CONNECTING CLOSED WORLD RESEARCH
INFORMATION SYSTEMS THROUGH
THE LINKED OPEN DATA WEB

BRIGITTE JOERG
German Research Center for Artificial Intelligence (DFKI GmbH)
Alt Moabit 91c, 10559 Berlin, Germany
brigitte.joerg@dfki.de

IVAN RUIZ-RUBE
Department of Computer Languages and Systems
University of Cádiz, Cádiz 11002, Spain
ivan.ruiz@uca.es

MIGUEL-ANGEL SICILIA
Computer Science Department, University of Alcalá
28871 Alcalá de Henares, Spain
msicilia@uah.es

JAN DVOŘÁK
Institute of Information Studies and Librarianship
Charles University of Prague, U Kráše 8
CZ-15800 Praha 5 — Jinonice, Czech Republic
jan.dvorak@ff.cuni.cz

KEITH JEFFERY
Science and Technology Facilities Council (STFC)
Rutherford Appleton Laboratory, Didcot, Oxfordshire OX11 0QX, UK
keith.jeffery@stfc.ac.uk

THORSTEN HOELLRIGL
AVEDAS AG, Waldstraße 65, 76133 Karlsruhe, Germany
T.Hoellrigl@avedas.com

HENRIK S. RASMUSSEN
Atira A/S, Niels Jernes Vej 10, 9220 Aalborg Ost, Denmark
hsr@atira.dk

ANDREAS ENGFER
AVEDAS AG, Waldstraße 65, 76133 Karlsruhe, Germany
A.Engfer@avedas.com
Research Information Systems (RIS) play a critical role in the sharing of scientific information and provide researchers, professionals and decision makers with the required data for their activities. Existing RIS standards have proposed data models to represent the main entities for storage and exchange. These account for the needs of multiple stakeholders through a high flexibility based on a formal syntax and declared semantics, but for techno-historical reasons they assume the completeness of information within system boundaries. The distributed nature of research information across systems calls for a mechanism to link the local entities from the closed world of concrete RISs with other possibly underspecified entities exposed through other means, as for example, the Linked Open Data Web. By transformation of a relational model into an open graph model, differences between the two system paradigms are revealed. The main principles and techniques for exposing CERIF-driven relational data as linked data will be provided as a first step demonstrating effective RISs interconnection through the linked open data (LOD) Web.

Keywords: Current Research Information Systems; CRIS; Common European Research Information Format; CERIF; Linked Open Data; LOD; LD; Ontologies; Conceptual Modeling.

1. Introduction

Research Information Systems (RIS) provide multiple stakeholders with the required data for their activities and therefore play a critical role in information sharing [1–5]. Existing standards provide suitable data models to represent the main research entities and their relationships for storage and exchange [6, 7]. These account for the needs of multiple stakeholders through a high flexibility by means of their closed world extensions based on a formal syntax and declared semantics [8]. But for techno-historical reasons, they assume the completeness of information within well-defined system boundaries. Increasingly, the distributed nature of research information across systems will require a mechanism to link local system entities with external and possibly underspecified entities exposed through mechanisms as the Linked Open Data Web [9]. Interlinking and sharing are effectively enabled through semantic clarity and therefore require foundations. These become of even more importance with the changing nature of science, crossing disciplines and producing ever-growing amounts of data and information [10–13]. Obviously, foundations [14–16] and standardizing efforts are becoming crucial and are increasingly accepted and applied during software and knowledge engineering processes. Numerous
information system models have been designed for particular applications. Unfortunately, they often lacked foundations or empirical evidence and therefore scalability. Besides formal models for representing the research domain like CERIF\textsuperscript{a} \cite{17} or VIVO \cite{18}, CASRAI\textsuperscript{b} develops vocabularies for research administration information, CYC \cite{19} extracts and learns research knowledge automatically on a large scale, and there exist models and reference models in the narrower and wider context of research, for scholarly repositories, digital libraries or open archival information systems, often managing inter-linkage or exchange through APIs as OAI-PMH\textsuperscript{c} or simply by file sharing. Recent technological developments with the Web and its extensions allow for an open linkage and deposit of data and information with the Linked Open Data (LOD)-Cloud, based on the common representational foundation provided by RDF \cite{9}. Connecting closed world systems through a naturally inherent open graph structure such as RDF \cite{20} requires model extensions that allow for a qualified open linkage.\textsuperscript{d}

We present here real world use cases to introduce the research domain before presenting the CERIF RIS model with its inherent entities, patterns and structures — representing an extensible closed world — to motivate the need for a qualified open linkage between RISs (Sec. 2). Then, a proposed CERIF model extension enabling system inter-connections through the LOD Web is described in Sec. 3. In Sec. 4, the main differences between the two system paradigms are revealed through the semi-automated transformation of the relational CERIF RIS model into a CERIF ontology within the context of the FP7-funded Virtual Open Access Agriculture & Aquaculture Repository (VOA3R) project, viewing the research domain with a linked open data eye. We conclude in Sec. 5 by presenting references to related work that may be of interest towards directing future work.

\textbf{2. Research Information: Entities–Stakeholders–Use-Cases}

For many decades, research information has mostly been interpreted as output in the format of publications. Increasingly however, the importance of research-related information in a wider context is recognized \cite{15,1,12}. In terms of modeling and interoperation this implies the consideration of additional entities beyond publication, such as funding, organization, project, person, data, patent, impact, or teaching, and variations in their changing relationships through time. The added value of rich and standardized models like CERIF is increasingly acknowledged \cite{21,22,23} and additionally reflected in the significant uptake through vendors. To

\textsuperscript{a} The Common European Research Information Format (CERIF); a EU Recommendation to Member States: http://cordis.europa.eu/cerif/, http://www.euroCRIS.org/

\textsuperscript{b} Consortia Advancing Standards in Research Administration Information (CASRAI): http://casrai.org/

\textsuperscript{c} Open Archive Initiative — Protocol for Metadata Harvesting (OAI-PMH): http://www.openarchives.org/OAI/openarchivesprotocol.html

\textsuperscript{d} Linkage between closed world systems is thus enabled, however, the LOD Web is at current mostly a rich means for data exposition and deposit, where authority and access patterns still have to be developed and where the primary data are still collected, maintained and declared in local systems.
account for openness and to ensure the timely, contextual and multilingual scalability within closed-system boundaries as it is required through the dynamics in Science, the CERIF model distinguishes between three kinds of modeling constructs (Fig. 1): research entities, link entities and multilingual entities. Research entities such as person, project, organization or publication host attributes such as identifier, acronym, gender, or ISBN; link entities account for relationships and their changes through time; multilingual entities allow for multiple language recordings [24].

In research, the reuse of information and data (e.g. for dissemination, exchange, evaluation) is crucial not least because research activities to a large extent are supported by public funding. The urgent need for sharing is obvious from the views of involved stakeholders as demonstrated in the following examples:

(1) Researchers, Research Organizations, and Funding Bodies require interlinking between publications within RISs and institutional repositories, e.g. via a Digital Object Identifier (DOI), where the DOI uniquely identifies a publication in both systems and as such links the two publication records, which can then be uniquely identified as being the same work.

(2) Researchers and Research Organizations need to interlink RISs and open publication databases e.g. PubMed, which all have their own publication identifiers. Enabling multiple qualified identifiers of the same work supports identification of duplicates and richer representations through merged metadata from different sources, resulting in improved data quality for the benefit of everyone.

(3) Researchers and Research Organizations will notice, that publication databases increasingly introduce unique identifiers for authors. In RISs these identifiers should also be connected to person records and as such be employed for comparison of author records from different publication databases. The same holds for citation information related to authors.

(4) Researchers, Research Organizations, and Funding Bodies will learn that research infrastructure portals such as Tools of Science allow for queries within a rich inventory of collected research infrastructures providers, where RISs should

---

Fig. 1. Three kinds of CERIF model constructs.

---


Tools of Science: [http://www.toolsofscience.se/](http://www.toolsofscience.se/)
provide a mechanism for linkage of local infrastructure-related information with such rich collections.

(5) Researchers, Research Organizations, and Funding Bodies highly benefit from large knowledge organization systems (KOS), such as AGROVOC,⁸ which are increasingly deployed in the LOD Cloud. Linking internal RIS records with the LOD’s AGROVOC vocabulary terms allows for a consistent and open record categorization without internal maintenance needs. LOD published ontologies and vocabularies are globally understood and re-used and allow for cross-system comparison; queries are supported through uniform resource identifiers (URIs).

(6) People in general want to know how the budgets allocated by Funding Bodies for funded research projects are finally turned into results, such as products, patents and papers. This also fosters the debate about the need for public institutions to open their data to citizens giving access for analysis and reuse.

From presented use-cases and preceding discussions we identified a threefold pattern for RIS linkage:

- RIS connecting to RIS,
- RIS connecting to a non-RIS system,
- RIS connecting to a KOS / Authority File.

We propose a (CERIF)-RIS model extension to enable the identified connection patterns through the LOD Web.

3. The CERIF Model and its LOD Extension

CERIF is a so-called conceptual model of the research domain; it identifies the main entities and defines them by their basic attributes and through the relationships they maintain with each other (Fig. 2). For historical reasons CERIF was developed and described in relational terms — although object-oriented, deductive database, information retrieval and hypertext implementations can be constructed — so that objects are identified through internal identifiers.

The closed world of CERIF assumes for publication records an internal identifier cfResPubId and a publication date cfResPublDate as mandatory attributes NN. Attributes such as number, volume, edition or cfNum, cfVol, cfEdition are defined but not mandatory. Missing attribute values in closed world systems entail their non-existence, whereas an open world assumes them to be unknown. In CERIF, with multilingual entities such as cfClassTerm and cfResPublTitle the title or the term values are mandatory, as well as their language and translation information.

In closed world systems, objects are identified through internal identifiers, most often composed of additional attributes constructing their primary key (PK). A

⁸AGROVOC is a comprehensive multilingual agricultural vocabulary belonging to the Agricultural Information Management Standards (AIMS), hosted by the Food and Agriculture Organization of the United Nations (FAO): http://aims.fao.org/website/AGROVOC-Thesaurus/sub.
primary key may contain identifiers of other system-internal entities — primary foreign keys (PFK). The identification of objects in closed world systems is thus different from the LOD world where objects are identified through globally valid and unique URIs. A RIS extension under closed world assumption must account for such differences and take note of openness with respect to any information that is in addition to an URI identifier.

As indicated before, CERIF distinguishes research, link, and multilingual entities. The suggested model extension requires knowledge over link entities because their structure partly applies to the structure of the proposed extension entity cfCERIF_LOD (described below). CERIF specifies relationships as entities so-called link entities, whose objects are identified through a composite identifier from the two linked entities identifiers. CERIF link entities are open and scalable in the sense that they do not maintain explicit role or type names, but are labeled with generic relationship names indicating the two linked entities (e.g. person-publication cfPers_ResPub, person-project cfProj_Pers). Explicit roles or types are then
expressed by identifier references (cfClassId, cfClassSchemeId) to the so-called CERIF Semantic Layer, a powerful classification mechanism. CERIF link entities map to RDF in that they also maintain kind of a triple structure.

Table 1 presents the generic structure of CERIF link entities and some examples. In a closed world system, the order of identifiers is fixed according to the underlying model definition (see also Fig. 1) if both of the identifiers are part of the primary key. In the CERIF model, the reading direction given through the name of the linking entity does not mean anything. In fact it only indicates a neutral linkage and is used in both reading directions, e.g. in row 2, person is linked with project, or in row 3, person is linked with organization. Each embodied relationship is furthermore classified through an assignment with a scheme. In Table 1, the role coordinator (via the cfClassId = “cf-coordinator-uuid”) is assigned to an activity classification scheme (via the cfClassSchemeId = “cf-activity-uuid”). In CERIF link entities, the classifications and corresponding scheme are part of the primary key as well as the timestamps indicated in column 4; all contribute to the composite primary key construct — for identification. It is important to note that a neutral reading direction requires neutral classification terms. To indicate the reading direction, CERIF therefore additionally allows for role expressions within the Semantic Layer. Relationships between research entities, that is, link entities in CERIF, are temporally based allowing for the provision of valid link time intervals.

Since its inception, the Web aimed at building a global community to share information, introducing the hypertext mechanism and enabling the linkage of documents by URLs. The vision of the Semantic Web towards a Web beyond documents, is through linkage of data by applying a basic set of principles [9], namely: use URIs as names for things; use HTTP URIs so that people can look up those names; when someone looks up a URI, provide useful information, using standards (RDF, SPARQL); and finally include links to other URIs.

The CERIF model extension allows for connection through the Linked Open Data world [9]. Structure-wise, an extension entity therefore has to be RDF compliant and support a triple structure. If we follow the examples from our motivation, the first asks for interconnecting RIS, implying that the linked objects can be of the same kind e.g. connecting the RIS’s project record with its sister record in the CORDIS RKB SPARQL Explorer through the project’s URI. A new cfCERIF_LOD entity has to

Table 1. Generic CERIF Link Entity structure (column 1), explicit examples in columns 2–4.

<table>
<thead>
<tr>
<th>CERIF Link Entity Name</th>
<th>Composite Identifier</th>
<th>Relationship Classification</th>
<th>Timestamps</th>
</tr>
</thead>
<tbody>
<tr>
<td>cfEntity1, Entity2</td>
<td>cfEntity1Id, cfEntity2Id</td>
<td>cfClassId, cfClassSchemeId</td>
<td>cfStartDate, cfEndDate</td>
</tr>
<tr>
<td>cfPers, Proj</td>
<td>cfPers-miguel-angel,sicilia-ID</td>
<td>cf-coordinator-uuid, cf-activity-uuid</td>
<td>2005-04-01T00:00:00, 2007-22-30T00:00:00</td>
</tr>
<tr>
<td>cfPers, OrgUnit</td>
<td>cfPers-miguel-angel,sicilia-ID</td>
<td>cf-board-member-uid, cf-activity-uuid</td>
<td>2012-01-01T00:00:00, 2014-12-30T00:00:00</td>
</tr>
</tbody>
</table>

hhttp://cordis.rkbexplorer.com/sparql/.
support the LOD underlying triple structure as well as the common CERIF link entity structure (which is also kind of a triple) — and as such has been developed with the construct in Table 2.

The first column in Table 2 shows the generic triple structure of the cfCERIF_LOD entity connecting internal RIS instances through predicates with objects of the external world. The second column shows the defined cfCERIF_LOD attributes upon which the third column is populated with data. Interconnecting objects requires internally the system internal instance identifier cfInstanceId = cfPers-miguel-angel_sicilia-ID and the entity type e.g. cfEntity = cfPerson preferably inline with a known ontology or scheme; Through the predicate we connect — inline with the CERIF link mechanism employing time-stamped classification and scheme references. For the cfCERIF_LOD entity the mechanism has been enhanced by cfPredicateClassId and cfPredicateClassSchemeId. Labels. For the linking function e.g. “professor” we employ terms from the CERIF1.3 Semantics. The example shows the link to an external entity of type organization, indicated by the class identifier cfClassId = cf-organisation-uuid, which belongs to the CERIF1.3 Semantics as indicated by cfClassSchemeId = cfCERIF1.3-all-uuid. External objects are identified by an object URI cfxObjectURI = http://otherRIS.org/resource/orgUnit/CarlosIII and a source URI cfxSourceURI = http://otherRIS.org. The specification of the cfCERIF_LOD entity presented in Table 2 allows for linkage from internal CERIF entity instances to external objects through predicates, and thus allows for the representation of a typical RDF triple structure:

\[
\langle \text{system-internal-object-key} \rangle \text{subject} \\
\langle \text{optionally-time-stamped-predicate} \rangle \text{predicate} \\
\langle \text{external-object-uri} \rangle \text{object}
\]
4. Transforming CERIF Data into Linked Open Data

CERIF is usually structured as “extended relational” representing a fully connected graph, where the linked open data world builds on an open graph structure, namely RDF [20]. The exposition of CERIF-driven LOD therefore requires a transformation of the underlying data model. That is, the relational schema needs to be transformed into a formal ontology [25]. The transformation process is not a straightforward mapping, but requires additional decisions towards the implementation of controlled vocabularies. The model transformation process will briefly be presented to demonstrate the difference between RIS and LOD structures. This is important to make sure that no information is lost. The then resulting ontology will provide insight into the graph structure in line with the VOA3R use case. Several examples of publication and consumption of LOD will be presented.

4.1. CERIF Ontology derived from relational CERIF 1.3 Schema

The publication of linked data required the definition of an ontology towards relating the CERIF entities conceptually. The current ontology resulted from a bottom-up approach of the transformation of a 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 CERIF1.3 Semantics have been translated into a vocabulary. The transformation was guided by a mapping, which at a meta-level is indicated in Table 3.

The CERIF 1.3 RDFS ontology was designed with the tool Neologism\(^1\) which will later be deployed at the euroCRIS website.\(^1\) The ontology incorporates research and link entities. The multilingual entities are not explicitly mapped into ontology classes, because RDF allows for enriching property values with used language information. A snapshot of the transformed ontology is presented in Fig. 3, employing additional vocabularies and formats such as FOAF\(^k\) or BIBO,\(^l\) and Dublin Core. It is

<table>
<thead>
<tr>
<th>Relational entity</th>
<th>RDF resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual entity</td>
<td>Abstract class</td>
</tr>
<tr>
<td>Research entity</td>
<td>Concrete class</td>
</tr>
<tr>
<td>Link entity</td>
<td>Concrete class</td>
</tr>
<tr>
<td>Multilingual entity</td>
<td>Multilingual property</td>
</tr>
<tr>
<td>Attribute</td>
<td>Property</td>
</tr>
<tr>
<td>CERIF roles (semantics)</td>
<td>Subproperties of generic relationship properties</td>
</tr>
<tr>
<td>CERIF types (semantics)</td>
<td>Subclasses of Concrete Class (from research entities)</td>
</tr>
</tbody>
</table>

\(^1\)Neologism: http://neologism.deri.ie/
\(^1\)CERIF Ontology: http://www.eurocris.org/…
\(^k\)Friend of a Friend Vocabulary: http://xmlns.com/foaf/spec/
\(^l\)Bibliographic Ontology: http://bibliontology.com/specification
a paradigm [26] in the LOD world to reuse existing elements from well-known and globally accepted ontologies and vocabularies.

The CERIF research entities have been transformed into RDF classes and their attributes into properties. For the class names we used conceptual names (e.g. person or project) instead of the physical CERIF table names (cfPers, cfProj), because there is no longer a limitation imposed by any database, and linked data are semantically inspired. The resulting properties and classes are described through common annotation properties like `rdfs:label` and `rdfs:comment` in the ontology. In addition to concrete RDF classes (person, project), we added a number of abstract classes to group the class concepts inline with CERIF specification documents, namely: Base, Result, SecondLevel, and Infrastructure; these share a `cerif:internalIdentifier` property.

In addition to the above ontology, we defined a controlled vocabulary to gather the relevant types and roles in the research context of the involved VOA3R entities. The vocabulary includes several RDF classes describing CERIF classification schemes, e.g. types of organizational units, as well as a number of RDF properties semantically enriching the relationships between CERIF entities, like roles of a person in a research project. The CERIF vocabulary will be available from the euroCRIS website.\footnote{CERIF Semantic Vocabulary: http://eurocris.org/semcerif}

4.2. The VOA3R case — transforming relational data into LOD

Due to the large coverage of the entire CERIF model, we concentrate on the VOA3R case study to demonstrate the exposition and inter-linkage of CERIF-driven Linked Open Data. The VOA3R project aims to improve the dissemination of European
agriculture and aquaculture research results through the re-use of metadata and
semantic technologies, enabling to share openly available research products in the
field. In addition to the collection of publications managed by the underlying
VOA3R platform, the project was in need of a database to capture the context
around publications through information about entities such as persons, projects or
organizational units. A CERIF-driven database has therefore been embedded. For
the exposition of corresponding VOA3R data in the LOD Cloud a transformation of
the information stored in the extended relational CERIF structure into an open RDF
graph structure was required. VOA3R maintains information about the entities:
project, person, organisation, postal-address, publication, their relationships and
classifications (see Fig. 2 above). A transformation into RDF is realized from the
CERIF metadata descriptions in most cases; for publications additional vocabularies
e.g. the LODE-BD\textsuperscript{a} recommendations have been employed.

The data publication process was automated and managed through D2R,\textsuperscript{o}
offering a Linked Data interface and including a so-called SPARQL endpoint to
publish open data, and supports in meeting the LOD principles [9]. In the remainder,
we describe the exposition of CERIF Research and Link Entities.

4.2.1. Exposing CERIF research entities

Research entities stored in CERIF tables are published as RDF resources, typed and
enriched with metadata according to the CERIF ontology and other external
vocabularies, as described below. Furthermore, internal entity identifiers are exposed
as a property cerif:internalIdentifier, which is currently a sub-property of the literal
dc:identifier property derived from Dublin Core.

Linked Data can be consumed by a wide variety of different client applications
and libraries. To ensure that our data can be consumed by applications that may be
turned to well-known vocabularies, the CERIF Ontology re-uses terms from those
categories. However, RDFS/OWL reasoners are not as widely spread as client
libraries for accessing RDF. In order to publish data for usage by clients with limited
reasoning capabilities, and following the Materialize Inferences pattern in [27] we
opted for exposing both, the original (CERIF) and some inferred triples (DC and
FOAF) as linked data. Table 4 provides an RDF metadata example for a CERIF
project, showing the employed ontologies through prefixed namespaces, e.g. rdfs, dc,
and cerif in the left-hand column. These are mapped to CERIF attributes within an
relational environment in the right-hand column. The full mapping of the CERIF
model will be described in an upcoming CERIF-LOD specification document.

Because metadata are generated at execution time while querying via the inter-
faces provided by the D2R Server, an explicit SQL to RDF mapping had to be
defined for the server configuration. Where as in Table 4, we indicate the RDF
metadata for a research project, the corresponding RDF representation of the

\textsuperscript{a}LODE Recommendations for bibliographical data: http://aims.fao.org/lode/bd

\textsuperscript{o}D2R Server: http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
The VOA3R project description itself is shown in Listing 1. The CERIF multilinguality features are considered by `xml:lang` attributes in the RDF/XML descriptions or similarly appending the string literal with `@` and the language tag of the RDF/Turtle descriptions.

```turtle
@prefix cerif:   <http://eurocris.org/cerif/0.1#> .
@prefix foaf:    <http://xmlns.com/foaf/0.1/> .
@prefix dc:      <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .

<http://example.org/resource/projects/VOA3R> a cerif:Project ;
    cerif:title "Virtual Open Access Agriculture and Aquaculture Repository"@en ;
    cerif:uri <http://voa3r.cc.uah.es> ;
    cerif:startDate "2010-06-01"^^xsd:date ;
    cerif:endDate "2013-05-01"^^xsd:date ;
    cerif:abstract "The general objective of the VOA3R project is to improve the spread of European agriculture and aquaculture research results by using an innovative approach to sharing open access research products"@en ;
    cerif:acronym "VOA3R" ;
    cerif:internalIdentifier "cfProj-voa3r-ID" ;

## Materialized inferred triples
rdfs:label "Virtual Open Access Agriculture and Aquaculture Repository"@en ;
dc:title "Virtual Open Access Agriculture and Aquaculture Repository"@en ;
foaf:homepage <http://voa3r.cc.uah.es> .
dcterms:abstract "The general objective of the VOA3R project is to improve the spread of European agriculture and aquaculture research results by using an innovative approach to sharing open access research products"@en ;
```

Listing 1. Project example in RDF Turtle syntax.

Table 4. Mapping CERIF/RDF for the project entity.

<table>
<thead>
<tr>
<th>RDF properties</th>
<th>CERIF entity attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>cerif:internalIdentifier</td>
<td>cfProj.cfProjId</td>
</tr>
<tr>
<td>dc:identifier</td>
<td></td>
</tr>
<tr>
<td>cerif:startDate</td>
<td>cfProj.cfStartDate</td>
</tr>
<tr>
<td>cerif:endDate</td>
<td>cfProj.cfEndDate</td>
</tr>
<tr>
<td>cerif:acronym</td>
<td>cfProj.cfAcro</td>
</tr>
<tr>
<td>cerif:uri</td>
<td>cfProj.cfURI</td>
</tr>
<tr>
<td>foaf:homepage</td>
<td></td>
</tr>
<tr>
<td>cerif:title [lang]</td>
<td>cfProjTitle.cfTitle</td>
</tr>
<tr>
<td>dc:title [lang]</td>
<td>[cfProjTitle.cfLangCode]</td>
</tr>
<tr>
<td>rdf:label [lang]</td>
<td></td>
</tr>
<tr>
<td>dcterms:abstract [lang]</td>
<td>[cfProjAbstr.cfLangCode]</td>
</tr>
<tr>
<td>cerif:keyword [lang]</td>
<td>cfProjKeyw.cfKeyw</td>
</tr>
<tr>
<td>dc:subject [lang]</td>
<td>[cfProjKeyw.cfLangCode]</td>
</tr>
</tbody>
</table>

VOA3R project description itself is shown in Listing 1. The CERIF multilinguality features are considered by `xml:lang` attributes in the RDF/XML descriptions or similarly appending the string literal with `@` and the language tag of the RDF/Turtle descriptions.
4.2.2. Exposing CERIF link entities

In CERIF, link entities provide the mechanism to maintain timely relationships between entities, that is, roles and types. The mechanism allows for valid time intervals through relationships (link entity attributes \textit{cfStartDate}, \textit{cfEndDate}), as well as for fractional values (\textit{cfFraction}) to express proportions or probabilities of recorded facts. There are two approaches when it comes to exposing these attributes in RDF:

- Adding annotations to facts as properties of a reified basic statement. This approach is described as the RDF Reification e.g. in [27].
- Generating a new RDF node for each fact, and link it with the involved entities. This approach is described as the Qualified Relation pattern in [28]. The additional information then goes into properties of the newly generated node.

The former approach avoids the need for having a node in the middle between the two entities of the fact. However, this comes at the cost of added complexity in reaching the additional temporal information. Since temporal information on facts is crucial with RISs and CERIF, we chose the latter approach. Hence, we generate a new RDF statement, relating two entities through a given predicate from an external LD vocabulary, such as the already described CERIF Vocabulary. For the re-use of standardized predicates, it is desirable to connect the classification schemes managed by a RIS with vocabularies that are publicly accessible on the Web.

Here, we present several examples for the exposition of CERIF link entities as linked data. The RDF example in Listing 2 describes that a person is the coordinator of a research project during a given time interval. In this case, both entities belong to the same dataset, although the procedure would be similar if the project was located in another dataset.

Besides (binary) link entities to express facts, CERIF supports (unary) classifications of entities, enhanced by time information as indicated in Listing 3, where the

```
<http://example.org/resource/persons/Miguel-Angel_Sicilia>
  cerif:linksToProject <http://example.org/resource/proj_pers/123XYZ>.

<http://example.org/resource/projects/VOA3R>
  cerif:isLinkedByPerson <http://example.org/resource/proj_pers/123XYZ>.

<http://example.org/resource/proj_pers/123XYZ>
  a cerif:Relationship;
  rdfs:label "Association between VOA3R and Miguel-Angel Sicilia";
  cerif:role <http://eurocris.org/semcerif#coordinator>;
  cerif:startDate "2010-06-01 00:00:00.0";
  cerif:endDate "2011-12-31 23:59:59.0".
```

Listing 2. Relationship between a person and project (without namespace prefixes).
above person is classified as a PhD (Doctor of Philosophy) through a vocabulary term (PhD) from the CERIF controlled vocabulary.

One of the basic principles of Linked Data is to set RDF links by URIs pointing into other data sources on the Web. These are the glue connecting isolated RIS into a globally interconnected research information space. In Listing 4, we demonstrate the linkage (“same as” relationship) of a given researcher from a CERIF-based RIS with the RDF description of that same researcher in external systems (non CERIF), such as DBLP or RKB Explorer. The former is the linked data version of DBLP bibliographic database, and the procedure to automatically generate links is simple due to the straightforward naming schema used in that dataset. The latter is a semantic web application that is able to present unified views of heterogeneous triple-stores, such as publications of ACM. In this case, the procedure to discover relationships between data items within different Linked Data sources is more complex. Hence, a proper matching tool such as SILK is required to tackle these issues.

In addition to the above, CERIF research systems increasingly require connections with KOS, such as AGROVOC [29]. In Listing 5, we see how a certain research project is classified by a given term, in this case, \( e_{550} \): Aquaculture.

\[ \]
4.2.3. Exposing additional metadata

A public exposition of data in the LOD Web requires additional metadata; the metadata describing an exposed dataset itself. It is highly important to provide information about the license or waiver under which the data have been published including the restrictions that apply to their use. With applications, it is important to indicate the origin of the data and publish provenance metadata with the primary data. Hence, we enriched the descriptions of all our CERIF RDF resources by common Dublin Core predicates, such as \texttt{dc:creator}, \texttt{dc:publisher}, \texttt{dc:date} and \texttt{dc:rights}, as indicated in Listing 6.

Finally, it is important to describe the published RDF dataset itself allowing crawlers to discover the data. We encourage the use of the VoID vocabulary for this purpose. VoID\(^9\) provides a RDF vocabulary for the expressing of metadata about datasets supporting users with their decisions over appropriate usage for their intended purposes. In our case, we adopt the predicates indicated in Listing 7 to specify the location of the SPARQL endpoint, the used vocabularies and a representative example entity for a research dataset.

4.3. Connecting CERIF RIS through the linked open data web

In this section we describe a key use case to illustrate the procedure to obtain the full CV of a given researcher, who has worked for several organizations during his career. Hence, his research information is spread across several RIS located at different universities. For information integration one could build a mashup web app such as in [30] consuming metadata about research information. We assume there is a registry of CERIF\(^9\) datasets publishing Linked Data based on our method. The new web

\(^9\) Describing Linked Datasets with the VoID: http://www.w3.org/TR/void/
application must follow steps: First, a consumer application will prepare a SPARQL query to ask what the (researcher’s) identifiers (URIs) of a given person are. This would require a query like the one shown in Listing 8 over each available CERIF dataset.

Once the identifiers from different datasets are gathered, the mashup should prepare new SPARQL queries to obtain all information related to the researcher. For example, we suppose that http://example.org/resource/persons/Miguel-Angel_Sicilia is one of the identifiers of the researcher in a specific dataset. Then, our mashup would prepare a query for retrieving the list of organizations (and the dates) with which the researcher maintained relationships (see Listing 9). This query will be issued to the SPARQL endpoint using the above ID as an input parameter in order to obtain the required information. This procedure should be repeated for each of the linked datasets from our registry.

Listing 10 presents an additional query example to retrieve the researcher’s set of publications. Finally, the mashup should merge the retrieved results from the several datasets and present it in a friendly way to the end-user.
5. Conclusion

This paper proposes the extension of a relational ERM standard model (CERIF) to enable the exposition and thus interconnection of RISs through LOD technologies. The extension showcases benefits to the different stakeholders involved in research activities. Several models, reference models or data models for managing information in the wider context of research have been developed; e.g. for scholarly repositories in a series of EU-funded DRIVER and follow-up projects [31] and are available for the digital libraries domain, e.g. through the DELOS reference model [32] and with Europeana as a data model [33] based on RDF, furthermore, e.g for the Language Technology domain [34]. ISO maintains a reference model for Open Archival Information Systems [35] and in the library domain multiple models [e.g. FRBR, IFLA, Marc, MODS] are circulating. Furthermore, in the research context there exist numerous classifications systems of Science. Where it will be interesting to investigate and relate the existing models in more details according to the underlying LOD paradigm, the proposed entity for extension may well be elaborated towards an even more generic concept of linkage, outside the LOD Web and even the possibility of linkages between relationships themselves.

```
SELECT DISTINCT ?organizationName, ?startDate, ?endDate
WHERE {
  <http://example.org/resource/persons/Miguel-Angel_Sicilia>
  cerif:linksToOrganisationUnit ?relationship .
  ?relationship cerif:startDate ?startDate .
  ?relationship cerif:endDate ?endDate .
  ?organizationId cerif:isLinkedByPerson ?relationship .
  ?organizationId foaf:name ?organizationName ;
}
ORDER BY ?organizationName
```

Listing 9. SPARQL query to obtain the organizations units with which the researcher maintains relationships through a particular time part to a certain amount.

```
SELECT DISTINCT ?publicationTitle
WHERE {
  <http://example.org/resource/persons/Miguel-Angel_Sicilia>
  cerif:linksToPublication ?relationship .
  ?publicationId cerif:isLinkedByPerson ?relationship .
  ?publicationId dc:title ?publicationTitle ;
}
ORDER BY ?publicationTitle
```

Listing 10. SPARQL query to obtain the publications of the a given researcher.
The Web of data provides an excellent opportunity for significant simplifications with data integration and federation problems [36]. In order to exploit these advantages, we plan to develop a registry and federation of CERIF datasets to enable the development of new and third-party applications which consume research data based on CERIF via LOD.

Acknowledgments

This work was supported by EC Funding under the VOA3R project through the grant agreement no. 250525, and META-NET through grant agreement no. 249119.

References

4. E. Zimmermann, CRIS-Cross: Current research information systems at a crossroads.
11. Y.-F. Li, G. Kennedy, F. Davies and J. Hunter, PODD: Towards an extensible, domain-agnostic scientific data management system, in Proc. 6th IEEE eScience Conference 2010, University of Queensland, Brisbane, Australia, pp. 137–144.
15. Y. Wand, R. Weber, P. Weingartner and G. J. W. Dorn, Mario Bunge’s ontology as a formal foundation for information systems concepts, Studies on Mario Bunge’s Treatise on Basic Philosophy (Rodopi, 1990), pp. 123–149.


