A mapping tool for semantic interoperability of research information systems

Ivanović Dragan*, Theodoridou Maria**, Remy Laurent***
* University of Novi Sad, Novi Sad, Serbia
** FORTH-ICS, Heraklion, Greece
*** IS4RI - Information Systems for Research and Innovation, Strasbourg, France
dragan.ivanovic@uns.ac.rs
maria@ics.forth.gr
lremy@is4ri.com

***Abstract*— The paper presents mapping of CKAN metadata model to CERIF RDF using the X3ML Toolkit. The mapping enables interoperability of CKAN data platforms and virtual research environments based on CERIF format. X3ML Toolkit enables describing schema mappings in such a way that it can be collaboratively created and discussed by metadata format experts.**

I. INTRODUCTION

Research management systems, institutional publication repositories, research data repositories, and virtual research environments contain research information. Interoperability of those systems is necessary in order to avoid duplicate input of data and enable creation of a unique digital space of research information, which can be exploited by researchers, managers, entrepreneurs, citizens and government. Beyond the ability of two or more computers systems to exchange information, semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems. Implementation of mappings between various data formats and harmonization of vocabularies are necessary in order to reach interoperability between systems. Mapping of metadata and harmonization of vocabularies are error- and time-prone tasks that require a close collaboration between experts in metadata models and software developers. Misunderstandings are a common problem in such collaborations. A tool which will enable metadata formats experts to define mappings could address previously stated issues, by solving communication issues and automating some operations. In this paper, we used the X3ML mapping toolkit to define the mapping of CKAN to CERIF. This toolkit allows experts to use a common language to express mappings and help them with automation tools like auto-completion and automatic transformation features.

The VRE4EIC project addresses key data and software challenges in supporting multidisciplinary data driven sciences (www.vre4eic.eu). 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 harmonization process is needed. This process will allow heterogeneous data (or metadata for this project) to be stored using a common data model. Platforms for data repositories and dataset metadata formats were analyzed for the needs of the VRE4EIC project. CKAN platform is a dominant one in this area. In

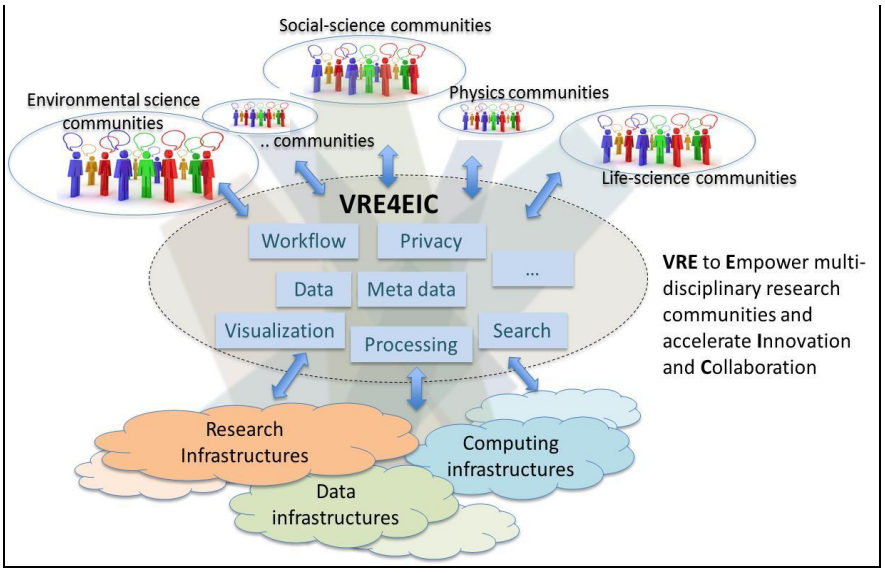

Figure 1. VRE4EIC project vision

the CKAN platform, datasets are described using an internal format. Metadata about datasets can be retrieved in JSON format using a REST API.

CERIF format is selected as the target research domain format because it is a conceptual metadata model which allows a representation of research entities, their activities and their output. It has high flexibility with formal (semantic) relationships, enables quality maintenance, archiving, access and interchange of research information covering all the research life-cycle. It is also a format recommended by European Commission for European Union countries research information systems. Some matching exists between CKAN and CERIF metadata format, expressed in tabular format [1]. Moreover, this matching doesn't use the actual version of CERIF. However, this matching is used as a starting point and updated to be in line with CERIF version 1.6. After that, the X3ML mapping tool is used to define mappings [3].

II. VRE4EIC

VRE4EIC (www.vre4eic.eu) develops a reference architecture and software components for VREs (Virtual Research Environments). One product of the VRE4EIC project is a prototype of VRE – e-VRE (enhanced Virtual Research Environments). The e-VRE platform increases the quality of VRE User Experiences by providing user centered, secure, privacy compliant, sustainable environments on searching data, composing workflows and tracking data publications. Moreover, it increases the deployment of the VRE on different clusters of research infrastructures by abstracting and reusing building blocks and workflows from existing VREs, infrastructures and projects. Also, e-VRE improves the contextual awareness and interoperability of the metadata across all layers of the resources in the VRE. Finally, it provides interoperation across 'silo' e-RIs.

III. X3ML Toolkit

The X3ML toolkit allows definition of a matching and 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. Matching has been defined as the process of finding relationships or correspondences between entities of different ontologies [4]. 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 [5]. 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. The X3ML toolkit is based on X3ML language, an XML based language, describing schema matchings in such a way that it can be collaboratively created and discussed by metadata format experts. The key components of the toolkit are:

a) Mapping Memory Manager, a tool for managing mapping definition files providing a number of administrative actions.

b) 3M Editor, a web application suite that assists users during the mapping definition process, using a human-friendly user interface and a set of sub-components that either suggest or validate user input.

c) X3ML Engine, a tool that realizes the transformation of the source records to the target format.

IV. CKAN

CKAN is a web-based open source and open architecture software platform for data management. This platform is in use by numerous governments, organizations and communities around the world. The CKAN platform can be easily installed, customized and extended for the specific needs of some organisation. Moreover, the CKAN platform can preserve various data types – datasets, source codes, documentations, etc.

There are three basic entities in the CKAN metadata model: Package, Resource, and Group. Also, there is the Organization entity which stores data about the institution being the data publisher. Besides these entities, CKAN also enables adding "tags" with predefined sets of values to a package entity and key-value pairs of additional package.

A simple example of CKAN JSON record is shown in the Listing 2.

```
{
 …
 "result":
 {
  "type": "dataset",
  "title": "UK: Adur District Council Spending Data",
  "author": "John Smith",
  "author_email": "john.smith@gmail.com"
 }
}
```

Listing 2. A CKAN JSON record

V. CERIF

The Common European Research Information Format (CERIF) is a flexible and a rich data model for representing information about research. The model was developed and is being maintained by euroCRIS (www.eurocris.org). The primary aim for the CERIF model is to support information interchange in the research domain. The various upgrades and extensions of the model are led by the CERIF Task Group. The CERIF model contains about 25 conceptual entities in the current version (CERIF 1.6 – Figure 2) dealing with the different concepts involved in research information:

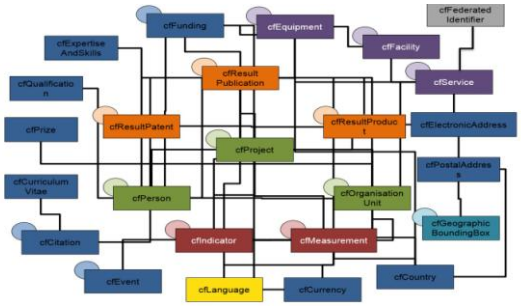

Figure 2. CERIF high level entities

- *Basic concepts:* persons, projects, organisation units
- *Concepts linked to scientific results:* products, patents, publications

- *Concepts linked to research infrastructure:* services, facilities, equipments
- *Indicators and measurements*
- *Federated identifiers*
- *And several other additional concepts*: funding, addresses, geographic bindings, languages, etc.

An encoding of CERIF in XML has been defined and used for the interoperability among CERIF-compliant systems [6]. The core technology for a wide spread, distributed and structured service for data is the semantic web technology in 2010s. According to the W3C, "*The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries*" [7]. One of the best ways to integrate and publish research information is to use the open W3C Web standard RDF (Resource Description Framework).

Definition of the CERIF RDF encoding was carried out in the context of the VRE4EIC project where a CERIF based semantic metadata catalogue has been built. It resulted from a bottom-up approach of the transformation of the relational structure to an ontological structure. 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 vocabularies. The CERIF research entities have been transformed into RDF classes and their attributes into properties.

The Listing 1 shows a simple example of CERIF RDF Product class with two properties (has_identifier and has_name) in the Turtle format.

```
<http://www.vre4eic.com/ckancerif/resultProduct/5843a7f8-f74e-48be-
918e-7b146b73e245>
    a                          cerif:Product ;
    cerif:has_identifier
<http://www.vre4eic.com/ckancerif/resultProductIdentifier/5843a7f8-
f74e-48be-918e-7b146b73e245> ;
    cerif:has_name        "UK: Adur District Council Spending Data" ;
```

Listing 1. A CERIF RDF record

The adoption by euroCRIS of this representation as official CERIF RDF expression is in progress.

VI. MAPPING

CKAN metadata used for definition and testing mappings have been retrieved from Open UK Government Data platform using CKAN REST API (https://data.gov.uk/api/3/action/package_list). CKAN JSON format has been converted to XML format and used as input source in 3M Editor of X3ML Toolkit. The previously discussed CERIF RDF encoding is used as the output format.

Flexibility of the CKAN platform makes much more complicated the mapping 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. 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.

Basic set of CKAN metadata representing a CKAN Package is mapped to CERIF Product instance. Subset of metadata grouped within the CKAN organization field is mapped to a CERIF OrganisationUnit instance which is linked to the previously mentioned Product instance. Furthermore, subset of metadata grouped within the CKAN resource field is mapped to a CERIF OrganisationUnit instance which is linked to the Product instance representing package. Moreover, CKAN tags are mapped using the semantic layer of CERIF model, i.e. using CERIF Classification instances. The 10 most used "extras" elements found in 133 CKAN instances all over the world [2] are mapped to CERIF RDF as it is shown in the table 1.

The mapping of CKAN metadata model to CERIF RDF is expressed in the X3ML Toolkit and the full mapping is accessible by opening the link http://www.ics.forth.gr/isl/3M-VRE4EIC, logging in using username vre4eicGuest and password vre4eic, finding mapping project with id 50 and selecting the icon .

There are a few basic options of X3ML menu for a mapping project.

In the Info section of the mapping we :

- uploaded as source a CKAN example which we previously transformed from JSON to XML format,
- uploaded as target the CERIF RDF,
- defined generators which we used in our mapping, and
- uploaded sample data which we use for verification of our mapping in the Transformation section.

In the Matching Table section we defined mapping of input source (CKAN) to the target output (CERIF RDF). A part of mapping definition using X3ML is shown in Figure 3. The first step is to define domain (the D line in Figure 3) by specifying source and target node. After that, path (the P line) correlation between source and target

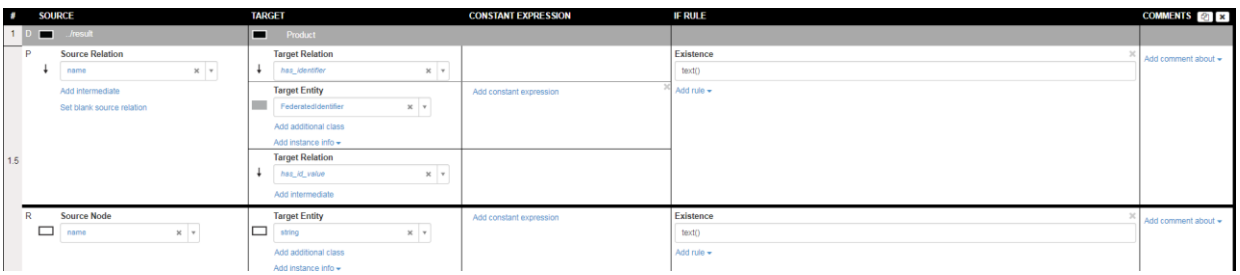

Figure 3. X3ML Matching table

format should be defined. This path correlation definition could include additional entities if it is needed (Add constant expression in Figure 3), as well as definition of some preconditions (if rules). At the end, mapping between range (the R line) of source and target node should be defined.

In the Generators section we assigned a generator for each entity which is a result of a mapping rule defined in the section Matching Table. Figure 4 shows assignment of generators for mapping shown in Figure 3. A generator is used to create a value: a literal, a constant, a URI or any other kind of value.

At the end, we can run the transformation and validate our mapping in the Transformation section. It is possible to define the output format: RDF/XML, N-triples, or

Figure 4. X3ML generators assignment

Turtle. Listing 2 shows part of result Turtle output format corresponding to mapping and assignment of generator shown in Figures 3 and 4.

```
<http://www.vre4eic.com/ckancerif/resultProduct/5843a7f8-f74e-48be-
918e-7b146b73e245>
      a                  cerif:Product ;
      cerif:has_identifier
         <http://www.vre4eic.com/ckancerif/resultProductName/housing-
affordability-data-system-hads>
…

<http://www.vre4eic.com/ckancerif/resultProductName/housing-
affordability-data-system-hads>
      a                  cerif:FederatedIdentifier;
      cerif:has_id_value "housing-affordability-data-system-hads" .
```

Listing 2. A CERIF RDF output

VII. CONCLUSION

Mapping of CKAN metadata model to CERIF RDF using the X3ML Toolkit has been presented in this paper. The mapping has been implemented for the needs of VRE4EIC project. Usually, a mapping of metadata formats is the result of a close collaboration between experts in metadata model and software developers. However, there could be misunderstanding in this collaboration and representing mapping in notation of some programming language is error- and time-prone. We used X3ML mapping tool to define mapping described in this paper in order to overcome previously stated issue. This tool enables metadata formats experts to define mappings. The tool can be also useful for verification of mapping, because transformation of a source sample data can be run and the target output can be shown in various formats. Moreover, there is Mapping Analyzer (Maze), a web-based tool which undertakes to serve experts in order to provide a complete management of mappings. Maze works as intermediary between expert users and X3ML language, providing a complete analysis for converting and publishing content as linked data. This tool enables experts to improve their mappings.

As further work, we will continue to use X3ML tool to define mappings of other metadata formats (DCAT-AP, DublinCore, ISO 19139) to CERIF RDF for the needs of creation of a unique research dataset catalogue which is one goal of the VRE4EIC project.

ACKNOWLEDGMENT

This work has been carried out within the VRE4EIC project and has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 676247.

REFERENCES

[1] Houssos, Nikos, Brigitte Jörg, and Brian Matthews. "A multi-level metadata approach for a Public Sector Information data infrastructure." In *Proceedings of the 11th International Conference on Current Research Information Systems*, pp. 19-31. 2012.

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

[3] Marketakis, Y, Minadakis, N, Kondylakis, H, Konsolaki, K, Samaritakis, G, Theodoridou, M, Flouris, G & Doerr, M 2016 X3ML Mapping Framework for Information Integration in Cultural Heritage and Beyond. *International Journal on Digital Libraries (IJDL), Special Issue on "Extending, Mapping and Focusing the CIDOC CRM"*, 1-19.

[4] Euzenat, J. and Shvaiko, P., 2007. *Ontology matching* (Vol. 18). Heidelberg: Springer.

[5] Do, H.H., 2017. *Schema matching and mapping-based data integration* (Doctoral dissertation, Universität Leipzig).

[6] Jörg, B., Dvořák, J. and Vestdam, T., 2012. Streamlining the CERIF XML Data Exchange Format Towards CERIF 2.0.

[7] Lassila, O. and Swick, R.R., 1999. Resource description framework (RDF) model and syntax specification.

TABLE I.
CKAN THE MOST USED EXTRAS ELEMENTS

CKAN element	CERIF element	Note
Package	Product	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]	Publication	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)]	Product	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]	Service	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 = api]	Service	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)]	Medium	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
Group	Product	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	Classification	The CKAN Tags are matched using the CERIF semantic layer and its cfClassification entity
Extras [key = spatial]	GeographicBoundingBox	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]	FederatedIdentifier	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]	FederatedIdentifier	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_ti tle]	FederatedIdentifier	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]	FederatedIdentifier	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 = contactemail]	Person_Product, Person, Person_ ElectronicAddress, ElectronicAddress	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 = spatialreference-system]	GeographicBoundingBox_ Classification	If the Extras key attribute has value "spatial-referencesystem", the Extras value attribute is matched with the cfGeographicBoundingBox linked CERIF semantic layer entity cfClassification
Extras [key = metadata-date]	Product_ Classification.startDate	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]	[Publication or Product or Service or Medium]_Classification	If the Extras key attribute has value "resource-type", the Extras value attribute is matched with the linked cfClassification entity
Extras [key = datasetreference-date]	Product_ Classification.startDate	If the Extras key attribute has value "dataset-referencedate", the Extras value attribute is matched with the linked cfClassification entity and its cfStartDate attribute
Organization	OrganisationUnit	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