Ref. Ares(2017)4778568 - 01/10/2017

# VRE4EIC

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

**Deliverable D4.2**

**Matching and mapping VRE Elements to CERIF**

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: | 31 May 2017 Amended Grant Agreement through amendment n°AMD-676247-8 |
| **Document** | |
| Period covered: | M7-24 |
| Deliverable number: | D4.2 |
| Deliverable title | Matching and mapping VRE Elements to CERIF |
| Contractual Date of Delivery: | 30/09/2017 |
| Actual Date of Delivery: | 01/10/2017 |
| Editor (s): | Laurent Remy (euroCRIS) |
| Author (s): | Dragan Ivanovic (euroCRIS), Jacco van Ossenbruggen (CWI), Theodore Patkos (FORTH), Athina Kritsotaki (FORTH), Manuella Sbarra (INGV), Paul Martin (UvA) |
| Reviewer (s): | Maria Theodoridou (FORTH), Laura Hollink (CWI) |
| Participant(s): | **euroCRIS**, CWI, FORTH, 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): | 104 |

# 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 virtualising 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 ..... 5**
**2 Methodology for matching and mapping ..... 6**
**3 Common elements ..... 8**
3.1 Identification of the common entities in the standards ..... 8
3.2 Summary of the common entities ..... 11
3.3 Focus on the dataset entity ..... 12
3.4 Notes ..... 14
3.4.1 OIL-E ..... 14
3.4.2 RDA (Research Data Alliance) ..... 14
**4 Vocabularies ..... 16**
**5 Identified metadata standards matching ..... 20**
5.1 Dublin Core ..... 20
5.1.1 Introduction ..... 20
5.1.2 Matching ..... 21
5.2 ISO19115 / INSPIRE ..... 24
5.2.1 Introduction ..... 24
5.2.2 Matching ..... 25
5.3 DCAT ..... 26
5.3.1 Introduction ..... 26
5.3.2 Matching ..... 27
5.4 CKAN ..... 29
5.4.1 Introduction ..... 29
5.4.2 Matching ..... 30
5.5 OIL-E ..... 33
5.5.1 Introduction ..... 33
5.5.2 Matching ..... 34
**6 Tools ..... 38**
6.1 CERIF RDF ..... 38
6.1.1 Basic Principles ..... 40
6.1.2 Current version CERIF 1.6 ..... 41
6.2 X3ML improvements ..... 42
**7 Conclusions ..... 44**
**8 References ..... 45**
**9 Annexes ..... 46**
9.1 Annexe 1: Dublin Core matching ..... 46
9.2 Annexe 2: ISO19115/INSPIRE mapping ..... 76
9.3 Annexe 3: DCAT-AP matching ..... 77
9.4 Annexe 4: CKAN matching and mapping ..... 88

# 1 Introduction

The goal of the VRE4EIC project is to allow users to search for data coming from various scientific datasets or repositories, which use heterogeneous data models. To ease the querying and the display of the results, a homogenization process is needed. This process will allow heterogeneous data (or metadata for this project) to be stored using CERIF as the common data model.

An analysis has been done previously regarding the standards being used and the metadata interoperability in *D4.1 - Review of existing VRE Metadata*. *D4.2* describes the result of task *T4.2* that takes the results of *D4.1* to define the mapping of these standards to the CERIF target schema.

In this document, we will first present the methodology used for the matching and mapping, also defining these two processes. We will then focus on the identified common elements, and the various vocabularies used or recommended by the standards, and that will be used in the semantic layer of CERIF. This will lead us to the definition of the matching of the standards to CERIF. We will end with a description of the tools that were built or tested during this phase of the project.

# 2 Methodology for matching and mapping

Before digging into the methodology used for this project, we need to define the various concepts we will be dealing with in this chapter.

First, the concept of **data model**, or **schema**, is an abstract model that represents different entities, the relations between them, and the properties to describe them. The schema also contains information about the format of the properties: whether it is a number, a date or a character string, etc. Since different groups of people can have different views of the same real world entities, they can have different ways to represent them, leading to different schemas representing the same concepts. This is where matching and mapping are needed.

**Matching** has been defined as the process of finding relationships or correspondences between entities of different ontologies (Euzenat & Shvaiko, 2007). Another definition says that schema matching aims at identifying semantic correspondences between elements of two schemas, e.g., database schemas, ontologies, and XML message formats (Do Hong Hai, 2005).

**Mapping** is the process of defining some transformation for the data to be compatible with the definition of the properties in the target data model.

In the context of this project, the core entities which need to be managed by the system have already been defined, and the target data model (CERIF) has already been chosen, as being a data model that is generic enough to handle most of the divergences to different domain semantics. Thus, the matching is constrained by the chosen target data model, which means that some specific properties may not be matched, as the correspondence does not exist in CERIF. The work of the data model expert is then to find the most exhaustive matching, to allow the integration of the most relevant information from the source data model.

The matching to CERIF is usually done in the following steps:

1. First we start with a study of the source schema: identifying the entities and the properties of these entities.
    - When integrating several source standards in one destination standard, it is a good idea to check the commonalities of the entities and properties of the various standards. If the standards contain too few elements in common, it will be hard to match the schemas so that the final user can easily query data from several sources.
2. Secondly, we identify the different vocabularies used in the source schema: either external controlled vocabularies (to define finished lists of terms for properties) or some specific terms used to describe the entities.
    - CERIF is semantically agnostic, which means that CERIF does not describe any domain-specific semantic for entities and relations between them. Instead, CERIF has a mechanism called a semantic layer that allows to store any vocabularies and use them as semantic for the entities and their relationships. The vocabularies and terms retrieved will be used in this semantic layer.
3. Then we try to match the source entity concepts to one of the entities in CERIF, not taking into account the properties, only the definition (i.e. the semantic) of the entities.
    - Indeed, we want to focus first on *the semantic*, not on the fact that the entities have the same number of properties of the correct format; properties are taken into account in the next step.
4. We finally match the properties, as exhaustively as possible, either by using the properties of the selected CERIF entity from the previous point, or by using additional CERIF entities and relationships with the selected CERIF entity.

In these last two steps, we define the semantics to be used to characterize the entities and the relationships between them.

The mapping process is facilitated by the X3ML toolkit developed by FORTH. Within this toolkit, the 3MEditor guides the user to specify the schema matching and the instance generators (the functions that will create the appropriate CERIF URIs). This tool also allows for some normalisation of the source data when required e.g. dates. Another tool is the X3ML engine that automatically transforms the source data into CERIF data instances applying the X3ML mapping definition file issued by the 3MEditor.

# 3 Common elements

To ease the study of the common elements, we take as a basis the work done by the EPOS metadata working group. To allow the integration of data coming from more than 240 research infrastructures in 25 countries, they developed a baseline composed of a minimal set of metadata elements[1]. These elements will be used as a basis to look for common elements in the chosen standards.

## 3.1 Identification of the common entities in the standards

The minimal set of metadata elements defined by EPOS is the following:

- Research Infrastructure
- Equipment
- Software
- Dataset
- Person
- Organisation
- Webservice
- Service

The purpose of the present section is to review the elements of the actual metadata standards, look for possible commonalities, and discuss the gaps.

The Publication entity has been added to the EPOS list, because it was well represented in the various standards, and the concept was relevant regarding the research context. Thus the minimal set of metadata elements (or entities) defined by EPOS has been refined and validated against the actual metadata standard, and is reported here:

- Research Infrastructure
- Equipment
- Software
- Dataset
- Person
- Organisation
- Webservice
- Service
- Publication

*Dublin Core*

Dublin Core does not define any entity in its schema, but we can derive some entities using the type property, and the recommendation to use a controlled vocabulary for values, such as the DCMI Type Vocabulary[2].

- Research Infrastructure: no mapping provided
- Equipment: no mapping provided
- Software: may be mapped using the Software class from the DCMI Type defined like ‘a computer program in source or compiled form’
- Dataset: is mapped using the Dataset class from the DCMI Type defined like ‘data encoded in a defined structure’

[1] http://dspacecris.eurocris.org/handle/11366/537
[2] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#type

- Person and Organisation: are mapped to the same dct:Agent class defined like 'a resource that acts or has the power to act'
- Webservice: is mapped using the Service class from the DCMI Type defined like 'a system that provides one or more functions'
- Service: is mapped using the Service class from the DCMI Type
- Publication: may be mapped using the Text class from the DCMI Type defined like 'A resource consisting primarily of words for reading'

*ISO19115 / INSPIRE*

The INSPIRE Metadata Technical Guidelines include rules for the description of resources based on ISO 19115 (datasets and series) and ISO 19119 (services). The main scope of those standards are environmental themes, covering also areas having cross-sector relevance.

In the above standards, the main metadata elements are

- Spatial datasets and data sets series
- Spatial Data Services

INSPIRE covers Dataset and Service entities; properties describe all attributes of the entity and its metadata.

- Research Infrastructure: no mapping provided
- Equipment: no mapping provided
- Software: no mapping provided
- Dataset: is a core concept in Inspire (scope code = dataset)
- Person: some properties contain information about the responsible party
- Organisation: some properties contain information about the responsible organisation
- Webservice: no mapping provided
- Service: is a core concept in Inspire (scope code = service)
- Publication: no mapping provided

*DCAT*

DCAT Application Profile uses terms from various existing specifications. Classes and properties within the standard have been taken from different namespaces. For this reason DCAT-AP is more expressive than other standards and more entities are covered, as shown below:

- Research Infrastructure: no mapping provided
- Equipment: no mapping provided
- Software: no mapping provided
- Dataset: is one of the mandatory classes in DCAT and is defined like a conceptual entity that represents the information published.
- Person: is mapped in a mandatory class using foaf vocabulary. The class is foaf:Person subclass of foaf:Agent
- Organisation: is mapped in a mandatory class using foaf vocabulary. The class is foaf:Organization subclass of foaf:Agent
- Webservice: may be mapped to the dcat:'Distribution' class. The class 'Distribution' is defined as a physical embodiment of the Dataset in a particular format. It is 'Recommended' to allow for cases where a particular Dataset does not have a downloadable Distribution
- Service: no mapping provided
- Publication: may be mapped with the optional "Document" class of the foaf vocabulary

Note that an extension of DCAT-AP XML Schema (http://www.w3.org/ns/dcat#) to describe the EPOS Baseline is being developed by the EPOS IT-team defining EPOS classes and properties not covered by

DCAT and other vocabularies. Such extension covers all the entities not mapped at the moment, for instance Equipment, Software, and Facility.

*CKAN*

CKAN is a web-based open source management system for the storage and distribution of open data. Dataset is the central domain element, as shown below:

- Research Infrastructure: no mapping provided
- Equipment: no mapping provided
- Software: may be mapped using Resource and an attribute resource_type = code
- Dataset: the Dataset entity is the object representing datasets
- Person: some properties contain information about the author (dc:creator) as the original creator of the dataset
- Organisation: the Organization entity is the object representing the organisation which is owner of the dataset
- WebService: may be mapped using Resource. This entity corresponds to a file, API or other online data resource and it is associated to a Dataset
- Service: no mapping provided
- Publication: may be mapped using Resource and an attribute resource_type = documentation

*CERIF*

CERIF is a 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 and supports knowledge transfer to decision makers for research evaluation, research managers, strategists, researchers, editors and the general public. All entities from the EPOS list can be mapped to the CERIF model, as shown below:

- Research Infrastructure: may be mapped using the cfFacility entity
- Equipement: may be mapped using the cfEquipment entity
- Software: may be mapped using the cfResultProduct entity with specific semantic
- Dataset: may be mapped using the cfResultProduct entity with specific semantic
- Person: may be mapped using the cfPerson entity
- Organisation: may be mapped using the cfOrganisationUnit entity
- WebService: may be mapped using the cfService entity with specific semantic
- Service: may be mapped using the cfService entity
- Publication: may be mapped using the cfResultPublication entity

## 3.2 Summary of the common entities

The table below shows the summary of the mapping of entities in the standard described above.

| Entity | Dublin Core | ISO19115 / INSPIRE | DCAT | CKAN | CERIF |
|---|---|---|---|---|---|
| *Research Infrastructure* | -- | -- | -- | -- | cfFacility |
| *Equipment* | -- | -- | -- | -- | cfEquipment |
| *Software* | dct:type=Software (DCMI Type) | -- | -- | Resource (resource_type= code) | cfResultProduct (with specific semantic) |
| *Dataset* | dct:type=Dataset (DCMI Type) | scope code=Dataset | dcat:Dataset | Dataset (Package) | cfResultProduct (with specific semantic) |
| *Person* | dct:Agent | (responsible party) | foaf:Person | creator | cfPerson |
| *Organisation* | dct:Agent | (responsible organisation) | foaf:Organization | organization (owner_org) | cfOrganisationUnit |
| *Webservice* | dct:type=Service (DCMI Type) | scope code=Service | dcat:Distribution | Resource | cfService (with specific semantic) |
| *Service* | dct:type=Service (DCMI Type) | scope code=Service | -- | -- | cfService |
| *Publications* | dct:type=Text (DCMI Type) | -- | foaf:Document | Resource (resource_type= documentation) | cfResultPublication |

Except for the Research Infrastructure and the Equipment entities, we can see that all entities are pretty well represented in all standards, which comfort us regarding the process of matching.

## 3.3 Focus on the dataset entity

*Dataset* is the class covered by every Standards, and a core concept for the VRE4EIC project. The table below shows the matching of the properties of a dataset. In this table, we selected only properties that were represented in more than one standard. Note that for the ISO19115 standard, we use the XML implementation ISO19139.

| Dataset-property | Dublin Core | INSPIRE metadata/ *ISO 19139 | DCAT | CKAN |
|---|---|---|---|---|
| *description* | dct:description | Resource abstract (M)<br>*Abstract describing the dataset (M) | dct:description | notes |
| *title* | dct:title (M) | Resource title (M)<br>*Resource title (M) | dct:title (M) | title |
| *contact point* | | Responsible party (M)<br>*Dataset responsible party (O) | dcat:contactPoint | maintainer_ email |
| *keyword/ tag* | | Keyword value (M)<br>*Descriptive Keywords (O) | dcat:keyword | tags: arbitrary textual tags for the dataset |
| *publisher* | dct:publisher | Responsible party (M)<br>*Dataset responsible party (O) | dct:publisher | maintainer: current maintainer or publisher of the dataset |
| *theme/ category* | | Topic Category | dcat:theme | |
| *access rights* | dct:accessRights | Limitations on public access(C)<br>*Resource Constraints (O) | dct:accessRights | |
| *conforms to* | dct:conformsTo | Conformity (M)<br>*not in ISO19139 | dct:conformsTo | |
| *documentation* | | Resource locator (C)<br>*On-line resource (O) | foaf:page | resources (resource_type= documentation) |
| *frequency* | dct:accrualPeriodicity | Maintenance information (R)<br>*Maintenance And Update Frequency (O) | dct:accrualPeriodicity | |
| *has version* | dct:hasVersion | | dct:hasVersion | version: dataset version |

| Dataset-property | Dublin Core | INSPIRE metadata/ *ISO 19139 | DCAT | CKAN |
|---|---|---|---|---|
| *identifier* | dct:identifier | Unique resource identifier (M)<br>*Identifier (O) | dct:identifier | id: unique id |
| *is version of* | dct:isVersionOf | | dct:isVersionOf | |
| *landing page* | | Resource locator (C)<br>*On-line resource (O) | dcat:landingPage | url |
| *language* | dct:language | Resource language (C)<br>*Language (O) | dct:language | |
| *provenance* | dct:provenance | Lineage(M)<br>*Lineage(O) | dct:provenance | |
| *related resource* | dct:relation | | dct:relation | resources: list of [[Domain Model/Resource\|Resources]] |
| *release date* | dct:issued | Date of publication(c)<br>*Dataset reference date (M) | dct:issued | extras [key="dataset-reference-date"] |
| *source* | dct:source | | | |
| *spatial/ geographical coverage* | dct:spatial | Geographic bounding box(M)<br>*Geographic location of the dataset(C) | | extras [key="spatial"] |
| *temporal coverage* | dct:temporal | Temporal extent(C)<br>*Additional extent information for the dataset | | |
| *type* | dct:type | Resource type(M)<br>*Hierarchy Level (O) | | type |
| *update/ modification date* | dct:modified | Date of last revision(C)<br>*Dataset reference date (M) | | metadata_modified |

We can see in this table that most of the basic properties describing a dataset are common in the various standards. These properties will be of importance, because they are properties that we can use to give a homogeneous result to the user, as we know they are represented almost everywhere.

## 3.4 Notes

### 3.4.1 OIL-E

OIL-E is one of the standards that have been listed as being relevant for the VRE4EIC project. OIL-E has not been integrated in this work on the common elements because its interest for this project is different from the other standards. OIL-E is a work in progress, and the first version of the matching reveals that the best way to integrate it into CERIF is to use it as part of the semantic layer. As we do not have data expressed in OIL-E yet, we cannot know whether this solution is a correct solution or the best one, or a solution that is not suitable. The dedicated part in chapter 5 “Identified metadata standards matching” will be more detailed about the matching of OIL-E to CERIF.

### 3.4.2 RDA (Research Data Alliance)

RDA is an international member-based organisation focusing on the development of infrastructure and community activities that reduce barriers to research data sharing and exchange across technologies, disciplines, and countries, and the acceleration of data driven innovation worldwide.

The RDA metadata working group recommends a set of metadata element for metadata interoperability and addresses issues related to coverage, ease of maintenance and sustainability.

The set of metadata elements is intended to specify the information that systems need in order to satisfy the use cases collected by the Data in Context Interest Group. The group had identified 17 high-level elements. These general elements also contain sub-elements that more closely resemble how the elements might be represented. All of these are still under a process of refinement.

The following high-level elements have been proposed and discussed:

- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Spatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- Provenance
- Citations
- Related publications (white or grey)
- Related software
- Schema
- Medium / format

This set of elements is still under development.

In the context of finding common elements, a first matching can be designed as follows:

- Research infrastructure: covered by Facility
- Equipment: covered by Equipment
- Software: covered by Related Software
- Dataset: covered by the whole set of elements, as the elements are used to describe a dataset
- Person: covered by Originator
- Organisation: covered by Originator

- Webservice: might be covered by Location (as a URL) or Schema (although schema element is questioned – not specified yet)
- Service: not covered
- Publication: covered by Related publications

The list of common entities seems also relevant regarding RDA. So we can presume that the matching of RDA elements to CERIF could be possible.

# 4 Vocabularies

Machine-processable metadata relies on the use of controlled vocabularies to restrict the range of values assigned to different metadata attributes. While certain metadata attributes are generally considered to be 'opaque', in the sense that the contained information only needs to be retrieved without any further examination (most notably 'human' annotations such as labels and comments), other attributes may need to be directly examined in order to direct the operation of e.g. catalogue services and e-VREs. For these attributes, the examining service needs to have some comprehension of the different possible values (for example for the different curation states of a scientific data file) and what they entail.

Most of the standards investigated within the context of VRE4EIC prescribe certain controlled vocabularies for certain selected metadata attributes—however it is also true that in many cases there is no single recommended vocabulary by which to construct the values assigned to some attributes, and so the community applying the standard is expected to agree upon a vocabulary suitable for their purposes. This is typically the case for many generic attributes in the Dublin Core standard, and also applies to CERIF, which relies on classification schemes to attach semantics to relations.

The purpose of this section is to identify those vocabularies that have been prescribed by certain standards (e.g. ISO 19115 for geospatial metadata) that therefore should either be adopted directly for any like-for-like metadata mapping, or need to be mapped themselves into the preferred vocabulary that will be used for the e-VRE system proposed by the project. The sets of such vocabularies already used in mappings described in this deliverable are listed in the table below. In each case, the vocabulary set (vocabulary) is identified as well as the standard that uses it. The table then specifies within the given standard which metadata attribute is involved (property), the type (range) assigned to the attribute (for which the vocabulary provides the range of permissible values), and the entity to which the attribute is given for (domain).

| Vocabulary | Standard | Domain | Property | Range |
|---|---|---|---|---|
| *CI_Presentation FormCode* | ISO 19139 | gmd:MD_Metadata | gmd:CI_Presentation FormCode | gmd:CI_Presentation FormCode |
| *CI_RoleCode* | ISO 19139 | gmd:contact, gmd:citedResponsibleParty, gmd:pointOfContact, gmd:distributorContact | gmd:CI_RoleCode | gmd:CI_RoleCode |
| *MD_CharacterSet Code* | ISO 19139 | gmd:MD_Metadata | gmd:MD_CharacterSet Code | gmd_MD_Character SetCode |
| *MD_Maintenance FrequencyCode* | ISO 19139 | gmd:MD_Metadata | gmd:MD_Maintenance FrequencyCode | gmd:MD_Maintenance FrequencyCode |
| *MD_ProgressCode* | ISO 19139 | gmd:MD_Metadata | gmd:MD_ProgressCode | gmd:MD_ProgressCode |
| *MD_ScopeCode* | ISO 19139 | gmd:MD_Metadata, gmd:dataQualityInfo | gmd:MD_ScopeCode | gmd:MD_ScopeCode |
| *MD_TopicCategory Code* | ISO 19139 | gmd:MD_Metadata | gmd:MD_TopicCategory Code | gmd:MD_TopicCategory Code |
| *ADMS (Asset Description Metadata Schema)* | DCAT-AP | Dataset, Catalogue Distribution (adms:Asset) | adms:identifier | adms:Identifier |

| Vocabulary | Standard | Domain | Property | Range |
|---|---|---|---|---|
| *ADMS* | DCAT-AP | ,, | adms:sample | dcat:Distribution |
| *ADMS scheme assettype/1.0* | DCAT-AP | ,, | dct:type | skos:Concept |
| *ADMS scheme status/1.0* | DCAT-AP | ,, | adms:status | skos:Concept |
| *ADMS* | DCAT-AP | ,, | adms:versionNotes | Literal |
| *ADMS scheme publishertype/1.0* | DCAT-AP | Agent | dct:type | skos:Concept |
| *EU Metadata Registry: File Type* | DCAT-AP | Distribution | dct:format | skos:Concept |
| *EU Metadata Registry: Corporate Body NAL* | DCAT-AP | Dataset, Catalogue | dct:publisher | skos:Concept<br>foaf:Agent |
| *EU Metadata Registry: Data Theme* | DCAT-AP | ,, | dcat:theme<br>dcat:themeTaxonomy | skos:Concept<br>skos:ConceptScheme |
| *EU Metadata Registry: Frequency* | DCAT-AP | ,, | dct:accrualPeriodicity | skos:Concept |
| *EU Metadata Registry: Language* | DCAT-AP | ,, | dct:language | skos:Concept |
| *EU Metadata Registry: Country, Place, Continent* | DCAT-AP | ,, | dct:spatial | skos:Concept |
| *Geonames* | DCAT-AP | ,, | dct:spatial | skos:Concept |

The standards identified above all serve different purposes, but can be summarised as follows. For the vocabularies used in ISO 19139 (INSPIRE) XML profiles, all are 'code lists' defined by in ISO 19115 (the geospatial data representation standard for which 19139 is the XML scheme).

- **CI_PresentationFormCode**: provides a vocabulary for different 'presentation forms' for data, including documents, images, models, maps, tables and audio, both in digital and hardcopy forms.
- **CI_RoleCode**: describes a function provided by a responsible party (e.g. author or publisher).
- **MD_CharacterSetCode**: identifies the character encoding being used in a resource (e.g. utf8 or usAscii).
- **MD_MaintenanceFrequencyCode**: denotes the frequency of data maintenance.
- **MD_ProgressCode**: provides the status of a dataset or progress of a review (e.g. completed or obsolete).
- **MD_ScopeCode**: classifies information regarding data (e.g. information applies to a dataset, to hardware, or to a service).
- **MD_TopicCategoryCode**: provides a (non-exhaustive) list of topics to which data can be associated.

- **ADMS**: Asset Description Metadata Schema[3] specifies controlled vocabularies used to describe EU digital assets. Properties used include adms:identifier (based on the UN/CEFACT Identifier complex type for non-literal identifiers), adms:sample (to link to a sample of a digital asset) and adms:versionNotes (to describe changes between current and previous versions of an asset). ADMS also provides small typing concept schemes for Asset and Publisher types.
- **MDR Corporate Body**: The EU Metadata Registry (MDR) provides a name authority list (NAL) listing European corporate bodies[4] that can be used as values for dct:publisher fields.
- **MDR Data Theme**: Also by the MDR: a Concept Scheme for dataset classification[5] to be used in values for dcat:theme fields.
- **MDR File Type**: MDR concept scheme for file types[6]
- **MDR Frequency**: MDR concept scheme for accrual frequencies[7] identifying how frequent (e.g. weekly, monthly) a dataset is updated.
- **MDR Language**: MDR concept scheme for official language definitions[8], to be used with the dct:language property.
- **MDR Country[9], Place[10] and Continent[11]**: MDR Concept Schemes/Named Authority Lists to be used to indicate geographic locations in values of dct:spatial properties.
- **Geonames[12]**: geographical database available under a CC-BY 3.0 license

The list above is only partial, in the sense that more vocabularies are expected to be identified over the course of the remainder of the project. In particular, the complete set of ISO 19115 code list dictionaries define a set of vocabularies that are very useful for characterising geospatial data and meeting the needs of the INSPIRE initiative, and more are likely to be imported in the VRE4EIC context. For the purposes of mapping into other frameworks (such as CERIF), these vocabularies provide a natural set of classification schemes that can be used to define the properties of various research concepts and relations. The main difficulty lies in the number of these vocabularies and determining which ones are key to the operation of an e-VRE and which ones overlap with other vocabularies from different sources. For example, we can group the vocabularies defined in ISO 19115 into three broad classes:

1. Vocabularies that are needed to precisely describe the content of geospatial data (e.g. MD_CellGeometryCode and MD_GeometricObjectTypeCode).
2. Vocabularies that define the format of geospatial data (the actual data artefact)—e.g. CI_PresentationFormCode and MD_CharacterSetCode.
3. Vocabularies that are used to define the context in which geospatial data is created and can be used (e.g. MD_KeywordTypeCode and MD_RestrictionCode).

Generally we can assume that the first class of vocabularies is intrinsic to data described using the ISO 19139 scheme, and can be preserved untouched when mapping the data into a different framework.

The second class of vocabularies may be replicated using an alternative vocabulary in another context, and so it must be determined whether there is one preferred vocabulary that can be used in mappings;

[3] https://www.w3.org/TR/vocab-adms/
[4] http://publications.europa.eu/mdr/authority/corporate-body/
[5] http://publications.europa.eu/mdr/authority/data-theme/
[6] http://publications.europa.eu/mdr/authority/file-type/
[7] http://publications.europa.eu/mdr/authority/frequency/
[8] http://publications.europa.eu/mdr/authority/language/
[9] http://publications.europa.eu/mdr/authority/country/
[10] http://publications.europa.eu/mdr/authority/place/
[11] http://publications.europa.eu/mdr/authority/continent/
[12] http://www.geonames.org

either the ISO 19115 recommendation, an alternative recommendation, or a new vocabulary that (preferably) subsumes all the vocabularies used by different information sources.

The third class of vocabularies is particularly relevant to the activity of an e-VRE, and is very likely to be directly referred to during its operation, necessitating the decision of whether to use the ISO 19115 recommendation or to subsume it with an alternative vocabulary.

# 5 Identified metadata standards matching

With common elements defined, the next step after the collection of controlled vocabularies is to define, for each standard, a precise matching for the properties of the entities. Depending on the source standard, some difficulties have been encountered for the matching, leading to some choices to get the best possible result.

## 5.1 Dublin Core

### 5.1.1 Introduction

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 simple resource description, to combining metadata vocabularies of different metadata standards, to providing interoperability for metadata vocabularies in the Linked Data cloud Semantic Web implementations.

Two characteristics of the Dublin Core elements are difficult to integrate in a matching to CERIF.

Firstly, in Dublin Core each element is optional and may be repeated as much as needed. For example, a resource can be described using Dublin Core with several titles, but CERIF (in its RDBMS expression) does not allow for several titles for most of its entities, except for the expression of the same title in different languages. For this mapping, we consider that Dublin Core source file only provides several titles for different languages of the same title.

Secondly, Dublin Core does not provide any top-level element, to encapsulate a resource description. For the mapping, we cannot rely on this information to know the type of the described resource. For this purpose, we can use the Dublin Core element “type”[13] that describes the nature of the resource. The recommended best practice is to use a controlled vocabulary to fill this element such as the DCMI Type Vocabulary[14]. For this matching, we consider that this element is mandatory, and filled with values coming from the DCMI Type Vocabulary.

[13] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#type
[14] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms#H7

### 5.1.2 Matching

The following table shows the DCMI Type Vocabulary terms, with the corresponding CERIF entity for each term. Using this table, we can now match the various Dublin Core elements with the CERIF properties linked to these main entities.

| DCMI Type Vocabulary term | DCMI Type Vocabulary description | Corresponding CERIF entity |
|---|---|---|
| *Collection* | An aggregation of resources. | The concept of group does not exist in CERIF and has not been further described; if the group contains elements of the same kind, one may use the same CERIF entity as the members of the group to represent the group (i.e. recursively); if the group contains elements of different entities, one may use the cfOrganisationUnit entity to represent the concept of a unit regrouping instances from different entities |
| *Dataset* | Data encoded in a defined structure. | cfResultProduct |
| *Event* | A non-persistent, time-based occurrence. | cfEvent |
| *Image* | A visual representation other than text. | cfMedium |
| *InteractiveResource* | A resource requiring interaction from the user to be understood, executed, or experienced. | cfService |
| *MovingImage* | A series of visual representations imparting an impression of motion when shown in succession. | cfMedium |
| *PhysicalObject* | An inanimate, three-dimensional object or substance. | cfResultProduct |
| *Service* | A system that provides one or more functions. | cfService |
| *Software* | A computer program in source or compiled form. | cfResultProduct |
| *Sound* | A resource primarily intended to be heard. | cfMedium |
| *StillImage* | A static visual representation. | cfMedium |
| *Text* | A resource consisting primarily of words for reading. | cfResultPublication<br>We consider research data only. |

*Simple Dublin Core to CERIF*

Simple Dublin Core[15] does not provide any range for the elements, thus providing users flexibility, allowing them to use any kind of structure as values for any element. This flexibility is a pain when talking about querying data, because we may not be able to deal properly with the data within the properties: it could be a simple text or numbered value, or it can be a complex entity, containing multiple structured values. For example, the dc:creator property can be filled with a basic rdfs:Literal containing the full name of the creator, or it can refer to a foaf:Person entity with sub-properties foaf:firstName and foaf:familyName, or a vcard:Individual with sub-property vcard:fn. The difficulty in this example is to retrieve, using one query, all the datasets with the name of their creators. Not knowing the structure of the properties or classes does not allow to provide any consistent mapping to another data format.

*Qualified Dublin Core to CERIF*

The qualified Dublin Core contains more information that further define the core 15 elements. This additional information allows us to have a better matching that could be automated, with some specific constraints.

The qualified Dublin Core[16] provides several additional Classes[17] that serve as ranges for most of the properties[18]. Some of these classes can be directly matched to CERIF whereas some others should be further constraint to allow a correct matching. The following table provides the various classes and their matching to CERIF entities.

| Dublin Core Class | CERIF entity | Comment |
|---|---|---|
| *Agent* | cfPerson<br>or cfOrganisationUnit<br>or cfService | An Agent should be described using foaf:Person for a person, foaf:Organization for an organisation. Any other case should be considered as a software. |
| *AgentClass* | cfClassificationScheme | This concept will be mapped in CERIF as a classification scheme where each classification term is a class of agent. |
| *BibliographicResource* | cfResultPublication | |
| *FileFormat* | | This concept will be mapped in CERIF using the cfMimeType field in the cfMedium entity. |
| *Frequency* | | This concept will be mapped in CERIF as a classification scheme where each classification term is a frequency. |
| *Jurisdiction* | | This concept will be mapped in CERIF as a classification scheme where each classification term is a jurisdiction. |
| *LicenseDocument* | cfResultPublication | We recommend to use foaf:Document as a structured content for the LicenseDocument. |
| *LinguisticSystem* | | This concept will be mapped in CERIF as a classification scheme where each classification term is a linguistic system. |

[15] http://dublincore.org/documents/dces/
[16] http://dublincore.org/documents/dcmi-terms/
[17] http://dublincore.org/documents/dcmi-terms/#H6
[18] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms#H2

| Dublin Core Class | CERIF entity | Comment |
|---|---|---|
| *Location* | cfClassification<br>or<br>cfGeographicBoundingBox | This concept will be mapped in CERIF as a classification scheme where each classification term is a location.<br>If the Location is described using coordinates, it can be stored in a cfGeographicBoundingBox instance. |
| *LocationPeriodOr Jurisdiction* | | This concept is a generalization of several concepts, and should not be used as is. The specialization should be used instead. |
| *MediaType* | | This concept will be mapped in CERIF as a classification scheme where each classification term is a file format. |
| *MediaTypeOrExtent* | | This concept is a generalization of several concepts, and should not be used as is. The specialization should be used instead. |
| *MethodOfAccrual* | | This concept will be mapped in CERIF as a classification scheme where each classification term is a method of accrual. |
| *MethodOfInstruction* | | This concept will be mapped in CERIF as a classification scheme where each classification term is a method of instruction. |
| *PeriodOfTime* | | This concept will be mapped to CERIF regarding the dcmi-period specification (http://www.dublincore.org/documents/dcmi-period/). Thus, the mapping is done using the X_Class entity where X is the original entity pointing to the period of time. The start date and end date are respectively mapped to cfStartDate and cfEndDate, and the scheme is mapped to the corresponding classification term. |
| *PhysicalMedium* | | This concept will be mapped in CERIF as a classification scheme where each classification term is a physical medium. |
| *PhysicalResource* | cfResultProduct | |
| *Policy* | cfResultPublication | The policy should be described as foaf:Document containing the description of the policy. |
| *ProvenanceStatement* | cfResultPublication | The provenance statement should be described as a foaf:Document containing the description of the provenance. |
| *RightStatement* | cfResultPublication | The rights should be described as a foaf:Document containing the description of the rights. |
| *SizeOrDuration* | cfMeasurement | |
| *Standard* | | This concept will be mapped in CERIF as a classification scheme where each classification term is a standard. |

The matching for the common elements identified in chapter 3 is described in annexe 1.

## 5.2 ISO19115 / INSPIRE

### 5.2.1 Introduction

The ISO19115 standard defines the schema required for describing geographic information. The schema provides information about the identification, the extent, the quality, the spatial and temporal data, spatial reference, and distribution of digital geographic data. The standard is applicable to cataloguing of datasets, especially geographic datasets, dataset series, and individual geographic features and feature properties. The ISO19115 standard can be used to provide the metadata about geographic datasets for INSPIRE[19].

The ISO standard format is complex and organized in sections and packages. A section is a grouping of similar information and a section may contain several packages. A package is a logical grouping of elements that can be found in multiple locations within the main sections. Packages are XML elements having a two-letter code followed by an underscore. The package abbreviations are shown in the following table.

| Package abbreviations | |
|---|---|
| CI | Citation |
| DQ | Data Quality |
| DS | Dataset |
| EX | Extent |
| FC | Feature Catalogue |
| GM | Geometry |
| LI | Lineage |
| MD | Metadata |
| RS | Reference System |
| SV | Services |
| QE | Data Quality Extended |

[19] http://inspire.ec.europa.eu/reports/ImplementingRules/metadata/MD_IR_and_ISO_20071210.pdf

### 5.2.2 Matching

The following table shows the main ISO 19115 sections and the CERIF elements with which these sections can be matched.

| ISO section | CERIF element | Comment |
|---|---|---|
| **Metadata** (*MD_Metadata*) - Root element that contains information about the metadata itself | cfResultProduct | MD_Metadata is the root ISO19115 format element which encloses all metadata about a dataset. Those metadata are matched with the CERIF cfResultProduct and linked entities cfPerson, cfPersonName, cfElectronicAddress, cfPostalAddress, cfMedium, cfOrganisationUnit, cfResultPublication, cfFederatedIdentifier, cfClassification, cfResultProduct, cfGeographicBoundingBox. |
| **Reference System Information** (*gmd:referenceSystemInfo*) - Information about the spatial and temporal reference systems used in the resource | cfResultProduct_ cfResultProduct | The reference system information is represented using a cfResultProduct CERIF entity instance which is linked to a cfResultProduct instance representing dataset metadata. The established link is classified as "Coordinate reference system" |
| **Identification Information** (*gmd:identificationInfo*) - Information required to uniquely identify a resource or resources. | cfResultProduct, cfPerson, cfOrganisationUnit, cfGeographicBoundingBox, cfClassification | The identification information includes the dataset name, abstract, authors, publishers, keywords, etc. |
| **Distribution Information** (*gmd:distributionInfo*) - Information about who makes a resource available and how to get it | cfMedium | The distribution information element represents a dataset resource and it contains information about the resources name, description, function, protocol for accessing resource, etc. It is matched with the CERIF cfMedium multilingual (cfMediumTitle, cfMediumDescription) and linked classification entities (cfClassification) |
| **Data Quality Information** (*gmd:dataQualityInfo*) - Information about the quality and lineage (including processing steps and sources) of a resource | cfResultPublication | The data quality information can be represented as a report with title, description, note, identifier and classification. It can be matched with the CERIF cfResultPublication multilingual entities (cfResultPublicationTitle, cfResultPublicationAbstract, cfResultPublicationBibliographicNote) and linked entities cfFederatedIdentifier and cfClassification. |

| ISO section | CERIF element | Comment |
|---|---|---|
| **Metadata Constraint Information** (*gmd:metadataConstraints*) - Information about constraints on the use of the metadata and the resource it describes | cfClassification | The metadata constraint information includes usage, access and other constraints which can be matched with classification of cfResultProduct element which represents dataset (cfResultProduct_Classification) |
| **Metadata Maintenance Information** (*gmd:metadataMaintenance*) - Information about maintenance of the metadata and the resource it describes | cfClassification | There are twelve possible approaches for maintenance frequency (MD_MaintenanceFrequencyCode): continual, daily, weekly, fortnightly, monthly, quarterly, biannually, annually, asNeeded, irregular, notPlanned, unknown. It is represented by the classification of the cfResultProduct element which represents dataset (cfResultProduct_Classification) |

The full matching and mapping between ISO19139 standard and CERIF format is expressed using the 3M tool, and part of this mapping can be found in annexe 2.

## 5.3 DCAT

### 5.3.1 Introduction

DCAT stands for Data CATalogue vocabulary. It is a recommendation by the W3C that is designed to facilitate interoperability between data catalogs published on the Web. The main entities managed by the DCAT vocabulary are Catalog, Dataset and Distribution. It relies on the FoaF vocabulary to describe persons and organisations.

A specification of DCAT for data portals in Europe has been designed: DCAT-AP[20]. 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 to use some properties for each class, whereas DCAT-AP defines sets of mandatory, recommended and optional properties for each class. As these constraints are helpful in terms of mapping the data, we use the DCAT-AP specification for the mapping.

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

### 5.3.2 Matching

The following table shows the DCAT-AP classes, with the corresponding CERIF entity for each term. Usage notes for the DCAT-AP classes are described in the official document[21].

| DCAT-AP class | URI | CERIF entity |
|---|---|---|
| *Agent* | foaf:Agent | cfOrganisationUnit<br>or<br>cfPerson<br>(an Agent should be described using foaf:Person for a person, foaf:Organization for an organisation, rather than foaf:Agent; foaf:Person and foaf:Organization are sub-classes of foaf:Agent) |
| *Catalogue* | dcat:Catalog | cfService |
| *Dataset* | dcat:Dataset | cfProduct |
| *Category* | skos:Concept | cfClassification |
| *Category scheme* | skos:ConceptScheme | cfClassificationScheme |
| *Distribution* | dcat:Distribution | cfMedium<br>Note: there is no consensus yet on the representation of temporal or spatial series of one Dataset in DCAT-AP; we choose to consider series as sub-datasets, being themselves Datasets; a Distribution then represent the expression of the dataset in one specific format. |
| *Licence document* | dct:LicenseDocument | cfClassification |
| *Catalogue record* | dcat:CatalogRecord | cfProduct |
| *Checksum* | spdx:Checksum | cfMeasurement |
| *Document* | foaf:Document | cfResultPublication<br>or<br>cfElectronicAddress |
| *Frequency* | dct:Frequency | cfClassification |
| *Identifier* | adms:Identifier | cfFederatedIdentifier |
| *Kind* | vcard:Kind | cfClassification |
| *Licence type* | skos:Concept | cfClassification |
| *Linguistic system* | dct:LinguisticSystem | cfClassification |
| *Location* | dct:Location | cfClassification<br>or<br>cfGeographicBoundingBox |
| *Media type or extent* | dct:MediaTypeOrExtent | cfClassification |

[21] https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11

| DCAT-AP class | URI | CERIF entity |
|---|---|---|
| *Period of time* | dct:PeriodOfTime | This concept will be matched with CERIF regarding the dcmi-period specification[22].<br>Thus, the matching is done using the X_Class entity where X is the original entity pointing to the period of time. The start date and end date are respectively matched with cfStartDate and cfEndDate, and the scheme is matched with the corresponding classification term. |
| *Publisher type* | skos:Concept | cfClassification |
| *Rights statement* | dct:RightsStatement | The rights statement should be described in the source as a foaf:Document containing the description of the rights. Rights statement can then be mapped to cfResultPublication. |
| *Standard* | dct:Standard | cfClassification |
| *Status* | skos:Concept | cfClassification |
| *Provenance statement* | dct:ProvenanceStatement | The provenance statement should be described in the source as a foaf:Document containing the description of the provenance. Provenance statement can then be mapped to cfResultPublication.<br>*Note: in CERIF, provenance is supposed to be represented as relations between entities, with temporal coverage. A "statement" is not suitable to be mapped as a rich relationship as it has no structured details.* |

Most of the DCAT-AP classes are external classes coming from the Simple Knowledge Organisation System schema [23] (skos prefix), Dublin Core (dct prefix), vCard ontology [24] (vcard prefix), Asset Description Metadata Schema [25] (adms prefix), Friend of a Friend vocabulary [26] (foaf prefix). The matching done for these classes can be easily reused for other standards based on these external vocabularies.

Some additional constraints have been described to allow the matching. For example, the concept of Agent cannot be matched to CERIF if described using a foaf:Agent object, as there is no concept in CERIF on the top of cfOrganisation and cfPerson. Thus, it is recommended to use either foaf:Person or foaf:Organization, so that the matching with CERIF can be done. For Rights or Provenance Statements, it is the same case as Dublin Core: the matching can only be done if the statement is contained within a document, and the document described using foaf:Document.

The matching described in annexe 3 of this document is focused on the common classes identified in chapter 3.

[22] http://www.dublincore.org/documents/dcmi-period/
[23] https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html
[24] https://www.w3.org/TR/vcard-rdf/
[25] https://www.w3.org/TR/vocab-adms/
[26] http://xmlns.com/foaf/spec/

## 5.4 CKAN

### 5.4.1 Introduction

CKAN is an open source and open architecture software platform for data management. 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.

Flexibility of the CKAN platform makes much more complicated the matching of the CKAN metadata model to the CERIF format. The set of metadata which describes a data file can be customised and extended. The vocabularies which contain the set of values available for some metadata can be also customised and extended.

There are three basic entities in the CKAN metadata model: Package, Resource, and Group. Moreover, there is the Organization CKAN entity which stores data about the institution being the data publisher. Besides these entities, CKAN also enables adding tags with available sets of values to a package and key-value pairs of additional package metadata – this is implemented in CKAN using "extras" elements which contain "key" and "value" attributes. Thus, any new metadata can be added to the CKAN instance in the form of key-value pair. For the sake of the VRE4EIC project matching of CKAN metadata with the CERIF format, we matched the 10 most used "extras" elements found in 133 CKAN instances all over the world[27]. Those 10 most used extras are shown in the following table.

| CKAN extras key | Number of CKAN instances |
|---|---|
| spatial | 33 |
| harvest_object_id | 30 |
| harvest_source_id | 29 |
| harvest_source_title | 29 |
| guid | 20 |
| contact-email | 17 |
| spatial-reference-system | 16 |
| metadata-date | 15 |
| resource-type | 15 |
| dataset-reference-date | 15 |

[27] Sebastian Neumaier and Jurgen Umbrich and Axel Polleres (2016). Challenges of mapping current CKAN metadata to DCAT. W3C Workshop on Data and Services Integration

### 5.4.2 Matching

The following table shows how previously mentioned CKAN entities can be matched with the CERIF elements.

| CKAN element | CERIF element | Note |
|---|---|---|
| *Package* | cfResultProduct | The Package CKAN element's attributes are matched with the cfResultProduct multilingual entities cfResultProductName, cfResultProductDescription, cfResultProductVersionInformation, as well as linked to the CERIF entities cfPerson, cfElectronicAddress, cfOrganisationUnit, cfClassification, and cfFederatedIdentifier |
| *Resource [resource_type = documentation]* | cfResultPublication | If resource_type has value "documentation", the Resource CKAN element's attributes are matched with the cfResultPublication multilingual entities cfResultPublicationTitle, cfResultPublicationAbstract, as well as linked to the CERIF semantic layer entity cfClassification and the linked entity cfFederatedIdentifier |
| *Resource [resource_type in (visualization, code)]* | cfResultProduct | If resource_type has value "visualization" or "code", the Resource CKAN element's attributes are matched with the cfResultProduct multilingual entities cfResultProductName, cfResultProductDescription, as well as linked to the CERIF semantic layer entity cfClassification and the linked entity cfFederatedIdentifier |
| *Resource [resource_type = api]* | cfService | If resource_type has value "api", the Resource CKAN element's attributes are matched with the cfService multilingual entities cfServiceName, cfServiceDescription, as well as linked to the CERIF semantic layer entity cfClassification and the linked entity cfFederatedIdentifier |
| *Resource [resource_type in (dataset, file, file.upload)]* | cfMedium | If resource_type has value "dataset", "file" or "file.upload", the Resource CKAN element's attributes are matched with the cfMedium multilingual entities cfMediumTitle, cfMediumDescription, as well as linked to the CERIF semantic layer entity cfClassification and the linked entity cfFederatedIdentifier |

| CKAN element | CERIF element | Note |
|---|---|---|
| *Group* | cfResultProduct | The Group CKAN element's attributes are matched with the cfResultProduct multilingual entities cfResultProductName, cfResultProductDescription, as well as linked to the CERIF semantic layer entity cfClassification and the linked entity cfFederatedIdentifier |
| *Tag* | cfClassification | The CKAN Tags are matched using the CERIF semantic layer and its cfClassification entity |
| *Extras [key = spatial]* | cfGeographicBoundingBox | If the Extras key attribute has value "spatial", the Extras value attribute is matched with the cfGeographicBoundingBox multilingual entity cfGeographicBoundingBoxDescr |
| *Extras [key = harvest_object_id]* | cfFederatedIdentifier | If the Extras key attribute has value "harvest_object_id", the Extras value attribute is matched with the cfFederatedIdentifier and URI contains the value of key attribute |
| *Extras [key = harvest_source_id]* | cfFederatedIdentifier | If the Extras key attribute has value "harvest_source_id", the Extras value attribute is matched with the cfFederatedIdentifier and URI contains the value of key attribute |
| *Extras [key = harvest_source_title]* | cfFederatedIdentifier | If the Extras key attribute has value "harvest_source_title", the Extras value attribute is matched with the cfFederatedIdentifier and URI contains the value of key attribute |
| *Extras [key = guid]* | cfFederatedIdentifier | If the Extras key attribute has value "guid", the Extras value attribute is matched with the cfFederatedIdentifier and URI contains the value of key attribute |
| *Extras [key = contact-email]* | cfPerson_ResultProduct, cfPerson, cfPerson_ ElectronicAddress, cfElectronicAddress | If the Extras key attribute has value "contact-email", the Extras value attribute is matched with the linked cfElectronicAddress entity and the established link is classified as "contact email" |
| *Extras [key = spatial-reference-system]* | cfGeographicBoundingBox _Classification | If the Extras key attribute has value "spatial-reference-system", the Extras value attribute is matched with the cfGeographicBoundingBox linked CERIF semantic layer entity cfClassification |

| CKAN element | CERIF element | Note |
|---|---|---|
| *Extras [key = metadata-date]* | cfResultProduct_ Classification.cfStartDate | If the Extras key attribute has value “metadata-date”, the Extras value attribute is matched with the linked cfClassification entity and its cfStartDate attribute |
| *Extras [key = resource-type]* | cfResultPublication_ Classification or cfResultProduct_ Classification or cfService_Classification or cfMedium_Classification | If the Extras key attribute has value “resource-type”, the Extras value attribute is matched with the linked cfClassification entity |
| *Extras [key = dataset-reference-date]* | cfResultProduct_ Classification.cfStartDate | If the Extras key attribute has value “dataset-reference-date”, the Extras value attribute is matched with the linked cfClassification entity and its cfStartDate attribute |
| *Organization* | cfOrganisationUnit | The Organization CKAN element’s attributes are matched with the cfOrganisationUnit multilingual entities cfOrganisationUnitName, cfOrganisationUnitResearchActivity, as well as linked to the CERIF entities cfMedium, cfClassification, and cfFederatedIdentifier |

The full matching and mapping between CKAN metadata model and CERIF format is expressed in 3M tool. The matching for the common elements identified in chapter 3 and part of mapping expressed using 3M tool can be found in annexe 4.

## 5.5 OIL-E

### 5.5.1 Introduction

Open Information Linking for Environmental science research infrastructures (OIL-E) 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 (ENVRI RM) and other concept models related to research infrastructure, architecture and scientific (meta)data. At present, the ENVRI RM models three viewpoints of research infrastructure: science, information and computation. These three viewpoints best capture the generic aspects across all e-RIs. The ENVRI 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 as a means to harmonise technical developments in research infrastructure, and an example of this is the examination of how best to map between the ENVRI RM and CERIF.

As a controlled vocabulary for describing environmental science research infrastructure, the most natural way to leverage the ENVRI reference model in the context of CERIF is to use it as a set of classification schemes for CERIF concepts and relations.

For many CERIF concepts/relations, a particular concept can be identified in OIL-E for which there are several 'archetypes' from the ENVRI RM modelled as sub-concepts within the OIL-E ontologies. These archetypes are drawn from http://www.oil-e.net/ontology/rm-archetypes.owl, with most schemes drawn http://www.oil-e.net/ontology/rm-core.owl (imported by rm-archetypes).

For the benefit of the reader, and as the IRIs for ontology concepts are not themselves required to carry semantic meaning out with their namespaces (i.e. not all concept names are consistently formatted or even necessarily humanly intelligible), all concepts are referred to by their preferred labels (in both CERIF RDF and OIL-E, this is their rdfs:label annotations).

### 5.5.2 Matching

The following concepts in OIL-E v2.1 were identified as possible classification schemes for CERIF concepts.

| CERIF concept | OIL-E concept | *Context in which classification is applicable* |
|---|---|---|
| | | **Examples of classifications from ENVRI-RM** |
| Event | behaviour | *An activity conducted by agents in an e-RI context.* |
| | | 'data collection [behaviour]', 'data replication [behaviour]' |
| Equipment | resource | *Equipment used in research interactions.* |
| | | 'sensor network', 'storage system' |
| Facility | resource | *Facilities used for research interactions.* |
| | | 'data repository', 'research infrastructure' |
| 'Organisation Unit' | actor | *Organisations acting as actors in a research interaction.* |
| | | 'data publisher', 'semantic mediator' |
| Person | actor | *People acting in a research interaction.* |
| | | 'environmental scientist', engineer |
| 'Result Entity' | 'persistent data' | *Results produced within an e-RI.* |
| | | 'QA-assessed data', 'annotated data' |
| Service | 'computational object' | *Computational services provided by e-RIs.* |
| | | 'catalogue service', 'data broker' |

It can be observed that not all suggested classification schemes fit as tightly to CERIF concepts as might be desired. For instance both Equipment and Facility in CERIF can be classified via archetypes subclassed from resource in OIL-E. Naturally, not all resources defined by the ENVRI RM fit the semantics for both of those two classes (though there is possible overlap depending on context). Although it is not generally appropriate to change an ontology based on the requirements of a specific matching between the ontology and another, this matching does provide some useful feedback into the next version of OIL-E in terms of what the most useful concept groupings might be.

Classifications can also be assigned to relations between concepts in CERIF (conceptualised by the 'Full Link Entity' concept in CERIF RDF). A number of ENVRI RM archetypes have been identified as possibly applicable for many such relations, as described below. Because of the great number of possible relations expressible in CERIF (essentially any class of concept to any other class of concept), we group the targets of the relations together for brevity where the same OIL-E concept is deemed applicable for all of them.

| CERIF relation | OIL-E concept | *Context in which classification is applicable* |
|---|---|---|
| | | **Examples of classifications from ENVRI-RM** |
| Event-[Equipment, Facility, Organisation Unit, Person, Result Entity, Service] | behaviour | *Activities involving e-RI equipment/facilities/organisations/people/results/services.* |
| | | ‘data collection [behaviour]’, ‘design experiment [behaviour]’ |
| Equipment-[Event, Measurement, Person, Result] | behaviour | *Activities involving e-RI equipment, particularly the generation of measurements.* |
| | | ‘data collection [behaviour]’, ‘instrument calibration [behaviour]’ |
| Equipment-[Measurement, Result Entity] | operation | *Operations to extract data from instruments e.g. sensors.* |
| | | ‘retrieve data [operation]’ |
| Equipment-Result Entity | information action | *The creation of new data from equipment.* |
| | | ‘perform measurement or observation [action]’ |
| Facility-[Event, Measurement, Result Entity] | behaviour | *Activities involving e-RI facilities such as the preservation of measurement data.* |
| | | ‘data preservation [behaviour]’, ‘data quality checking [behaviour]’ |
| Facility-Result Entity | information action | *Actions performed on data at e-RI facilities.* |
| | | ‘carry out backup [action]’, ‘publish data [action]’ |
| Federated Identifier-Service | information action | *Actions taken by making use of federated identifiers.* |
| | | ‘cite data [action]’ |
| Measurement-[Event, Equipment, Person, Project] | behaviour | *Activities involving measurement data collected by/within an e-RI.* |
| | | ‘data collection [behaviour]’, ‘data quality checking [behaviour]’ |
| Measurement-Result Entity | information action | *Actions taken to generate results from measurements.* |
| | | ‘add metadata [action]’, ‘process data [action]’ |
| Measurement-Equipment | operation | *Operations to extract data from instruments e.g. sensors.* |
| | | ‘retrieve data [operation]’ |
| Organisation Unit-[Event, Measurement, Result | behaviour | *Activities involving organisations, including research communities, within the context of a research infrastructure.* |

| CERIF relation | OIL-E concept | *Context in which classification is applicable* |
|---|---|---|
| | | **Examples of classifications from ENVRI-RM** |
| Entity, Service] | | ‘design measurement model [behaviour]’, ‘semantic harmonisation [behaviour]’ |
| Organisation Unit[Result Entity, Service] | information action | *Information actions taken to produce/query results or service specifications by organisational units.* |
| | | ‘query provenance [action]’, ‘describe service [action]’ |
| Person-[Event, Equipment, Facility, Measurement, Result, Service] | behaviour | *Activities involving individuals within the context of a research infrastructure.* |
| | | ‘design experiment [behaviour]’, ‘instrument calibration [behaviour]’ |
| Person-[Result Entity, Service] | information action | *Actions taken by individuals to produce/query results or service specifications.* |
| | | ‘annotate data [action]’, ‘describe service [action]’ |
| Person-Service | operation | *Operations invoked by individuals on online computational services.* |
| | | ‘query data [operation]’, ‘authorise action [operation]’ |
| Result Entity-[Event, Equipment, Facility, Person, Service] | behaviour | *Activities involving the use or generation of results.* |
| | | ‘coordinate service [operation]’, ‘data publication [operation]’ |
| Result Entity-Measurement | information action | *Actions taken to generate results from measurements.* |
| | | ‘add metadata [action]’, ‘process data [action]’ |
| Result Entity-Equipment | information action | *Operations to extract data from instruments e.g. sensors.* |
| | | ‘retrieve data [operation]’ |
| Service-[Event, Organisation Unit, Person, Result Entity] | operation | *Operations performed via online services provided by an e-RI.* |
| | | ‘prepare data transfer [operation]’, ‘update catalogue [operation]’ |
| Service-Result Entity | information action | *Actions taken via online services to generate/retrieve result data.* |
| | | ‘assign unique identifier [action]’, ‘convert data [action]’ |

There are broadly speaking three classes of ‘activity’ concepts represented at present in OIL-E, split across the three main ENVRI RM viewpoints of science (behaviours), information (information actions) and computation (operations). Relationships in CERIF can (in specific contexts as identified in the above table) be classified via one of these three activity classes depending on whether the desired semantics can be most comfortably placed at the level of agent behaviours, transformations (or creation) of data, or operations invoked by computational services. In this context the ENVRI RM provides a quite rich set of semantics for different kinds of activity or event, albeit not a complete set.

# 6 Tools

One of the goals of task 4.2 of the VRE4EIC project is to test the X3ML toolkit provided by FORTH to formulate mapping definitions.

The X3ML toolkit allows to define a mapping between a source schema expressed in XML and a target schema expressed in RDF, and transform the instances of the source schema to instances of the target. So the current version of this toolkit only allows transformation from XML to RDF. As there was no available RDF version of the current CERIF model, the first sub-task regarding the tools was to create an RDF version of CERIF. The first part of this section describes the methodology used to create the RDF version of CERIF from the relational model.

Using this new RDF version of CERIF, the X3ML toolkit has been tested in the context of the VRE4EIC project by implementing the mappings for some standards (like CKAN and ISO19139 for example) and running the tool to get transformed data. The second part describes the various improvements that have been implemented in the tool since the beginning of the project.

## 6.1 CERIF RDF

CERIF started as a relational model that has several exchange formats – it has many releases: from version 1.3 (fig.1) to 1.6. It started with CERIF 2006 release in XML format as a popular interchange format for data exchange operations at physical level without the knowledge of database formats and SQL scripts and corresponding to the W3C XML recommendation from August 2006.

Since the publication of the first version of the CERIF model, it has grown in quantity and quality of the concepts represented in it. The model became a recommendation of the European Commission in 1991 and 2000. It is targeted to providers of research information systems, interested on publishing data according to the Linked Data approach.

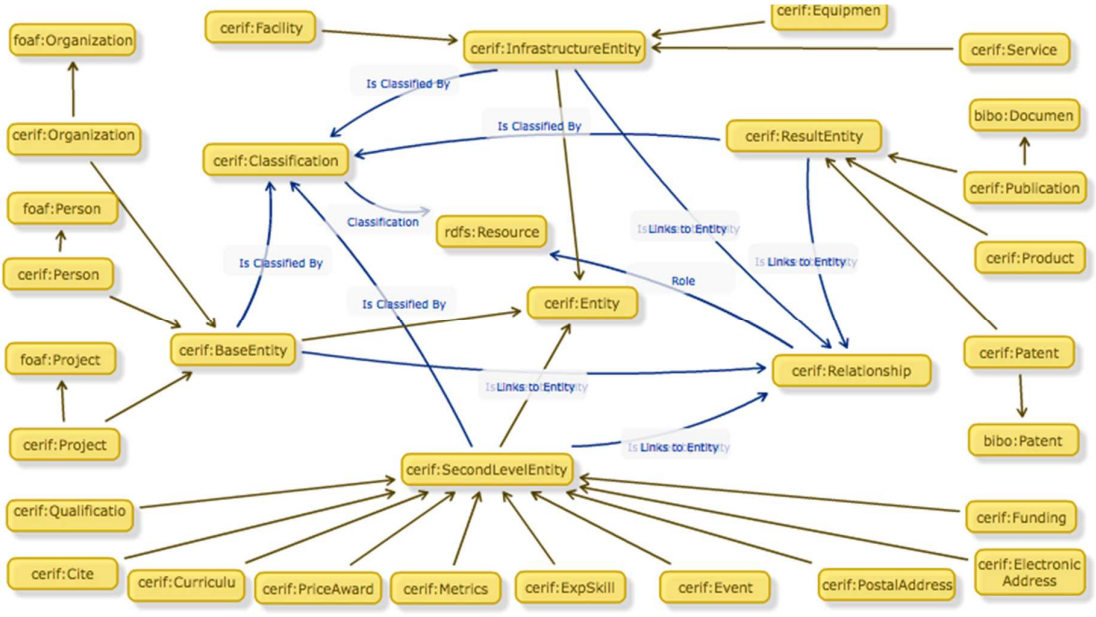



*Figure 1 A snapshot of CERIF 1.3*

The core technology for a wide spread, distributed and structured service for research information is the semantic web technology to integrate database contents with HTML/XML web pages for being provided to the research interested public. One of the best ways to integrate and publish research data is to use RDF (Resource Description Framework) which is also recommended by the W3C consortium, an open Web standard that can be freely used by anyone.

The RDF technology solution has a number of advantages over the XML solution: 1) semantics of new elements and attributes or relations can be described, 2) subsumption of properties allows generalized queries without loss of coverage 3) use of triples of type "anything can say anything about anything", 4) not only research data, but also about statements about research data /reification, 5) extensible structure without changing UI and query.

The transformation to the CERIF RDF model resulted from a bottom-up approach of the transformation of a relational structure to an ontological structure. In order to produce an effective RDF version of the model, basic principles of ontology modelling and knowledge representation rules have been applied. Entities and attributes of the extended relational CERIF model have been transformed into ontology axioms; the types and roles as defined in the CERIF Semantics have been translated into a vocabulary. The CERIF research entities have been transformed into RDF classes and their attributes into properties.

We started with the analysis of CERIF 1.5 since it was documented in greater detail than 1.6 (the current status of CERIF 1.6 version was not finalized; instead it was used for testing and review) model used for the XML data instances that we have obtained from the EKT-hosted CC-REFIM[28] system. The approach was first to finalise and agree on the CERIF 1.5 RDF encoding in order to have a transition to 1.6 that will require very few changes and will be straight forward. The identification and the specification of the rules for the transition from one version to the other and how to preserve backwards compatibility was an important part of the process.

In the beginning, the relational model of CERIF1.5 was analysed and we distinguished in the CERIF entities two main categories. Entities that represent Individuals (cf:BaseEntity, cf:InfrastructureEntity, cf:ResultEntity, cf:SecondLevelEntity) and entities that represent Relationships (cfCite_Medium, cfProj_Equip, cfFund_Fund etc.). We exempted from the Individuals the classes cfClassification and cfClassificationScheme and treated them separately.

We systematically proceeded first with Relationships. The same work was done with Individuals too. We listed all the Relationships in a table and for each one of them we recorded all its fields from the relational model CERIF 1.5. We then followed the principles described in the next section in order to group them and produce a reduced model.

After engineering of the RDF version 1.5 (mapping from the relational 1.5 to the RDF 1.5), the same procedure was followed for CERIF 1.6 version too. A mapping from the relational CERIF version 1.6 to a derivative CERIF version RDF 1.6 was employed and the transition was done without changing any semantics.

Finally, the latest RDF version was used as the basis of the semantic improvement of the model (an RDF model proposal).

[28] CC-REFIM = CERIF-CRIS REFerence Implementation, the result of an internal euroCRIS project, available at http://cc-refim.ekt.gr/cerifcris/

### 6.1.1 Basic Principles

In order to produce the RDF version, we took into account the following principles, a sort of "ontology learning algorithm", which does not entail any arbitrariness.

*Principles for strict preservation of semantics*

1. All relational fields (attributes) and JOINS with the **same semantics** become one property.
2. Every property has only ONE domain and ONE range class. We do not assign a property with same semantics to more than one domain. A property that is repeatedly assigned to more than one classes should be attached to the least general superclass that can have that property and inherit it to all those classes that have at least this property. This "least general superclass" has to be "invented". This results in a dramatic reduction of complexity of a relational source and better comprehension. The labels of the "invented" classes are preliminary working titles.
3. If a group of properties always co-occur (with the same semantics) in more than one class and never occur alone, they are assigned to the same least general superclass.
4. N-ary (3 and more) relationships become classes. Binary relationships become properties.
5. Every RDF class instance represents one instance of the class in the real world which is identified by its own URI (althouqh any identification scheme can be used, we chose URI as the one widely accepted by W3C and the Semantic Web community). This URI can also be regarded as the value of a property "has preferred ID" not represented. Therefore, all fields in the Relational schema or XML that describe a primary URI are not mapped to properties, but used to instantiate the respective RDF class. If more than one URI is described in the source, one may be selected as primary, and not been mapped, but the others are mapped to a property of type "has identifier"; an appropriate classification can be used to preserve the semantics of each identifier, if any.
6. We do not model internal identifiers since they do not make sense in context of linked data. For entities without a primary URI in the source, suitable URIs have to be generated at data transformation time, either UUID or combinations of local keys and constant prefixes.
7. Rules 1, 2, 3, 4, 5, 6 build "automatically" an IsA hierarchy. At the leaves of this hierarchy remain classes which have properties no other class has, and a lot of classes without properties anymore (because their properties are all inherited). All classes without properties can be deleted ("cleaned") and replaced by classification terms.

Exception of the above statement is the following: classes that are multiply classified and have no properties (only inherited) are not removed, in case they ontologically make sense on a top level categorization or they reduce the complexity of the model.

*Semantic complexity reduction rules*

1. Relationships that express combinations of entities without specific semantics are generalized to a new superclass with generalized properties to a superclass subsuming the respective set of combined entities. If there is any systematic constraint in these combinations, a respective inheritance structure with adequate subproperties of the generalized properties can be engineered.
2. Properties are raised to higher superclass if it is meaningful (and the expressive power is increased).
3. Redundant properties are cleaned up.
4. A more reduced and simplified model is achieved by converting the invented general classes to general properties

The cleaned-up IsA hierarchy resulting from this process does not contain any change of semantics from the source. It is an exact representation of the same meaning. As mentioned above, general relationships between entity combinations are ontologically meaningless, i.e., nothing can be concluded from them. Such a combinatorial system has to be inspected, if it is complete, i.e., if it contains all combinations. Then it is sufficiently represented by a generalized property. If certain combinations should be excluded for semantic reasons, and have not just been forgotten by the designer, they can be represented by specializing the general relationships to subconcepts. The decision about these semantic reasons is not "automatic". Besides these, this cleaning step produces a semantic equivalent of the source.

The effect of systematically introducing these "invented" superclasses is an ontology-learning process which results in a dramatic reduction of complexity of a Relational Source. The complete graphic representation becomes readable, and can be used as a basis to further normalize the source semantics. From our experience, an "invented" superclass either has a relevant real world equivalent, or reveals a bad conceptual modelling feature. In the first case, more meaningful labels can be assigned, in the second, the concept should be replaced by better constructs.

### 6.1.2 Current version CERIF 1.6

The major updates in the current CERIF release have been made in order to support recording and thus an understanding of datasets of CKAN, DCAT and eGMS. A dataset is represented in CERIF as a cfResultProduct instance that is classified as Dataset. So, major updates in the CERIF 1.6 release have been made around the CERIF entity cfResultProduct. Therefore, all entities, including the cfResultProduct entity, in addition to their naming (syntax) are enhanced with semantic (contextual) information. The formal CERIF entity cfResultProduct, is in fact a container to aggregate all potential types. So, new link entities from 'dataset' to geographic bounding box and to Patent (cfResultProduct_GeographicBoundingBox, cfResultProduct_ResultPatent) have been added.

A new cfOrder attribute has also been added in links from Results to Persons and Organisations. Additionally, an informing about handling of dates is part of this version; for example, new DateTime attribute with measurements or deprecated dates from Project, Funding and Patent and instead recommendations of putting the dates in a classification or a link to a relevant concept, have been proposed. Other changes include depreciation of attributes or entities such as the Dublin Core part of the model with future releases.

RDFS version for CERIF version 1.6 should overcome some limitations of current semantics in terms of query capability. We used the 1.6 RDF version in order to produce an improved and more simplified version (see fig.2, 3) of the model by applying the ontological rules we have mentioned above.

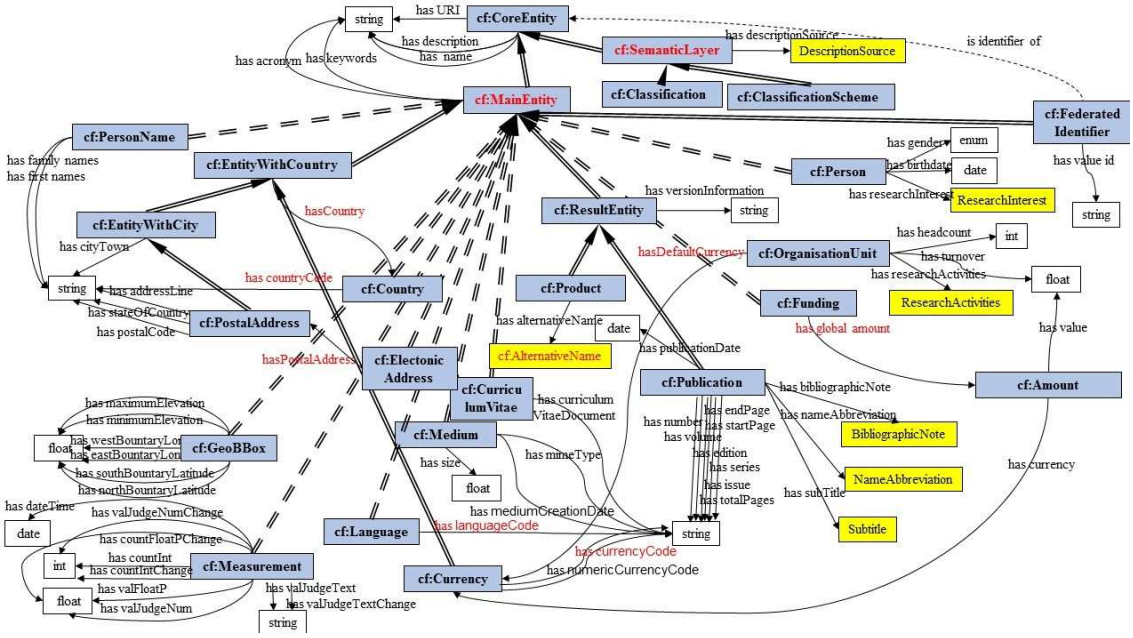



*Figure 2 A simplified representation of individuals (CERIF1.6)*

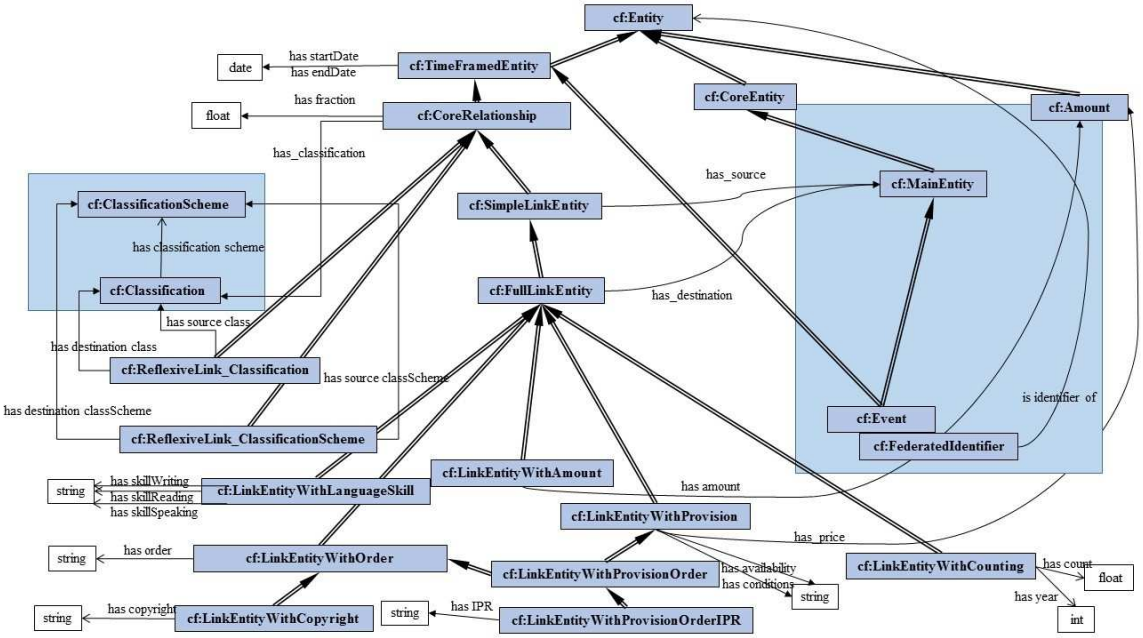



*Figure 3 Representation of proposed relationships of CERIF1.6*

## 6.2 X3ML improvements

As already mentioned in Section 2, in order to allow data transformation and aggregation, it is required to produce mappings, to relate equivalent concepts or relationships from the source schemata to CERIF, the common aggregation schema, i.e. the target schema, in a way that facts described in terms of the source schema can automatically be translated into descriptions in terms of CERIF (target schema). This is the mapping definition process and the output of this task is the mapping, i.e., a collection of mapping rules.

It was decided very early in the project that the mapping process will be supported by the X3ML Toolkit, a set of open source tools developed by the VRE4EIC partner FORTH.

The X3ML Toolkit consists of the following components:

- X3ML mapping definition language, an XML based language which describes schema mappings in such a way that they can be collaboratively created and discussed by experts.
- 3M - the Mapping Memory Manager is a tool for managing mapping definition files.
- 3M Editor is a web application assisting users during the mapping definition process, using a human-friendly UI and a set of components that either suggest or validate the user input.
- X3ML Engine realizes the transformation of the source records to the target format.

During the last two years continuous improvements and debugging of the X3ML Toolkit have taken place. As the mappings are progressing, bugs are revealed, needs for special generators emerge, improved UI functionality is requested.

The X3ML mapping definition language schema has been updated to support source and target namespaces. The following changes have been adopted:

- Added element source under element info containing also the elements source_info and source_collection
- Added element target under element info containing also the elements target_info and target_collection
- Changed the cardinality constraints of the children of the additional element (entity and relationship). More specifically the upper bound is now unlimited.

Apart from their declaration in the info element, we preserve the ability to add namespaces (in general) under the namespace elements (for being backwards compatible with older mappings).

The latest updated version of the X3ML schema (X3ML.xsd) is 1.2 and is being used by the X3ML engine version 1.7.5

In December 2016 Release 3.1 of the 3M Editor has been deployed.

Taking into account the feedback of the users, the User Interface has been redesigned to reflect better the requirements and needs of the users. The release also includes a major update in the Instance Generation process. The editing of the Instance Generators has been redesigned and the default UUID generator is inserted automatically in the mapping definition file. This functionality significantly speeds the mapping process and allows a preliminary testing of the mapping before the actual Instance Generation functions are added. Moreover, an RDFvisualizer component has been implemented and integrated in 3M Editor, allowing the user to examine the result of his mapping in real time.

Release 3.1 also handles better variables and supports both global and map specific variables.

The current release of the X3ML engine is 1.7.5 while three releases took place during the last two years. The engine supports XML2RDF and RDF2RDF mappings. The implementation of XML2XML mapping has been considered but the complexity of such a mapping is considerable, cannot be generic and is beyond the resources of the project.

All the code is available on github https://github.com/isl.

In parallel, FORTH is completing the implementation of the Mapping Suggester (http://139.91.183.3/testMappingSuggester/), a tool that facilitates the mapping process by suggesting relevant mappings to the user. These suggestions make use of “mapping memories” of similar cases collected from the user community and are recalculated with each new mapping decision. The user can either accept or reject the suggestion. When the user creates a new mapping file, the Mapping Suggester runs a schema matching with the source schema provided and the existing mapped source schemata in the mapping memory. This tool will be integrated in the X3ML Toolkit by the end of the project.

# 7 Conclusions

The standards identified in D4.1 have now been analysed and matched to CERIF. The controlled vocabularies used or recommended by the standards have been collected. They can be viewed as a core set of vocabularies that can be ingested into the metadata catalogue as part of the semantic layer of CERIF. This core set of vocabularies will then be available for all other standards that need to be mapped to CERIF.

3M has been used successfully to express various mappings. ISO19115/Inspire and CKAN have been described directly into 3M, and the description of mapping for DCAT-AP and Dublin Core has already been started. These mappings should now be validated and tested against real data to verify their adequacy on specific practical requirements (*T4.3 - Validation of Matching and Mapping*). Once done, the mappings can be integrated to the global architecture of the platform.

# 8 References

Euzenat, J., & Shvaiko, P. (2007). Ontology Matching

Do Hong Hai (2005). Schema Matching and mapping-based data integration

# 9 Annexes

## 9.1 Annexe 1: Dublin Core matching

*Dataset and Software*

As the Dublin Core types Dataset and Software both are mapped to cfResultProduct, the matching is the same.

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *audience* | cfClassification + cfResultProduct_Classification | a vocabulary needs to be used for various audiences used, and stored as a cfClassificationScheme in CERIF |
| *educationLevel* | cfClassification + cfResultProduct_Classification | a vocabulary needs to be used for various education levels used, and stored as a cfClassificationScheme in CERIF |
| *mediator* | cfClassification + cfResultProduct_Classification | a vocabulary needs to be used for various mediators used, and stored as a cfClassificationScheme in CERIF |
| *contributor* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_ResultProduct (classification = Contributor)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_ResultProduct (classification = Contributor)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfResultProduct_Service (classification = Contributor) | cfPersonName only accept structured names, so if the contributor is a person, the name should be structured in Dublin-Core too to automatize the mapping |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *creator* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_ResultProduct (classification = Creator)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_ResultProduct (classification = Creator)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfResultProduct_Service (classification = Creator) | cfPersonName only accept structured names, so if the creator is a person, the name should be structured in Dublin-Core too to automatize the mapping |
| *coverage* | cfClassification + cfResultProduct_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *spatial* | cfClassification + cfResultProduct_Classification<br>or<br>cfGeographicBoundingBox + cfResultProduct_GeographicBoundingBox (classification scheme = "DCMI properties", classification = "spatial coverage") | a vocabulary can be used for a predefined list of spatial references, and stored as a cfClassificationScheme in CERIF<br>or<br>the location can be stored as a cfGeographicBoundingBox, if it respects the DCMI Box Encoding Scheme (http://dublincore.org/documents/dcmi-box/) |
| *temporal* | cfMeasurement.cfValJudgeText + cfResultProduct_Measurement (classification scheme = "DCMI properties", classification = "temporal coverage")<br>or<br>cfClassification + cfResultProduct_Classification | the temporal coverage can be stored in the cfResultProduct_Classification entity, if it respects the DCMI Period Encoding Scheme (http://dublincore.org/documents/dcmi-period/) with classification scheme = "DCMI properties", classification = "temporal coverage" |
| *available* | cfClassification + cfResultProduct_Classification .cfStartDate | classification scheme = "lifecycle", classification = "available"<br>*Note: in CERIF, provenance is supposed to be represented as relations between entities, with* |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| | | *temporal coverage. As we don't have more information about the Person or Organization that made the dataset available, we choose the solution of a classification scheme to manage the lifecycle.* |
| *created* | If the creator is a Person:<br>- cfPerson_ResultProduct.cfStartDate (classification = Creator)<br>If the creator is an Organization:<br>- cfOrganisationUnit_ResultProduct.cfStartDate (classification = Creator)<br>If the creator is a Service:<br>- cfResultProduct_Service.cfStartDate (classification = Creator) | |
| *date* | cfClassification + cfResultProduct_Classification.cfStartDate + cfResultProduct_Classification.cfEndDate | * the lifecycle of the resource needs to be stored as a cfClassificationScheme in CERIF ("lifecycle"), with corresponding steps as cfClassification (or "other" if no steps indicated)<br>* cfResultProduct_Classification.cfEndDate needed only in case of a period |
| *dateAccepted* | cfClassification + cfResultProduct_Classification.cfStartDate | classification scheme = "lifecycle", classification = "accepted"<br>*Note: same comment as for "available" regarding the provenance.* |
| *dateCopyrighted* | cfClassification + cfResultProduct_Classification.cfStartDate | classification scheme = "lifecycle", classification = "copyrighted"<br>*Note: same comment as for "available" regarding the provenance.* |
| *dateSubmitted* | cfClassification + cfResultProduct_Classification.cfStartDate | classification scheme = "lifecycle", classification = "submitted"<br>*Note: same comment as for "available" regarding the provenance.* |
| *issued* | cfClassification + cfResultProduct_Classification.cfStartDate | classification scheme = "lifecycle", classification = "issued"<br>*Note: same comment as for "available" regarding the provenance.* |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *modified* | cfClassification + cfResultProduct_Classification .cfStartDate | classification scheme = "lifecycle", classification = "modified"<br>*Note: same comment as for "available" regarding the provenance.* |
| *valid* | cfClassification + cfResultProduct_Classification .cfStartDate | classification scheme = "lifecycle", classification = "valid"<br>*Note: same comment as for "available" regarding the provenance.* |
| *description* | cfResultProductDescription.cf Description | only a text description will be supported by CERIF |
| *abstract* | cfMeasurement.cfValJudgeTe xt + cfResultProduct_Measureme nt (classification scheme = "DCMI properties", classification = "abstract") | only a text description will be supported by CERIF |
| *tableOfContents* | cfMeasurement.cfValJudgeTe xt + cfResultProduct_Measureme nt (classification scheme = "DCMI properties", classification = "table of contents") | only a text description will be supported by CERIF |
| *extent* | cfMeasurement.cfValJudgeTe xt + cfResultProduct_Measureme nt (classification scheme = "DCMI properties", classification = "extent") | |
| *format* | cfClassification + cfResultProduct_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *identifier* | cfFederatedIdentifier.cfFeder atedIdentifier + cfFederatedIdentifier_Service | the formal identification system needs to be provided as a cfService in CERIF |
| *instructionalMethod* | cfClassification + cfResultProduct_Classification | a vocabulary needs to be used for various types of methods used |
| *language* | cfClassification + cfResultProduct_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *provenance* | cfMeasurement.cfValJudgeText + cfResultProduct_Measurement (classification scheme = "DCMI properties", classification = "provenance") | *Note: in CERIF, provenance is supposed to be represented as relations between entities, with temporal coverage. The provenance field does not allow to provide this kind of rich information that's why we choose to match it this way.* |
| *publisher* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_ResultProduct (classification = Publisher)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_ResultProduct (classification = Publisher)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfResultProduct_Service (classification = Publisher) | cfPersonName only accept structured names, so if the creator is a person, the name should be structured in Dublin-Core too to automatize the mapping |
| *conformsTo* | cfClassification + cfResultProduct_Classification | a vocabulary needs to be used for various types of standards used |
| *hasFormat* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = has format)<br>for a Service:<br>- cfResultProduct_Service (classification = has format)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = is format of) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that is a format of the described resource |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *hasPart* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = has part)<br>for a Service:<br>- cfResultProduct_Service (classification = has part)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = is part of) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that is a part of the described resource |
| *hasVersion* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = has version)<br>for a Service:<br>- cfResultProduct_Service (classification = has version)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = is version of) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that is a version of the described resource |
| *isFormatOf* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = is format of)<br>for a Service:<br>- cfResultProduct_Service (classification = is format of)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = has format) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that has the described resource as a format |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *isPartOf* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = is part of)<br>for a Service:<br>- cfResultProduct_Service (classification = is part of)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = has part) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that has the described resource as a part |
| *isReferencedBy* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = is referenced by)<br>for a Service:<br>- cfResultProduct_Service (classification = is referenced by)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = references) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that references the described resource |
| *isReplacedBy* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = is replaced by)<br>for a Service:<br>- cfResultProduct_Service (classification = is replaced by)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = replaces) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that replaces the described resource |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *isRequiredBy* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = is required by)<br>for a Service:<br>- cfResultProduct_Service (classification = is required by)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = requires) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that requires the described resource |
| *isVersionOf* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = is version of)<br>for a Service:<br>- cfResultProduct_Service (classification = is version of)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = has version) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the resource that is a version of the described resource |
| *references* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = references)<br>for a Service:<br>- cfResultProduct_Service (classification = references)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = is referenced by) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the referenced resource |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *relation* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = related to)<br>for a Service:<br>- cfResultProduct_Service (classification = related to)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = related to) | |
| *replaces* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = replaces)<br>for a Service:<br>- cfResultProduct_Service (classification = replaces)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = is replaced by) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the replaced resource |
| *requires* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = requires)<br>for a Service:<br>- cfResultProduct_Service (classification = requires)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = is required by) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the required resource |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *source* | for a Dataset or Software:<br>- cfResultProduct_ResultProduct (classification = derived from)<br>for a Service:<br>- cfResultProduct_Service (classification = derived from)<br>for a Text:<br>- cfResultPublication_ResultProduct (classification = source of) | for cfResultProduct_ResultProduct, cfResultProductId1 must be the reference of the described resource, and cfResultProductId2 the reference of the source |
| *accessRights* | cfMeasurement.cfValJudgeText + cfResultProduct_Measurement (classification scheme = "DCMI properties", classification = "access rights") | |
| *license* | cfResultPublication_ResultProduct (classification scheme = "Inter-Output Relations", classification = "Relation") | The use of foaf:Document is recommended to represent a LicenseDocument in the source. |
| *rights* | cfClassification + cfResultProduct_Classification | a vocabulary needs to be used for various types of license that can be linked to indicate intellectual property rights |
| *rightsHolder* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_ResultProduct (classification = Rights holder)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_ResultProduct (classification = Rights holder)<br>for a Service:<br>- cfServiceName.cfName + cfService + | |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| | cfResultProduct_Service (classification = Rights holder) | |
| *subject* | cfClassification + cfResultProduct_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *alternative* | cfResultProductAltName.cfAltName | |
| *title* | cfResultProductName.cfName | |
| *type* | cfClassification + cfResultProduct_Classification | cfClassification is the term corresponding to "Dataset"/"Software" within cfClassificationScheme "Dublin-Core Types" |

*Service*

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *audience* | cfClassification + cfService_Classification | a vocabulary needs to be used for various audiences used, and stored as a cfClassificationScheme in CERIF |
| *educationLevel* | cfClassification + cfService_Classification | a vocabulary needs to be used for various education levels used, and stored as a cfClassificationScheme in CERIF |
| *mediator* | cfClassification + cfService_Classification | a vocabulary needs to be used for various mediators used, and stored as a cfClassificationScheme in CERIF |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *contributor* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_Service (classification = Contributor)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_Service (classification = Contributor)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfService_Service (classification = Contributor) | * cfPersonName only accept structured names, so if the contributor is a person, the name should be structured in Dublin-Core too (for automating)<br>* for cfService_Service, cfServiceId1 must be the reference of the contributor and cfServiceId2 must be the reference of the described resource |
| *creator* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_Service (classification = Creator)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_Service (classification = Creator)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfService_Service (classification = Creator) | * cfPersonName only accept structured names, so if the creator is a person, the name should be structured in Dublin-Core too (for automating)<br>* for cfService_Service, cfServiceId1 must be the reference of the creator and cfServiceId2 must be the reference of the described resource |
| *coverage* | cfClassification + cfService_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *spatial* | cfClassification + cfService_Classification<br>or<br>cfGeographicBoundingBox + cfService_GeographicBoundingBox (classification scheme = "DCMI properties", classification = "spatial coverage") | a vocabulary can be used for a predefined list of spatial references, and stored as a cfClassificationScheme in CERIF<br>or<br>the location can be stored as a cfGeographicBoundingBox, if it respects the DCMI Box Encoding Scheme (http://dublincore.org/documents/dcmi-box/)<br>note: cfService_GeographicBoundingBox does not exist in CERIF 1.6 |
| *temporal* | cfMeasurement.cfValJudgeText + cfService_Measurement (classification scheme = "DCMI properties", classification = "temporal coverage")<br>or<br>cfClassification + cfService_Classification | the temporal coverage can be stored in the cfMedium_Classification entity, if it respects the DCMI Period Encoding Scheme (http://dublincore.org/documents/dcmi-period/) with classification scheme = "DCMI properties", classification = "temporal coverage" |
| *available* | cfClassification + cfService_Classification.cfStartDate | classification scheme = "lifecycle", classification = "available" |
| *created* | cfClassification + cfService_Classification.cfStartDate | |
| *date* | cfClassification + cfService_Classification.cfStartDate + cfService_Classification.cfEndDate | * the lifecycle of the resource needs to be stored as a cfClassificationScheme in CERIF ("lifecycle"), with corresponding steps as cfClassification (or "other" if no steps indicated)<br>* cfService_Classification.cfEndDate needed only in case of a period |
| *dateAccepted* | cfClassification + cfService_Classification.cfStartDate | classification scheme = "lifecycle", classification = "accepted" |
| *dateCopyrighted* | cfClassification + cfService_Classification.cfStartDate | classification scheme = "lifecycle", classification = "copyrighted" |
| *dateSubmitted* | cfClassification + cfService_Classification.cfStartDate | classification scheme = "lifecycle", classification = "submitted" |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *issued* | cfClassification + cfService_Classification.cfStartDate | classification scheme = "lifecycle", classification = "issued" |
| *modified* | cfClassification + cfService_Classification.cfStartDate | classification scheme = "lifecycle", classification = "modified" |
| *valid* | cfClassification + cfService_Classification.cfStartDate | classification scheme = "lifecycle", classification = "valid" |
| *description* | cfServiceDescription.cfDescription | only a text description will be supported by CERIF |
| *abstract* | cfMeasurement.cfValJudgeText + cfService_Measurement (classification scheme = "DCMI properties", classification = "abstract") | only a text description will be supported by CERIF |
| *tableOfContents* | cfMeasurement.cfValJudgeText + cfService_Measurement (classification scheme = "DCMI properties", classification = "table of contents") | only a text description will be supported by CERIF |
| *extent* | cfMeasurement.cfValJudgeText + cfService_Measurement (classification scheme = "DCMI properties", classification = "extent") | |
| *format* | cfClassification + cfService_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *identifier* | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service | the formal identification system needs to be provided as a cfService in CERIF |
| *instructionalMethod* | cfClassification + cfService_Classification | a vocabulary needs to be used for various types of methods used |
| *language* | cfClassification + cfService_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *provenance* | cfMeasurement.cfValJudgeText + cfService_Measurement (classification scheme = "DCMI properties", classification = "provenance") | |
| *publisher* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_Service (classification = Publisher)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_Service (classification = Publisher)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfService_Service (classification = Publisher) | * cfPersonName only accept structured names, so if the creator is a person, the name should be structured in Dublin-Core too (for automating)<br>* for cfService_Service, cfServiceId1 must be the reference of the publisher and cfServiceId2 must be the reference of the described resource |
| *conformsTo* | cfClassification + cfService_Classification | a vocabulary needs to be used for various types of standards used |
| *hasFormat* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = is format of)<br>for a Service:<br>- cfService_Service (classification = has format)<br>for a Text:<br>- cfResultPublication_Service (classification = is format of) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that is a format of the described resource |
| *hasPart* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = is part of)<br>for a Service:<br>- cfService_Service (classification = has part)<br>for a Text:<br>- cfResultPublication_Service (classification = is part of) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that is a part of the described resource |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *hasVersion* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = is version of)<br>for a Service:<br>- cfService_Service (classification = has version)<br>for a Text:<br>- cfResultPublication_Service (classification = is version of) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that is a version of the described resource |
| *isFormatOf* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = has format)<br>for a Service:<br>- cfService_Service (classification = is format of)<br>for a Text:<br>- cfResultPublication_Service (classification = has format) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that has the described resource as a format |
| *isPartOf* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = has part)<br>for a Service:<br>- cfService_Service (classification = is part of)<br>for a Text:<br>- cfResultPublication_Service (classification = has part) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that has the described resource as a part |
| *isReferencedBy* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = references)<br>for a Service:<br>- cfService_Service (classification = is referenced by)<br>for a Text:<br>- cfResultPublication_Service (classification = references) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that references the described resource |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *isReplacedBy* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = replaces)<br>for a Service:<br>- cfService_Service (classification = is replaced by)<br>for a Text:<br>- cfResultPublication_Service (classification = replaces) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that replaces the described resource |
| *isRequiredBy* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = requires)<br>for a Service:<br>- cfService_Service (classification = is required by)<br>for a Text:<br>- cfResultPublication_Service (classification = requires) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that requires the described resource |
| *isVersionOf* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = has version)<br>for a Service:<br>- cfService_Service (classification = is version of)<br>for a Text:<br>- cfResultPublication_Service (classification = has version) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the resource that is a version of the described resource |
| *references* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = is referenced by)<br>for a Service:<br>- cfService_Service (classification = references)<br>for a Text:<br>- cfResultPublication_Service (classification = is referenced by) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the referenced resource |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *relation* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = related to)<br>for a Service:<br>- cfService_Service (classification = related to)<br>for a Text:<br>- cfResultPublication_Service (classification = related to) | |
| *replaces* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = is replaced by)<br>for a Service:<br>- cfService_Service (classification = replaces)<br>for a Text:<br>- cfResultPublication_Service (classification = is replaced by) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the replaced resource |
| *requires* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = is required by)<br>for a Service:<br>- cfService_Service (classification = requires)<br>for a Text:<br>- cfResultPublication_Service (classification = is required by) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the required resource |
| *source* | for a Dataset or Software:<br>- cfResultProduct_Service (classification = source of)<br>for a Service:<br>- cfService_Service (classification = derived from)<br>for a Text:<br>- cfResultPublication_Service (classification = source of) | for cfService_Service, cfServiceId1 must be the reference of the described resource, and cfServiceId2 the reference of the source |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *accessRights* | cfMeasurement.cfValJudgeText + cfService_Measurement (classification scheme = "DCMI properties", classification = "access rights") | |
| *license* | cfClassification + cfService_Classification | a vocabulary needs to be used for various types of license that can be linked to indicate intellectual property rights |
| *rights* | cfClassification + cfService_Classification | a vocabulary needs to be used for various types of license that can be linked to indicate intellectual property rights |
| *rightsHolder* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_Service (classification = Rights holder)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_Service (classification = Rights holder)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfService_Service (classification = Rights holder) | * cfPersonName only accept structured names, so if the creator is a person, the name should be structured in Dublin-Core too (for automating)<br>* for cfService_Service, cfServiceId1 must be the reference of the publisher and cfServiceId2 must be the reference of the described resource |
| *subject* | cfClassification + cfService_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *alternative* | cfMeasurement.cfValJudgeText + cfService_Measurement (classification scheme = "DCMI properties", classification = "alternative title") | |
| *title* | cfServiceName.cfName | |

| Dublin Core Property | CERIF | Comments |
|---|---|---|
| *type* | cfClassification + cfService_Classification | cfClassification is the term corresponding to "InteractiveResource"/"Service" within cfClassificationScheme "Dublin-Core Types" |

*Text*

| Property | CERIF | Comments |
|---|---|---|
| *audience* | cfClassification + cfResultPublication_Classification | a vocabulary needs to be used for various audiences used, and stored as a cfClassificationScheme in CERIF |
| *educationLevel* | cfClassification + cfResultPublication_Classification | a vocabulary needs to be used for various education levels used, and stored as a cfClassificationScheme in CERIF |
| *mediator* | cfClassification + cfResultPublication_Classification | a vocabulary needs to be used for various mediators used, and stored as a cfClassificationScheme in CERIF |
| *contributor* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_ResultPublication (classification = Contributor)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_ResultPublication (classification = Contributor)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfResultPublication_Service (classification = Contributor) | cfPersonName only accept structured names, so if the contributor is a person, the name should be structured in Dublin-Core too (for automating) |

| Property | CERIF | Comments |
|---|---|---|
| *creator* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_ResultPublication (classification = Creator)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_ResultPublication (classification = Creator)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfResultPublication_Service (classification = Creator) | cfPersonName only accept structured names, so if the creator is a person, the name should be structured in Dublin-Core too (for automating) |
| *coverage* | cfClassification + cfResultPublication_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *spatial* | cfClassification + cfResultPublication_Classification<br>or<br>cfGeographicBoundingBox + cfResultPublication_GeographicBoundingBox (classification scheme = "DCMI properties", classification = "spatial coverage") | a vocabulary can be used for a predefined list of spatial references, and stored as a cfClassificationScheme in CERIF<br>or<br>the location can be stored as a cfGeographicBoundingBox, if it respects the DCMI Box Encoding Scheme (http://dublincore.org/documents/dcmi-box/)<br>note: cfResultPublication_GeographicBoundingBox does not exist in CERIF 1.6 |
| *temporal* | cfMeasurement.cfValJudgeText + cfResultPublication_Measurement (classification scheme = "DCMI properties", classification = "temporal coverage")<br>or<br>cfClassification + cfResultPublication_Classification | the temporal coverage can be stored in the cfMedium_Classification entity, if it respects the DCMI Period Encoding Scheme (http://dublincore.org/documents/dcmi-period/) with classification scheme = "DCMI properties", classification = "temporal coverage" |

| Property | CERIF | Comments |
|---|---|---|
| *available* | cfResultPublication.cfResultPublicationDate | classification scheme = "lifecycle", classification = "available" |
| *created* | cfClassification + cfResultPublication_Classification.cfStartDate | |
| *date* | cfResultPublication_Classification.cfStartDate + cfResultPublication_Classification.cfEndDate | * the lifecycle of the resource needs to be stored as a cfClassificationScheme in CERIF ("lifecycle"), with corresponding steps as cfClassification (or "other" if no steps indicated)<br>* cfResultPublication_Classification.cfEndDate needed only in case of a period |
| *dateAccepted* | cfClassification + cfResultPublication_Classification.cfStartDate | classification scheme = "lifecycle", classification = "accepted" |
| *dateCopyrighted* | cfClassification + cfResultPublication_Classification.cfStartDate | classification scheme = "lifecycle", classification = "copyrighted" |
| *dateSubmitted* | cfClassification + cfResultPublication_Classification.cfStartDate | classification scheme = "lifecycle", classification = "submitted" |
| *issued* | cfClassification + cfResultPublication_Classification.cfStartDate | classification scheme = "lifecycle", classification = "issued" |
| *modified* | cfClassification + cfResultPublication_Classification.cfStartDate | classification scheme = "lifecycle", classification = "modified" |
| *valid* | cfClassification + cfResultPublication_Classification.cfStartDate | classification scheme = "lifecycle", classification = "valid" |
| *description* | cfMeasurement.cfValJudgeText + cfResultPublication_Measurement (classification scheme = "DCMI properties", classification = "description") | only a text description will be supported by CERIF |
| *abstract* | cfResultPublicationAbstr.cfAbstr | only a text description will be supported by CERIF |
| *tableOfContents* | cfMeasurement.cfValJudgeText + cfResultPublication_Measurement (classification scheme = "DCMI properties", | only a text description will be supported by CERIF |

| Property | CERIF | Comments |
|---|---|---|
| | classification = "table of contents") | |
| *extent* | cfMeasurement.cfValJudgeText + cfResultPublication_Measurement (classification scheme = "DCMI properties", classification = "extent") | |
| *format* | cfClassification + cfResultPublication_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *bibliographicCitation* | cfCiteDescription.cfDescription + cfCite + cfResultPublication_Cite (classification = Citation) | |
| *identifier* | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service | the formal identification system needs to be provided as a cfService in CERIF |
| *instructionalMethod* | cfClassification + cfResultPublication_Classification | a vocabulary needs to be used for various types of methods used |
| *language* | cfClassification + cfResultPublication_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *provenance* | cfMeasurement.cfValJudgeText + cfResultPublication_Measurement (classification scheme = "DCMI properties", classification = "provenance") | |

| Property | CERIF | Comments |
|---|---|---|
| *publisher* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_ResultPublication (classification = Publisher)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_ResultPublication (classification = Publisher)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfResultPublication_Service (classification = Publisher) | cfPersonName only accept structured names, so if the creator is a person, the name should be structured in Dublin-Core too (for automating) |
| *conformsTo* | cfClassification + cfResultPublication_Classification | a vocabulary needs to be used for various types of standards used |
| *hasFormat* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = has format)<br>for a Service:<br>- cfResultPublication_Service (classification = has format)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = has format) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that is a format of the described resource |
| *hasPart* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = has part)<br>for a Service:<br>- cfResultPublication_Service (classification = has part)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = has | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that is a part of the described resource |

| Property | CERIF | Comments |
|---|---|---|
| | part) | |
| *hasVersion* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = has version)<br>for a Service:<br>- cfResultPublication_Service (classification = has version)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = has version) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that is a version of the described resource |
| *isFormatOf* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = is format of)<br>for a Service:<br>- cfResultPublication_Service (classification = is format of)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = is format of) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that has the described resource as a format |

| Property | CERIF | Comments |
|---|---|---|
| *isPartOf* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = is part of)<br>for a Service:<br>- cfResultPublication_Service (classification = is part of)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = is part of) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that has the described resource as a part |
| *isReferencedBy* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = is referenced by)<br>for a Service:<br>- cfResultPublication_Service (classification = is referenced by)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = is referenced by) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that references the described resource |
| *isReplacedBy* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = is replaced by)<br>for a Service:<br>- cfResultPublication_Service (classification = is replaced by)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = is replaced by) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that replaces the described resource |

| Property | CERIF | Comments |
|---|---|---|
| *isRequiredBy* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = is required by)<br>for a Service:<br>- cfResultPublication_Service (classification = is required by)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = is required by) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that requires the described resource |
| *isVersionOf* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = is version of)<br>for a Service:<br>- cfResultPublication_Service (classification = is version of)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = is version of) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the resource that is a version of the described resource |
| *references* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = references)<br>for a Service:<br>- cfResultPublication_Service (classification = references)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = references) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the referenced resource |

| Property | CERIF | Comments |
|---|---|---|
| *relation* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = related to)<br>for a Service:<br>- cfResultPublication_Service (classification = related to)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = related to) | |
| *replaces* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = replaces)<br>for a Service:<br>- cfResultPublication_Service (classification = replaces)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = replaces) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the replaced resource |
| *requires* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = requires)<br>for a Service:<br>- cfResultPublication_Service (classification = requires)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = requires) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the required resource |

| Property | CERIF | Comments |
|---|---|---|
| *source* | for a Dataset or Software:<br>- cfResultPublication_ResultProduct (classification = derived from)<br>for a Service:<br>- cfResultPublication_Service (classification = derived from)<br>for a Text:<br>- cfResultPublication_ResultPublication (classification = derived from) | for cfResultPublication_ResultPublication, cfResultPublicationId1 must be the reference of the described resource, and cfResultPublicationId2 the reference of the source |
| *accessRights* | cfMeasurement.cfValJudgeText + cfResultPublication_Measurement (classification scheme = "DCMI properties", classification = "access rights") | |
| *license* | cfClassification + cfResultPublication_Classification | a vocabulary needs to be used for various types of license that can be linked to indicate intellectual property rights |
| *rights* | cfClassification + cfResultPublication_Classification | a vocabulary needs to be used for various types of license that can be linked to indicate intellectual property rights |
| *rightsHolder* | for a Person:<br>- cfPersonName + cfPersonName_Person + cfPerson + cfPerson_ResultPublication (classification = Rights holder)<br>for an Organization:<br>- cfOrganisationUnitName.cfName + cfOrganisationUnit + cfOrganisationUnit_ResultPublication (classification = Rights holder)<br>for a Service:<br>- cfServiceName.cfName + cfService + cfResultPublication_Service (classification = Rights holder) | cfPersonName only accept structured names, so if the creator is a person, the name should be structured in Dublin-Core too (for automating) |

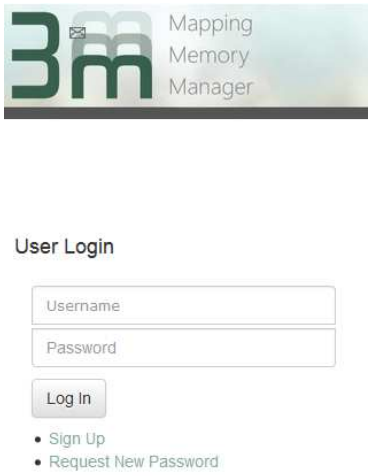
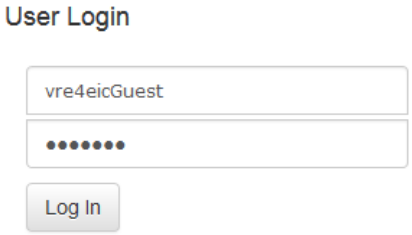
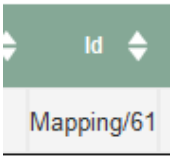
| Property | CERIF | Comments |
|---|---|---|
| *subject* | cfClassification + cfResultPublication_Classification | best practice recommends to use a controlled vocabulary, which needs to be stored as a cfClassificationScheme in CERIF, with terms stored as cfClassification |
| *alternative* | cfResultPublicationSubtitle.cfSubtitle | |
| *title* | cfResultPublicationTitle.cfTitle | |
| *type* | cfClassification + cfResultPublication_Classification | cfClassification is the term corresponding to "Text" within cfClassificationScheme "Dublin-Core Types" |

## 9.2 Annexe 2: ISO19115/INSPIRE mapping

Part of mapping between ISO19115/INSPIRE and CERIF expressed in the 3M tool is shown on the following image.

| # | | SOURCE | TARGET | CONSTANT EXPRESSION | |
|---|---|---|---|---|---|
| 1 | D | ../gmd:MD_Metadata | Product | | |
| 1.1 | P | gmd:fileIdentifier | has_identifier<br>FederatedIdentifier<br>has_id_value | | |
| | R | gmd:fileIdentifier | string | | |
| 1.2 | P | ../gmd:MD_CharacterSetCode | is_source_of<br>SimpleLinkEntity<br>has_classification<br>Classification<br>has_term | [has_startDate]<br>[has_endDate]<br>[has_URI] | [date = ""]<br>[date = ""]<br>[string = ""] |
| | R | ../gmd:MD_CharacterSetCode | PlainLiteral | | |
| 1.3 | P | gmd:parentIdentifier | has_identifier<br>FederatedIdentifier<br>has_id_value | | |
| | R | gmd:parentIdentifier | string | | |
| 1.4 | P | ../gmd:MD_ScopeCode | is_source_of<br>SimpleLinkEntity<br>has_classification<br>Classification<br>has_term | [has_startDate]<br>[has_endDate]<br>[has_URI] | [date = ""]<br>[date = ""]<br>[string = ""] |
| | R | ../gmd:MD_ScopeCode | PlainLiteral | | |
| 1.5 | P | gmd:hierarchyLevelName | is_source_of<br>SimpleLinkEntity<br>has_classification<br>Classification<br>has_term | [has_startDate]<br>[has_endDate] | [date = ""]<br>[date = ""] |
| | R | gmd:hierarchyLevelName | PlainLiteral | | |
| 1.6 | P | ../gmd:LanguageCode | is_source_of<br>SimpleLinkEntity<br>has_classification<br>Classification<br>has_term | [has_startDate]<br>[has_endDate]<br>[has_URI] | [date = ""]<br>[date = ""]<br>[string = ""] |
| | R | ../gmd:LanguageCode | PlainLiteral | | |
| 1.7 | P | gmd:contact | is_destination_of | | |
| | R | gmd:contact | LinkEntityWithProvisionOrderIPR | | |
| 1.8 | P | gmd:contact | is_destination_of | | |
| | R | gmd:contact | FullLinkEntity | | |
| 1.9 | P | gmd:metadataStandardName | is_source_of<br>SimpleLinkEntity<br>has_classification<br>Classification | [has_startDate]<br>[has_endDate] | [date = ""]<br>[date = ""] |

To view the full mapping available in 3M:

| | |
|---|---|
| • Open a web browser at the following location: http://www.ics.forth.gr/isl/3M-VRE4EIC |  |
| • Fill in the username with and the password with the following credentials and click on the Log In button<br>• Username: vre4eicGuest<br>• Password: vre4eic |  |
| • Filter the table of mappings with the number 61 | |
| • Click on Mapping/61 |  |
| • Click on the view icon in the top menu | |

A new window opens with the full mapping description.

## 9.3 Annexe 3: DCAT-AP matching

*Agent*

**Person**

The matching is based on the FoaF vocabulary. Properties “name” and “type” are marked as respectively mandatory and recommended in DCAT-AP. All other properties come from the foaf vocabulary, either having foaf:Agent or foaf:Person as a domain, either being relevant for the concept of person. Amongst these properties, some have not been mapped, either because they need some additional concepts to be consistent (“current project” for example), either because the definition is too vague to define a consistent matching to CERIF concepts (“made” for example).

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *name* | foaf:name | rdfs:Literal | cfPersonName.cfOtherNames + cfPersonName_Person (classification term = fullname) | |
| *type* | dct:type | skos:Concept | link to classification (ADMS publisher type vocabulary) | ADMS publisher type vocabulary http://purl.org/adms/publishertype/ |
| *account* | foaf:account | foaf:OnlineAccount | | foaf:OnlineAccount has not been mapped, this property will not be mapped in the context of this project |
| *current project* | foaf:currentProject | owl:Thing | | There is no definition of a project that can be mapped to CERIF, this property will not be mapped in the context of this project |
| *image* | foaf:img | foaf:Image | | foaf:Image has not been mapped, this property will not be mapped in the context of this project |
| *interest* | foaf:interest | foaf:Document | cfPerson_ResultPublication (classification = interest in) | |
| *knows* | foaf:knows | foaf:Person | cfPerson_Person (classification scheme = related to) | |
| *made* | foaf:made | owl:Thing | | This property will not be mapped. |
| *personal mailbox* | foaf:mbox | owl:Thing | cfPerson_ElectronicAddress (classification term = personal mailbox) + cfElectronicAddress.cfURI | |
| *homepage* | foaf:homepage | foaf:Document | cfPerson_ElectronicAddress (classification term = homepage) + cfElectronicAddress.cfURI | |
| *openid* | foaf:openid | foaf:Document | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = OpenID) | |
| *tipjar* | foaf:tipjar | foaf:Document | | This property will not be mapped. |

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *weblog* | foaf:weblog | foaf:Document | cfPerson_ElectronicAddress (classification term = weblog) + cfElectronicAddress.cfURI | |
| *past project* | foaf:pastProject | owl:Thing | | There is no definition of a project that can be mapped to CERIF, this property will not be mapped in the context of this project |
| *phone* | foaf:phone | | cfPerson_ElectronicAddress (classification term = phone) + cfElectronicAddress.cfURI | |
| *publications* | foaf:publications | foaf:Document | cfPersonResultPublication (classification term = author) | |
| *schoolHomepage* | foaf:schoolHomepage | foaf:Document | | The school a person attended is supposed to be stored as a cfOrganisationUnit in CERIF, with the school homepage attached to it; as there is no more information than this homepage, we cannot fill properly the cfOrganisationUnit; thus, this property will not be mapped |
| *topic_interest* | foaf:topic_interest | owl:Thing | cfPerson.cfResearchInterest | |
| *work info homepage* | foaf:workInfoHomepage | foaf:Document | | The organisation a person works for is supposed to be stored as a cfOrganisationUnit in CERIF. The work info homepage should then be linked to the link between the cfPerson and the cfOrganisationUnit. CERIF does not provide any mechanism to store such data. As the cfOrganisationUnit is not known, this property will not be mapped |

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *workplace homepage* | foaf:workplaceHomepage | foaf:Document | | The organisation a person works for is supposed to be stored as a cfOrganisationUnit in CERIF, with the workplace homepage attached to it; as there is no more information than this homepage, we cannot fill properly the cfOrganisationUnit; thus, this property will not be mapped |
| *age* | foaf:age | rdfs:Literal | | derived from cfPerson.cfBirthdate |
| *birthday* | foaf:birthday | rdfs:Literal | cfPerson.cfBirthdate | |
| *familyName* | foaf:familyName | rdfs:Literal | cfPersonName.cfFamilyNames + cfPersonName_Person (classification term = current name) | |
| *firstName* | foaf:firstName | rdfs:Literal | cfPersonName.cfFirstNames + cfPersonName_Person (classification term = current name) | |
| *gender* | foaf:gender | rdfs:Literal | cfPerson.cfGender | |
| *Given name* | foaf:givenName | | cfPersonName.cfFirstNames + cfPersonName_Person (classification term = current name) | if firstName not filled |
| *jabber ID* | foaf:jabberID | rdfs:Literal | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = Jabber) | |
| *lastName* | foaf:lastName | rdfs:Literal | cfPersonName.cfFamilyNames + cfPersonName_Person (classification term = current name) | if familyName not filled |
| *sha1sum of a personal mailbox URI name* | foaf:mbox_sha1sum | rdfs:Literal | cfElectronicAddress_Measurement (classification term = sha1sum) + cfMeasurement.cfValJudgeText | cfElectronicAddress is the one corresponding to foaf:mbox |

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *myersBriggs* | foaf:myersBriggs | rdfs:Literal | cfPerson_Classification (classification = MBTI) | |
| *nickname* | foaf:nick | | cfPersonName.cfOtherNames + cfPersonName_Person (classification term = nickname) | |
| *AIM chat ID* | foaf:aimChatID | rdfs:Literal | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = AIM) | |
| *ICQ chat ID* | foaf:icqChatID | rdfs:Literal | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = ICQ) | |
| *MSN chat ID* | foaf:msnChatID | rdfs:Literal | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = MSN) | |
| *Skype ID* | foaf:skypeID | rdfs:Literal | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = Skype) | |
| *Yahoo chat ID* | foaf:yahooChatID | rdfs:Literal | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = Yahoo) | |
| *plan* | foaf:plan | rdfs:Literal | | More information needed about what a .plan comment is. |
| *status* | foaf:status | rdfs:Literal | cfPerson_Classification (classification term = status) | |
| *title* | foaf:title | | cfPerson_Classification (classification term = title) | |

**Organisation**

The matching is based on the FoaF vocabulary. Properties “name” and “type” are marked as respectively mandatory and recommended in DCAT-AP. All other properties come from the foaf vocabulary, either having foaf:Agent or foaf:Organization as a domain, either being relevant for the concept of organisation. Amongst these properties, some have not been mapped, either because they need some additional concepts to be consistent (“account” for example), either because they are not relevant for the concept of organisation (“gender” for example).

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *name* | foaf:name | rdfs:Literal | cfOrganisationUnit.cfName | |
| *type* | dct:type | skos:Concept | link to Classification (ADMS publisher type vocabulary) | ADMS publisher type vocabulary http://purl.org/adms/publishertype/ |
| *account* | foaf:account | foaf:OnlineAccount | | foaf:OnlineAccount has not been mapped, this property will not be mapped in the context of this project |
| *AIM chat ID* | foaf:aimChatID | | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = AIM) | |
| *ICQ chat ID* | foaf:icqChatID | | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = ICQ) | |
| *interest* | foaf:interest | foaf:Document | cfOrganisationUnit_ResultPublication (classification = interest in) | |
| *jabber ID* | foaf:jabberID | | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = Jabber) | |
| *made* | foaf:made | owl:Thing | | This property will not be mapped. |
| *personal mailbox* | foaf:mbox | owl:Thing | cfOrganisationUnit_ElectronicAddress (classification term = personal mailbox) + cfElectronicAddress.cfURI | |

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *sha1sum of a personal mailbox URI name* | foaf:mbox_sha1sum | | cfElectronicAddress_Measurement (classification term = sha1sum) + cfMeasurement.cfValJudgeText | cfElectronicAddress is the one corresponding to foaf:mbox |
| *MSN chat ID* | foaf:msnChatID | | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = MSN) | |
| *homepage* | foaf:homepage | foaf:Document | cfOrganisationUnit_ElectronicAddress (classification term = homepage) + cfElectronicAddress.cfURI | |
| *openid* | foaf:openid | foaf:Document | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = OpenID) | |
| *tipjar* | foaf:tipjar | foaf:Document | | This property will not be mapped. |
| *weblog* | foaf:weblog | foaf:Document | cfOrganisationUnit_ElectronicAddress (classification term = weblog) + cfElectronicAddress.cfURI | |
| *phone* | foaf:phone | | cfOrganisationUnit_ElectronicAddress (classification term = phone) + cfElectronicAddress.cfURI | |
| *topic_interest* | foaf:topic_interest | owl:Thing | cfOrganisationUnit.cfResearchActivity | |
| *Yahoo chat ID* | foaf:yahooChatID | rdfs:Literal | cfFederatedIdentifier.cfFederatedIdentifier + cfFederatedIdentifier_Service (classification term = issued by) + cfService (cfName = Yahoo) | |
| *age* | foaf:age | rdfs:Literal | | This property will not be mapped. |
| *birthday* | foaf:birthday | rdfs:Literal | | This property will not be mapped (property inherited from foaf:Agent). |

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *gender* | foaf:gender | rdfs:Literal | | This property will not be mapped (property inherited from foaf:Agent). |
| *nickname* | foaf:nick | | cfOrganisationUnit.cfAcronym | |
| *status* | foaf:status | rdfs:Literal | cfOrganisationUnit_Classification (classification term = status) | |

*Dataset*

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *description* | dct:description | rdfs:Literal | cfResultProduct.cfDescription | |
| *title* | dct:title | rdfs:Literal | cfResultProduct.cfName | |
| *contact point* | dcat:contactPoint | vcard:Kind | link to cfPerson (classification term = Contact) | we recommend to use foaf:Person instead of vcard:Kind |
| *dataset distribution* | dcat:distribution | dcat:Distribution | link to cfMedium (classification term = Distribution) | |
| *keyword/ tag* | dcat:keyword | rdfs:Literal | cfResultProduct.cfKeywords (all keywords should be concatened with a coma as separator) | |
| *publisher* | dct:publisher | foaf:Agent | link to cfOrganisationUnit (classification term = Publisher) | MDR Corporate bodies Named Authority List http://publications.europa.eu/mdr/authority/corporate-body/index.html |
| *theme/category* | dcat:theme, subproperty of dct:subject | skos:Concept | link to cfClassification (Dataset Theme Vocabulary) | Dataset Theme Vocabulary http://publications.europa.eu/resource/authority/datatheme (http://publications.europa.eu/mdr/resource/authority/data-theme/html/data-theme-eng.html) |
| *access rights* | dct:accessRights | dct:RightsStatement | link to cfClassification (classification term = [public, restricted, non-public]) | dataset access rights [public, restricted, non-public] |
| *conforms to* | dct:conformsTo | dct:Standard | link to cfClassification (classification term = conforms to) | |

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *documentation* | foaf:page | foaf:Document | link to cfResultPublication (classification term = document) | |
| *frequency* | dct:accrualPeriodicity | dct:Frequency | link to cfClassification (MDR Frequency Named Authority List) | MDR Frequency Named Authority List http://publications.europa.eu/mdr/authority/frequency/index.html |
| *has version* | dct:hasVersion | dcat:Dataset | link to cfProduct as source (classification term = has version) | |
| *identifier* | dct:identifier | rdfs:Literal | link to cfFederatedIdentifier (cfFederatedIdentifier = identifier, cfStartDate = release date) | |
| *is version of* | dct:isVersionOf | dcat:Dataset | link to cfResultProduct as destination (classification term = has version) | |
| *landing page* | dcat:landingPage | foaf:Document | link to cfElectronicAddress (classification term=Homepage) | |
| *language* | dct:language | dct:LinguisticSystem | link to cfClassification (MDR Languages Named Authority List) | MDR Languages Named Authority List http://publications.europa.eu/mdr/authority/language/ |
| *other identifier* | adms:identifier | adms:Identifier | link to cfFederatedIdentifier (cfFederatedIdentifier = other identifier) | |
| *provenance* | dct:provenance | dct:ProvenanceStatement | link to cfResultPublication (classification term = provenance) | *Note: in CERIF, provenance is supposed to be represented as relations between entities, with temporal coverage. The provenance field does not allow to provide this kind of rich information that's why we choose to match it this way.* |
| *related resource* | dct:relation | rdfs:Resource | for a relation to another dataset, link to cfResultProduct (classification term = related to) | |

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *release date* | dct:issued | rdfs:Literal typed as xsd:date or xsd:dateTime | link to cfClassification (term = issued, cfStartDate = release date) | |
| *sample* | adms:sample | dcat:Distribution | link to cfMedium (classification term = has sample) | |
| *source* | dct:source | dcat:Dataset | link to cfResultProduct as destination (classification term = derived from) | |
| *spatial/geographical coverage* | dct:spatial | dct:Location | link to cfClassification ([MDR Continents Named Authority List, MDR Countries Named Authority List, MDR Places Named Authority List, Geonames]) | MDR Continents Named Authority List<br>http://publications.europa.eu/mdr/authority/country/<br>MDR Countries Named Authority List<br>http://publications.europa.eu/mdr/authority/place/<br>MDR Places Named Authority List<br>http://publications.europa.eu/mdr/authority/continent/<br>Geonames<br>http://sws.geonames.org/ |
| *temporal coverage* | dct:temporal | dct:PeriodOfTime | link to cfClassification (classification term = temporal coverage) | |
| *type* | dct:type | skos:Concept | link to cfClassification (classification term = ?) | |
| *update/modification date* | dct:modified | rdfs:Literal typed as xsd:date or xsd:dateTime | link to cfClassification (classification term = modified, cfStartDate = modification date) | |
| *version* | owl:versionInfo | rdfs:Literal | link to cfMeasurement (classification term = version number) | |
| *version notes* | adms:versionNotes | rdfs:Literal | cfResultProduct.cfVersionInformation | |

*Distribution*

| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| *access URL* | dcat:accessURL | dfs:Resource | cfMedium.cfURI | |
| *description* | dct:description | rdfs:Literal | cfMedium.cfDescription | |
| *format* | dct:format | dct:MediaTypeOrExtent | link to cfClassification (MDR File Type Named Authority List) | MDR File Type Named Authority List<br>http://publications.europa.eu/mdr/authority/file-type/ |
| *licence* | dct:license | dct:LicenseDocument | cfResultPublication_ResultProduct (classification scheme = "Inter-Output Relations", classification = "Relation") | The use of foaf:Document as LicenseDocument is recommended in the source. |
| *byte size* | dcat:bytesize | rdfs:Literal typed as xsd:decimal | link to cfMeasurement (classification term = size) | |
| *checksum* | spdx:checksum | spdx:checksum | link to cfMeasurement (classification term = checksum) | |
| *documentation* | foaf:page | foaf:document | link to cfMedium (classification term = document) | |
| *download url* | dcat:downloadurl | rdfs:Resource | link to cfElectronicAddress (classification term=download) | |
| *language* | dct:language | dct:Linguisticsystem | link to cfClassification (MDR Languages Named Authority List) | MDR Languages Named Authority List<br>http://publications.europa.eu/mdr/authority/language/ |
| *linked schemas* | dct:conformsto | dct:Standard | link to cfClassification (classification term = conforms to) | |
| *media type* | dcat:mediatype, subproperty of dct:format | dct:MediaTypeOrExtent | link to cfClassification (classification term = type) | IANA Media Types<br>http://www.iana.org/assignments/media-types/mediatypes.xhtml |
| *release date* | dct:issued | rdfs:Literal typed as xsd:date or xsd:datetime | link to cfClassification (classification term = issued, cfStartDate = release date) | |
| *rights* | dct:rights | dct:RightsStatement | link to cfResultPublication (classification term = rights) | |
| *status* | adms:status | skos:Concept | link to cfClassification (ADMS status vocabulary) | ADMS status vocabulary<br>http://purl.org/adms/status/ |

| | | | | |
|---|---|---|---|---|
| *title* | dct:title | rdfs:Literal | cfMedium.cfTitle | |
| *update/modification date* | dct:modified | rdfs:Literal typed as xsd:date or xsd:datetime | link to cfClassification (classification term = modified, cfStartDate = modification date) | |

*Document*

The matching is based on the FoaF vocabulary. All properties come from the foaf vocabulary, either having foaf:Document as a domain, either being relevant for the concept of document.

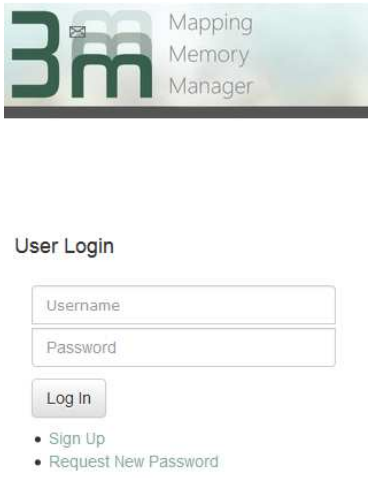
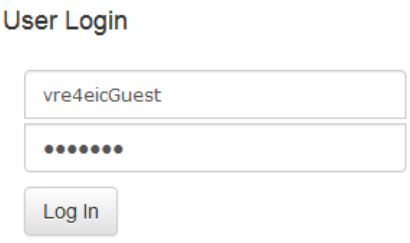
| Property | URI | Range | CERIF | Comment |
|---|---|---|---|---|
| *name* | foaf:name | rdfs:Literal | cfResultPublication.cfTitle | |
| *sha1 sum (hex)* | foaf:sha1 | | cfResultPublication_Measurement (classification term = sha1sum) + cfMeasurement.cfValJudgeText | |
| *maker* | foaf:maker | foaf:Agent | link to cfPerson (classification term = author) | we recommend to use foaf:Person or foaf:Organization |
| *homepage* | foaf:homepage | foaf:Document | cfResultPublication.cfURI | |
| *primary topic* | foaf:primaryTopic | owl:Thing | cfResultPublication.cfKeywords | primary topic and topic can be joined in cfKeywords as comma separated values |
| *topic* | foaf:topic | owl:Thing | cfResultPublication.cfKeywords | primary topic and topic can be joined in cfKeywords as comma separated values |

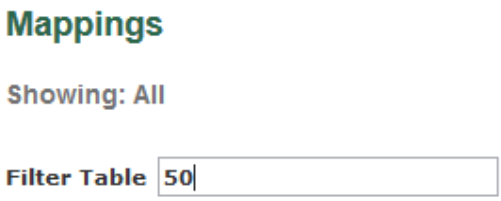
## 9.4 Annexe 4: CKAN matching and mapping

Part of mapping between CKAN and CERIF expressed in the 3M tool is shown on the following image.

| # | | SOURCE | TARGET | CONSTANT EXPRESSION |
|---|---|---|---|---|
| 1 | D | ../result | Product | |
| 1.1 | P | id | *has_identifier*<br>FederatedIdentifier<br>*has_id_value* | |
| | R | id | string | |
| 1.2 | P | title | *has_name* | |
| | R | title | PlainLiteral | |
| 1.3 | P | url | *has_URI* | |
| | R | url | string | |
| 1.4 | P | notes | *has_description* | |
| | R | notes | PlainLiteral | |
| 1.5 | P | name | *has_identifier*<br>FederatedIdentifier<br>*has_id_value* | |
| | R | name | string | |
| 1.6 | P | revision_id | *has_identifier*<br>FederatedIdentifier [revisionFedId]<br>*has_id_value* | |
| | R | revision_id | string | |
| | | | *is_destination_of*<br>LinkEntityWithProvisionOrderIPR<br>[authorPersonProduct] | [has_classification] = [Classification = "author"]<br>[has_startDate] = [date = ""]<br>[has_endDate] = [date = ""] |

To view the full mapping available in 3M:

| | |
|---|---|
| • Open a web browser at the following location: http://www.ics.forth.gr/isl/3M-VRE4EIC |  |
| • Fill in the username with and the password with the following credentials and click on the Log In button<br>  • Username: vre4eicGuest<br>  • Password: vre4eic |  |

| | |
|---|---|
| • Filter the table of mappings with the number 50 |  |
| • Click on Mapping/50 | |
| • Click on the view icon in the top menu | |

A new window opens with the full mapping description.

The following tables describe the matching for the common elements identified in chapter 3.

*Dataset*

Mappings of CKAN Package elements which represent datasets to the CERIF elements are shown in the following table.

| CKAN element | CERIF | Comments |
|---|---|---|
| *id* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier=cfResultProduct.cfResultProductIdentifier, classified using cfFederatedIdentifier_Classification entity where classification is "result product identifier" | Dataset ID |
| *revision_id* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier=cfResultProduct.cfResultProductIdentifier, classified using cfFederatedIdentifier_Classification entity where classification is "result product revision identifier", and cfFederatedIdentifier_Classification. cfStartDate is value of the CKAN attribute "metadata_modified" | ID of the last revision for the dataset object (doesn't include tags, groups, extra fields, relationships) |

| CKAN element | CERIF | Comments |
|---|---|---|
| *name* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier=cfResultProduct.cfResultProductIdentifier, classified using cfFederatedIdentifier_Classification entity where classification is "result product local name" | Dataset local name harvested from data provider |
| *title* | stored in cfResultProductName.cfName, with cfResultProductName.cfLangCode="en", cfResultProductName.cfTrans="o" | * cfTrans indicates if the text has been translated: "o" it is the original language, "m" it has been translated by a machine, "h" it has been translated by a human. |
| *url* | cfResultProduct.cfURI | Web home page for this dataset |
| *author* | should be split and stored in cfPersonName.cfFirstNames and cfPersonName.cfFamilyNames + connected with the certain dataset using cfPerson_ResultProduct | * Original author of the dataset<br>* A connection between cfPersonName and cfPerson entity should be established using the Link entity cfPersonName_Person which should be classified using the classification scheme "Person Names" and the term "Short Name"<br>* A connection between cfResultProduct and cfPerson entity should be established using the Link entity cfPerson_ResultProduct which should be classified using the classification scheme "Person Output Contributions" and the term "Author" |
| *author_email* | cfElectronicAddress.cfURI + cfPerson_ElectronicAddress | * A connection between cfPers and cfElectronicAddress entity should be established using the Link entity cfPerson_ElectronicAddress which should be classified using the classification scheme "Person Contact Details" and the term "Email" |

| CKAN element | CERIF | Comments |
|---|---|---|
| *maintainer* | should be split and stored in cfPersonName.cfFirstNames and cfPersonName.cfFamilyNames + connected with the certain dataset using cfPerson_ResultProduct | * Current maintainer or publisher of the dataset<br>* A connection between cfPersonName and cfPerson entity should be established using the Link entity cfPersonName_Person which should be classified using the classification scheme "Person Names" and the term "Short Name"<br>* A connection between cfResultProduct and cfPerson entity should be established using the link entity cfPerson_ResultProduct which should be classified using the classification scheme "Person Output Contributions" and the term "Publisher" |
| *maintainer_email* | cfElectronicAddress.cfURI + cfPerson_ElectronicAddress | * A connection between cfPerson and cfElectronicAddress entity should be established using the link entity cfPerson_ElectronicAddress which should be classified using the classification scheme "Person Contact Details" and the term "Email" |
| *type* | cfClassification + cfResultProduct_Classification | * define classification (name="Dataset Type") and all terms for this classification describing "type" values<br>*connect cfResultProduct and the cfClassification entity which represent the certain "type" value<br>* add value from the the CKAN attribute metadata_created as a startdate (cfResultProduct_Classification.cfStartDate), and we recemmend to add 2099-12-31T23:59:59-01:00 as an enddate, in case of unknown. |
| *version* | cfResultProduct.cfResultProductVersionInformation | |

| CKAN element | CERIF | Comments |
|---|---|---|
| *state* | cfClassification + cfResultProduct_Classification | * define classification (name="Dataset State") and three terms for this classification describing states: active, deleted, pending<br>*connect cfResultProduct and the cfClassification entity which represent the certain dataset state<br>* "We recommend to add 1901-01-01T00:0000-01:00 as a start date (cfResultProduct_Classification.cfStartDate), in case of unknown, and we recommend to add 2099-12-31T23:59:59-01:00 as an end date (cfResultProduct_Classification.cfEndDate), in case of unknown." |
| *notes* | stored in cfResultProductDescription.cfDescription with cfResultProductDescription.cfLangCode="en", cfResultProductDescription.cfTrans="o" | |
| *private* | cfClassification + cfResultProduct_Classification | *dataset with no organization can't be private (if owner_org is null, private is false)<br>* define classification (name="Dataset is Private") and two terms for this classification: true, false<br>*connect cfResultProduct and the cfClassification entity which represent the certain dataset private attribute value<br>* "We recommend to add 1901-01-01T00:0000-01:00 as a start date (cfResultProduct_Classification), in case of unknown, and we recommend to add 2099-12-31T23:59:59-01:00 as an end date (cfResultProduct_Classification), in case of unknown." |

| CKAN element | CERIF | Comments |
|---|---|---|
| *owner_org* | if (organization.id==owner_org) { the connection between the certain dataset and connected organization which is owner of dataset should be established using cfOrganisationUnit_ResultProduct} | * Organization should be stored using the mapping in the table CKAN Organization<br>* define classification name="Dataset Owner Organization Relations"<br>* The linked entity cfOrganisationUnit_ResultProduct should be classified using the classification "Dataset Owner Organization Relations" |
| *isopen* | cfClassification + cfResultProduct_Classification | * define classification (name="Dataset is Open") and two terms for this classification: true, false<br>*connect cfResultProduct and the cfClassification entity which represent the certain dataset isopen attribute value<br>* "We recommend to add 1901-01-01T00:0000-01:00 as a startdate (cfResultProduct_Classification), in case of unknown, and we recemmend to add 2099-12-31T23:59:59-01:00 as an enddate (cfResultProduct_Classification), in case of unknown." |
| *license_id* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier=cfResultPublication.cfResultPublicationIdentifier, classified using cfFederatedIdentifier_Classification entity where classification is "CERIF Entities" and term "Publication" | * A connection between cfResultPublication representing a license and the cfResultProduct representing a dataset should be established using the Link entity cfResultPublication_ResultProduct which should be classified using the classification scheme "Inter-Output Relations" and the term "Relation" |
| *license_title* | stored in cfResultPublicationTitle.cfTitle, with cfResultPublicationTitle.cfLangCode="en", cfProjTitle.cfTrans="o" | * A connection between cfResultPublicationTitle.cfTitle and cfResultPublication entity should be established using the foreign key cfResultPublicationId |
| *license_url* | cfResultPublication.cfURI | |

| CKAN element | CERIF | Comments |
|---|---|---|
| *resources* | the connection between the certain dataset and connected resources should be established using cfResultProduct_Medium, cfResultProduct_Service, cfResultProduct_ResultProduct, or cfResultProduct_ResultPublication | * Resource should be stored using the mapping in the table CKAN Resource<br>* define classification (name="Dataset Resource Relations") and a term for this classification: Relation<br>* The linked entity cfResultProduct_Medium, , cfResultProduct_Service, cfResultProduct_ResultProduct, or cfResultProduct_ResultPublication should be classified using the classification scheme "Dataset Resource Relations" and the term "Relation" |
| *groups* | connection between the certain dataset and connected groups should be established using cfResultProduct_ResultProduct | * Group should be stored using the mapping in the table CKAN Group<br>* The linked entity cfResultProduct_ResultProduct should be classified using the CERIF classification scheme "Inter-Product Relations" and the term "Part" |
| *tags* | cfClassification + cfResultProduct_Classification | * Tags should be stored using the mapping in the table CKAN Tag<br>* define classification (name="Dataset tag") and a term for each tag (see the mapping in the tabel for mapping CKAN Tag elements) |
| *extras [key = spatial]* | cfGeographicBoundingBox | If the extras key attribute has value “spatial”, the extras value attribute is matched with the cfGeographicBoundingBox multilingual entity cfGeographicBoundingBoxDescr |
| *extras [key = harvest_object_id]* | cfFederatedIdentifier | If the extras key attribute has value “harvest_object_id”, the extras value attribute is matched with the cfFederatedIdentifier and URI contains the value of key attribute |
| *extras [key = harvest_source_id]* | cfFederatedIdentifier | If the extras key attribute has value “harvest_source_id”, the extras value attribute is matched with the cfFederatedIdentifier and URI contains the value of key attribute |

| CKAN element | CERIF | Comments |
|---|---|---|
| *extras [key = harvest_source_title]* | cfFederatedIdentifier | If the extras key attribute has value “harvest_source_title”, the extras value attribute is matched with the cfFederatedIdentifier and URI contains the value of key attribute |
| *extras [key = guid]* | cfFederatedIdentifier | If the extras key attribute has value “guid”, the extras value attribute is matched with the cfFederatedIdentifier and URI contains the value of key attribute |
| *extras [key = contact-email]* | cfPerson_ResultProduct, cfPers, cfPerson_ElectronicAddress, cfElectronicAddress | If the extras key attribute has value “contact-email”, the extras value attribute is matched with the linked cfElectronicAddress entity and the established link is classified as “contact email” |
| *extras [key = spatial-reference-system]* | cfGeographicBoundingBox_Classification | If the extras key attribute has value “spatial-reference-system”, the extras value attribute is matched with the cfGeographicBoundingBox linked CERIF semantic layer entity cfClassification |
| *extras [key = metadata-date]* | cfResultProduct_Classification.cfStartDate | If the extras key attribute has value “metadata-date”, the extras value attribute is matched with the linked cfClassification entity and its cfStartDate attribute |
| *extras [key = resource-type]* | cfResultPublication_Classification or cfResultProduct_Classification or cfService_Classification or cfMedium_Classification | If the extras key attribute has value “resource-type”, the extras value attribute is matched with the linked cfClassification entity |
| *extras [key = dataset-reference-date]* | cfResultProduct_Classification.cfStartDate | If the extras key attribute has value “dataset-reference-date”, the extras value attribute is matched with the linked cfClassification entity and its cfStartDate attribute |

Mappings of the CKAN Tag elements intended to annotate dataset with tags to the CERIF elements are shown in the following table.

| CKAN element | CERIF | Comments |
|---|---|---|
| *id* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier=cfClassification.cfClassificationId, classified using cfFederatedIdentifier_Classification entity where classification "Tag identifier" | |

| CKAN element | CERIF | Comments |
|---|---|---|
| *name* | stored in cfClassificationTerm.cfTermSrc, with cfClassificationTerm.cfLangCode="en",cfClassificationTerm.cfTrans="o" | |
| *display_name* | stored in cfClassificationTerm.cfTerm, with cfClassificationTerm.cfLangCode="en",cfClassificationTerm.cfTrans="o" | |
| *state* | cfClassification + cfClassification_Classification | * define classification (name="Tag State") and three terms for this classification describing states: active, deleted, pending<br>*connect the certain cfClassification representing a tag and the cfClassification entity which represents the certain tag state<br>* "We recommend to add 1901-01-01T00:0000-01:00 as a start date (cfClassification_Classification.cfStartDate), in case of unknown, and we recommend to add 2099-12-31T23:59:59-01:00 as an end date (cfClassification_Classification.cfEndDate), in case of unknown." |
| *vocabulary_id* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier=cfClassification.cfClassificationId, classified using cfFederatedIdentifier_Classification entity where classification is "Vocabulary identifier" | |

Mappings of the CKAN Group element intended to group datasets to the CERIF elements are shown in the following table.

| CKAN element | CERIF | Comments |
|---|---|---|
| *id* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier=cfResultProduct.cfResultProductIdentifier, classified using cfFederatedIdentifier_Classification entity where classification is "CERIF Entities" and term | Group ID |

| CKAN element | CERIF | Comments |
|---|---|---|
| | "Product" | |
| *name* | stored in cfResultProductAlternativeName.cfAlternativeName, with cfResultProductAlternativeName.cfLangCode="en", cfResultProductAlternativeName.cfTrans="o" | * cfTrans indicates if the text has been translated: "o" it is the original language, "m" it has been translated by a machine, "h" it has been translated by a human. |
| *title* | stored in cfResultProductName.cfName, with cfResultProductName.cfLangCode="en", cfResultProductName.cfTrans="o" | |
| *description* | stored in cfResultProductDescription.cfDescription, with cfResultProductDescription.cfLangCode="en", cfResultProductDescription.cfTrans="o" | |
| *image_display_url* | stored in cfMedium.cfURI | * A connection between cfResultProduct representing a group and the cfMedium representing a group image should be established using the Link entity cfResultProduct_Medium which should be classified using the classification scheme "Media Relations" and the term "Logo" |

*Software, WebService and Publication*

Mappings of the CKAN Resource element which represent software (code), web service (file, api) or publication (documentation) to the CERIF elements are shown in the following table.

| CKAN element | CERIF | Comments |
|---|---|---|
| *id* | switch(resource_type)<br>case file, file.upload:<br>cfMedium.cfMediumId<br>case api:<br>cfService.cfServiceId<br>case visualization, code:<br>cfResultProduct.cfResultProductIdentifier<br>case documentation:<br>cfResultPublication.cfResultPublicationId | Resource Id |

| CKAN element | CERIF | Comments |
|---|---|---|
| *url* | switch(resource_type)<br>case file, file.upload:<br>cfMedium.cfURI<br>case api:<br>cfService.cfURI<br>case visualization, code:<br>cfResultProduct.cfURI<br>case documentation:<br>cfResultPublication.cfURI | |
| *name* | switch(resource_type)<br>case file, file.upload:<br>stored in cfMediumTitle.cfTitle, with cfMediumTitle.cfLangCode="en", cfMediumTitle.cfTrans="o"<br>case api:<br>stored in cfServiceName.cfName, with cfServiceName.cfLangCode="en", cfServiceName.cfTrans="o"<br>case visualization, code:<br>stored in cfResultProductName.cfName, with cfResultProductName.cfLangCode="en", cfResultProductName.cfTrans="o"<br>case documentation:<br>stored in cfResultPublicationTitle.cfTitle, with cfResultPublicationTitle.cfLangCode="en", cfResultPublicationTitle.cfTrans="o" | |
| *description* | switch(resource_type)<br>case file, file.upload:<br>stored in cfMediumDescription.cfDescr, with cfMediumDescription.cfLangCode="en", cfMediumDescription.cfTrans="o"<br>case api:<br>stored in cfServiceDescription.cfDescr, with cfServiceDescription.cfLangCode="en", cfServiceDescription.cfTrans="o"<br>case visualization, code: | |

| CKAN element | CERIF | Comments |
|---|---|---|
| | stored in cfResultProductDescription.cfDescr, with cfResultProductDescription.cfLangCode="en", cfResultProductDescription.cfTrans="o"<br>case documentation:<br>stored in cfResultPublicationAbstract.cfAbstract, with cfResultPublicationAbstract.cfLangCode="en", cfResultPublicationAbstract.cfTrans="o" | |
| *resource_type* | switch(resource_type)<br>case file, file.upload:<br>cfClassification + cfMedium_Classification<br>case api:<br>cfClassification + cfService_Classification<br>case visualization, code:<br>cfClassification + cfResultProduct_Classification<br>case documentation:<br>cfClassification + cfResultPublication_Classification | * define classification (name="Resource Type") and all terms for this classification describing "resource_type" values: file, file.upload, api, visualization, code, documentation<br>*connect cfMedium, cfService, cfResultProduct or cfResultPublication and the cfClassification entity which represent the certain "resource_type" value<br>* add value from the CKAN resource attribute "created" as a start date (*_Classification.cfStartDate), and we recommend to add 2099-12-31T23:59:59-01:00 as an end date, in case of unknown. |
| *format* | switch(resource_type)<br>case file, file.upload:<br>cfClassification + cfMedium_Classification<br>case api:<br>cfClassification + cfService_Classification<br>case visualization, code:<br>cfClassification + cfResultProduct_Classification<br>case documentation:<br>cfClassification + cfResultPublication_Classification | * define classification (name="Resource Format") and all terms for this classification describing "format" values<br>*connect cfMedium, cfService, cfResultProduct or cfResultPublication and the cfClassification entity which represent the certain "format" value<br>* We recommend to add 1901-01-01T00:0000-01:00 as a start date, in case of unknown, and we recommend to add 2099-12-31T23:59:59-01:00 as an end date, in case of unknown. |

| CKAN element | CERIF | Comments |
|---|---|---|
| *mimetype* | cfMedium.cfMimeType | * mimetype is define only if resource_type is file or file.upload |
| *mimetype_inner* | cfMedium.cfMimeType | * mimetype_inner is the mimetype of a 'contained' file. e.g. when resource.mimetype='application/zip', resource.mimetype_inner='text/csv'. If resource.mimetype is not a container, resource.mimetype_inner is empty<br>* mimetype and mimetype inner should be stored in cfMedium.cfMimeType in the format: "mimetype"; inner="mimetype_inner" |
| *size* | cfMedium.cfSize | * size is define only if resource_type is file or file.upload |
| *revision_id* | switch(resource_type)<br>case file, file.upload:<br>*id* = cfMedium.cfMediumId<br>*term* = Medium<br>case api:<br>*id* = cfService.cfServiceId<br>*term* = Service<br>case visualization, code:<br>*id* = cfResultProduct.cfResultProductId<br>*term* = Product<br>case documentation:<br>*id* = cfResultPublication.cfResultPublicationId<br>*term* = Publication<br>* stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier=*id*, classified using cfFederatedIdentifier_Classification entity where classification is "CERIF Entities" and term *term*, and cfFederatedIdentifier_Classification.cfStartDate is value of the CKAN attribute last_modified | ID of the last revision for the resource |

| CKAN element | CERIF | Comments |
|---|---|---|
| *state* | switch(resource_type)<br>case file, file.upload:<br>cfClassification +<br>cfMedium_Classification<br>case api:<br>cfClassification +<br>cfService_Classification<br>case visualization, code:<br>cfClassification +<br>cfResultProduct_Classification<br>case documentation:<br>cfClassification +<br>cfResultPublication_Classification | * define classification (name="Resource State") and three terms for this classification describing states: active, deleted, pending<br>*connect cfMedium, cfService, cfResultProduct or cfResultPublication and the cfClassification entity which represent the certain dataset state<br>* "We recommend to add 1901-01-01T00:0000-01:00 as a start date, in case of unknown, and we recommend to add 2099-12-31T23:59:59-01:00 as an end date, in case of unknown." |

*Organisation*

Mappings of the CKAN Organization elements to the CERIF elements are shown in the following table.

| CKAN element | CERIF | Comments |
|---|---|---|
| *id* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier= cfResultProduct.cfResultProductIdentifier, classified using cfFederatedIdentifier_Classification entity where classification is "Oragnization Identifier" | Organization ID |
| *title* | stored in cfOrganisationUnitName.cfName, with cfOrganisationUnitName.cfLangCode="en" , cfOrganisationUnitName.cfTrans="o" | * cfTrans indicates if the text has been translated: "o" it is the original language, "m" it has been translated by a machine, "h" it has been translated by a human. |
| *name* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier= cfResultProduct.cfResultProductIdentifier, classified using cfFederatedIdentifier_Classification entity where classification is "Organization local name" | Organization local name harvested from data provider |
| *description* | stored in cfOrganisationUnitResearchActivity.cfResearchActivity, with cfOrganisationUnitResearchActivity.cfLang | |

| CKAN element | CERIF | Comments |
|---|---|---|
| | Code="en", cfOrganisationUnitResearchActivity.cfTrans="o" | |
| *type* | cfClassification + cfOrganisationUnit_Classification | * define classification (name="Organization Type") and all terms for this classification describing "type" values *connect cfOrganisationUnit and the cfClassification entity which represent the certain "type" value * add value from the the CKAN attribute created as a startdate (cfOrganisationUnit_Classification.cfStartDate), and we recemmend to add 2099-12-31T23:59:59-01:00 as an enddate, in case of unknown. |
| *state* | cfClassification + cfOrganisationUnit_Classification | * define classification (name="Organization State") and three terms for this classification describing states: active, deleted, pending *connect cfOrganisationUnit and the cfClassification entity which represents the certain organization state * "We recommend to add 1901-01-01T00:0000-01:00 as a start date (cfOrganisationUnit_Classification.cfStartDate), in case of unknown, and we recommend to add 2099-12-31T23:59:59-01:00 as an end date (cfOrganisationUnit_Classification.cfEndDate), in case of unknown." |
| *revision_id* | stored in cfFederatedIdentifier.cfFederatedIdentifier where cfFederatedIdentifier.cfInstanceIdentifier= cfOrganisationUnit.cfOrganisationUnitId, classified using cfFederatedIdentifier_Classification entity where classification is "Organization revision identifier" | ID of the last revision for the dataset object (doesn't include tags, groups, extra fields, relationships) |

| CKAN element | CERIF | Comments |
|---|---|---|
| *approval_status* | cfClassification + cfOrganisationUnit_Classification | * define classification (name="Organization Approval Status") and two terms: true, false *connect cfOrganisationUnit and the cfClassification entity which represents the certain organization approval status * "We recommend to add 1901-01-01T00:0000-01:00 as a start date (cfOrganisationUnit_Classification. cfStartDate), in case of unknown, and we recommend to add 2099-12-31T23:59:59-01:00 as an end date (cfOrganisationUnit_Classification. cfEndDate), in case of unknown." |
| *image_url* | stored in cfMedium.cfURI | * A connection between cfOrganisationUnit representing a group and the cfMedium representing an organization image should be established using the link entity cfOrganisationUnit_Medium which should be classified using the classification scheme "Media Relations" and the term "Logo" |