

Many tabular and statistical datasets are available online for research purposes, many of them published in comma or tab separated text formats. Metadata about these datasets and metadata interoperability seems less mature. BlueBRIDGE[31] reports the use of Statistical Data and Metadata eXchange (SDMX[32], or ISO 17369:2013[33]). While often used in combination with an XML serialisation syntax, in the linked data world the RDF Data Cube Vocabulary[34] is often used for publishing this type of (meta)data. Note that Data Cube is a W3C Recommendation and compatible with the SDMX information model.

Workflow-oriented metadata

While many of the VREs discussed in D3.2 report the intention of supporting scientific workflow execution services, metadata-level interoperability and interoperability of the workflow specifications themselves seems still to be in a very early phase. Standardisation efforts such as the Common Workflow Language[35] are currently addressing this issue, but standardisation on the VRE level seems premature at the time of writing.

Provenance-oriented metadata

In many domains, and especially in those where data is the result of multi step pipeline or workflow, explicitly recording which processing steps have been applied with which parameters is often crucial. This type of metadata is often required, both to be able to assess the applicability of data in new contexts, and to allow for reproducibility of the dataset when needed. While this requirement is widely recognized, and many of the standards discussed above indeed include metadata elements to define the provenance or lineage of scientific datasets, in practice the availability of provenance data is still lacking for many datasets. More widespread adoption of scientific workflow systems might improve this situation in the future. For example, the Apache Taverna[36] system supports generating provenance metadata[37] for all its computations using the W3C PROV Ontology[38] Recommendation.

[30] https://www.w3.org/TR/skos-reference/
[31] http://www.bluebridge-vres.eu/
[32] https://sdmx.org/
[33] https://www.iso.org/standard/52500.html
[34] https://www.w3.org/TR/vocab-data-cube/
[35] http://www.commonwl.org/
[36] https://taverna.incubator.apache.org/
[37] https://taverna.incubator.apache.org/documentation/provenance/
[38] https://www.w3.org/TR/prov-o/

5 Summary and conclusions

5.1 Summary

Based on the VRE characterizations given in deliverable D3.2, this document provides an overview of the key metadata standards currently used in the VRE landscape. We discussed Dublin Core, ISO9115/INSPIRE, DCAT, CKAN, OIL-E and CERIF in terms of their design, typical usage and potential for usage within a VRE system. We provided a metadata analysis, first by discussing metadata elements that are commonly found across domains and systems, and are modelled in ways that provide a good basis for interoperable services within a VRE system. Then we discussed elements that are more domain specific and/or require more standardization. These are often elements for which the same level of interoperability is much harder to achieve. Fortunately, because these elements are often specific to a limited number of domains, there is also lesser need for interoperability across wider range of other domains.

5.2 Impact on metadata matching and mapping

The metadata elements discussed in 4.1 are either truly domain-independent, including many of the organisational and administrative metadata elements, or are so common across domains and so important for discovery services that we envision that interoperability support for these elements needs to be available as a core service in the VRE catalogue.

The EPOS project has already gained some experience in how the research communities can effectively provide the required metadata needed to meet the key INSPIRE requirements, and on how to map the provided data into CERIF as a common data model in the back-end catalog. Since many of the common elements can be directly expressed in CERIF, the best approach seems to be to provide mappings from the other models into CERIF, and vice versa. The usage of CERIF as a *lingua franca* also prevents the need to support mappings among all mapping formats.