Journal of Universal Computer Science, vol. 19, no. 12 (2013), 1854-1867
submitted: 20/11/12, accepted: 10/5/13, appeared: 28/6/13 © J.UCS

Promoting International Interoperability of Research Information Systems: VIVO and CERIF

Leonardo Lezcano
(Information Engineering Research Unit, Computer Science Department
University of Alcalá, Alcalá de Henares, Madrid, Spain
leonardo.lezcano@uah.es)

Brigitte Jörg
(Jisc Innovation Support Center, CERIF National Co-ordinator
UKOLN, University of Bath, UK
b.joerg@ukoln.ac.uk)

Brian Lowe
(Albert R. Mann Library, Cornell University, New York, USA
bjl23@cornell.edu)

Jon Corson-Rikert
(Albert R. Mann Library, Cornell University, New York, USA
jc55@cornell.edu)

Abstract: Institutional repositories (IR) and Current Research Information Systems (CRIS) store and manage information on the context in which research activity takes place. Several models, standards and ontologies have been proposed to date as solutions to provide coherent semantic descriptions of research information. These present a large degree of overlap but also present very different approaches to modelling. This paper introduces a contrast of two of the more widespread models, the VIVO ontology and the CERIF standards, and provides guidance for mapping them in a way that enables clients to integrate data coming from heterogeneous sources. The majority of mapping challenges have risen from the representation of VIVO sub-hierarchies in CERIF as well as from the representation of CERIF attributes in VIVO. In addition, the paper illustrates features for linking data across the Web, for querying of geographically distributed data stores and for aggregating data described using different data models in a common store. These features are supported by semantic web technologies including RDF, OWL and SWRL.

Keywords: CERIF, VIVO, CRIS, research information, scientific information, ontologies, knowledge representation, mapping, semantic interoperability, linked data, OWL, SPARQL
Categories: H.2.5, H.1.0, H.4.2

1 Introduction

Traditionally, most research has been curiosity-led, discipline-oriented, and motivated and executed by a small group of individuals following a particular hypothesis, experiment or proven method. Nevertheless, the complex problems that science is facing nowadays require large teams with each member providing a specialized contribution to the whole, and in many cases requiring external collaboration [Toral, 2013]. These collaborative teams are often geographically dispersed and represent different disciplines. Gibbons et al. refer to the fact that science has been shifting from discipline-oriented to cross-disciplinary research as Mode Two [Gibbons, 94]. Also, data science is requiring new services to be available [Wang, 2013].

Many factors including the increased knowledge, the above mentioned paradigm shift, the recognition of economic stimulus and new patterns of collaborative interdisciplinary science lead inexorably to the need for systems to assist researchers, administrators, strategists, opinion-formers, entrepreneurs and also the general public [Zimmerman, 02]. Current Research Information Systems (CRIS) are expected to provide a wide range of information on the conduct of science at the local, national, and international level.

In order to support decision making and knowledge creation, CRIS can be used to find specialized equipment or facilities, recognize innovations and results (to avoid duplication of effort), manage the grant process, produce statistics and reports, evaluate projects and assess science, promote science in society and to locate funding sources, among other applications. In order for research information systems to properly represent the content and context of research work, Sicilia provides examples that could serve as a point of departure to develop an upper ontology for research methods and tools [Sicilia, 10].

CERIF [Jörg, 10] is the common European research information model for the development of new CRISs and a template both for data exchange between CRISs and for mediating access to multiple heterogeneous distributed CRISs. CERIF was released as an EC Recommendation to European Member States in 2000.

Across the Atlantic, the VIVO project [Krafft, 10] has been creating an open, Semantic Web-based network of institutional ontology-driven databases to enable national discovery, networking, and collaboration via information sharing about researchers and their activities. Work began on VIVO at Cornell University in 2003 and has scaled up to the national and international level since 2009 with receipt of a major National Institutes of Health grant[1] under the American Recovery and Reinvestment Act of 2009[2].

The purpose of the present research is to study the overlaps and differences between these two widespread approaches to research information modeling. Section 2 provides a background of both models. Then, Section 3 explains the directions for mapping them in a way that enables clients to integrate data coming from heterogeneous sources. Section 4 introduces the use of Semantic Web technologies to the mapping exercise and allowing to answer particular queries without necessarily having to perform a wholesale conversion of data from one format to the other. Conclusions are finally presented in Section 5.

[1] http://www.nih.gov/news/health/nov2009/ncrr-02.htm

[2] http://www.recovery.gov/

2 Background

This Section introduces the VIVO ontology and the CERIF model.

2.1 The CERIF model

CERIF is considered a standard recommended by the European Union to its Member States[3]. The CERIF model represents information about entities such as *Publication*, *Project*, *Organization*, *Person*, *Product*, *Patent*, *Service*, *Equipment* and *Facility* as well as semantically enhanced relationships among these entities in a formalized way. The physical model is a relational database model available as SQL scripts based on common ERM (Entity Relationship Model) constructs. The latest releases include a formalized, so called “Semantic Layer,” and an XML interchange format [Jörg, 09].

The CERIF model is conceptually structured into entity types and features. Among the types, *core*, *result*, *link*, and “2nd level entities” are distinguished. Multiligualism and semantics are considered as features. Further details can be found in [Jörg, 12]. A mapping between the CERIF part related to published results of scientific research and the MARC 21 bibliographic standard is studied in [Ivanovic, 11], and a CERIF data model extension for the evaluation of scientific research results is proposed by Ivanovic et al. in [Ivanovic, 10].

2.2 The VIVO ontology

All data in VIVO are represented as Resource Description Framework (RDF)[4] statements using classes and properties in the Web Ontology Language (OWL)[5]. These ontologies specify the types of resources described in VIVO and their relationships. The VIVO core ontology[6] models the people, organizations, and activities involved in scientific research. In accordance with the principles of the Linked Data[7] initiative [Berners-Lee, 09], the VIVO core ontology extends existing ontologies such as the Friend-of-a-Friend (FOAF) ontology[8], which provides the basis for describing persons and organizations, and the Bibliographic Ontology (BIBO)[9]. A comprehensive list of the ontologies integrated into the VIVO ontology can be found in the VIVO Project Wiki[10]. A description of VIVO ontology design principles including remaining independent of specific domains and representing temporal relationships is also available in the VIVO Project Wiki.

[3] http://cordis.europa.eu/cerif/

[4] http://www.w3.org/RDF/

[5] http://www.w3.org/TR/owl2-overview/

[6] http://vivoweb.org/download

[7] http://linkeddata.org

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

[9] http://bibliontology.com/

[10] https://wiki.duraspace.org/display/VIVO/VIVO+Ontology+main+page

3 Mapping CERIF and VIVO

This section is intended to provide mapping recommendations for the elements of the CERIF model described in the FDM specification document [Jörg, 12] to the VIVO 1.4 ontology. While both the entire CERIF model and VIVO allow for many more types of relationships and entities than discussed here, it is expected that the approach required to create any mapping between knowledge artifacts in the two models can be derived from the following recommendations. General metrics from CERIF and VIVO are provided in Table 1 and 2, and discussed below in the Conclusions section.

CERIF			
Entities	Attributes	Link Entities	Language Entities
56	1766	120	61

Table 1: CERIF model metrics

VIVO		
Classes	Datatype Properties	Object Properties
209	94	218

Table 2: VIVO ontology metrics

3.1 CERIF Base, Result and Infrastructure Entities ([base], [result] & [infra])

Mapping CERIF *Base*, *Result* and *Infrastructure* entities to VIVO is a straightforward process given the fact they have no foreign key (FK)[11] and therefore most of their attributes can be mapped as datatype properties between a given class in VIVO and a data literal. It should be noted that a minority of the attributes, such as *cfURI* in the *cfProj* table, are mapped to an object property in VIVO. The *vivo:webpage* property has an instance of *vivo:URLLink* as its object; a *URLLink* simply pairs the URL of the webpage with a string to serve as the HTML anchor text.

To represent complex objects, CERIF uses "2nd level entities" connected to base entities through link entities. These have been mapped to VIVO as shown in Section 3.3. From a conceptual perspective, 2nd level entities in CERIF represent the environment in which the base entities act, communicate and produce results. The *cfEvent* table is currently classified by CERIF as a 2nd level entity while VIVO considers the *Event* class as another first-order entity.

Result entities like *cfResPubl[ication]*, *cfResPat[ent]* and *cfResProd[uct]* can be mapped to specializations of the VIVO *InformationResource* class. Table 3 includes mapping examples[12]. As in the first two cases, once the mapped classes and tables have been identified, CERIF attributes must be mapped to VIVO properties. It should be noted that most of the Multiple Language CERIF features are not explicitly modeled in VIVO but can be accommodated natively in RDF via language tags on untyped literals. The VIVO application has been in version 1.5 to recognize language tags and render a preferred literal based on the user's current browser language settings.

[11] The Currency Code attribute (*cfCurrCode*) is an exception.

[12] The CERIF group indicates that, in the next major CERIF release, measurement attributes such as *headcount* and *turnover* will no longer be supported explicitly. The recommended alternative will be the new and generic measurement entities for calculations, and other inferred data.

CERIF	VIVO
Table	**Class**
cfPers	foaf:Person
cfResPubl	bibo:Document
dfResPat	bibo:Patent
cfResProd	vivo:CaseStudy vivo:Dataset
cfFacil	vivo:Facility
cfSrv	vivo:Service

Table	Attribute	Class	Property
cfProj	cfURI	vivo:Project	vivo:webpage only vivo:URLLink
	cfAcro		vivo:description only Literal
	cfStartDate		vivo:dateTimeInterval only vivo:DateTimeInterval
	cfEndDate		
cfOrgUnit	cfAcro	foaf: Organization	vivo:abbreviation only Literal
	cfURI		vivo:webpage only vivo:URLLink
	cfHeadcount		not modeled, but can be inferred by counting the number of Person instances which are related to a given Organization through an appropriate object property or Position node
	cfTurn		not modeled (not even in vivo:PrivateCompany)

Table 3: Examples of mappings between CERIF Base and Result Entities and VIVO classes and properties.

3.2 CERIF Semantics [class]

In the CERIF model, the semantics of a given record within a broad entity like *Project* (*cfProj_Class*) are enriched by a time-stamped reference to the CERIF Semantic Layer to host any vocabulary, e.g. the CERIF 1.3 Vocabulary[13]. The VIVO ontology uses a sub-hierarchy to accomplish such specialization of concepts, e.g., *Human Study* is a subclass of *Research Project* that in turn is a subclass of the top classification *Project*. More examples are included in Table 4. In addition, VIVO uses references to external vocabulary terms less to define what something is than to document subject or topic associations; external references expressed as Linked Data

[13] http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Semantics/CERIF1.3_Vocabulary.xls

retain their original URIs to facilitate direct data interoperability across institutions and platforms.

CERIF		VIVO	
Table	**Class Term**	**SubClass**	**Top Class**
cfProj_Class	Discipline Codes, Application Codes	eagle-i:ResearchProject, eagle-i:HumanStudy, eagle-i:Clinical Trial	vivo:Project
cfPers_Class	Consultant, Lecturer, Research Fellow, etc	vivo:FacultyMember, vivo:Librarian, vivo:Non-AcademicStaff, etc.	foaf:Person
cfOrgUnit_Class	Private non-profit, University College, etc.	vivo:Association, vivo:College, vivo:Consortium, etc.	foaf:Organization
cfResPubl_Class	Book, Review, etc.	bibo:Collection, bibo:Article and other BIBO classes merged with VIVO classes such as vivo:CaseStudy, vivo:Catalog	vivo:Information Resource

Table 4: Examples of mappings between CERIF Semantics and VIVO classes and subclasses.

3.3 CERIF Link Entities [link]

CERIF defines every relationship between two entities using a pair of record identifiers (*cfId1* and *cfId2*) taken from the tables representing those entities. The semantics of the pair are then enriched by a time-stamped reference to the CERIF Semantic Layer to host vocabularies of any structure. The VIVO ontology fulfills this relation classification task by means of a hierarchy of *Object Properties* combined with the taxonomy of VIVO classes. The *cfFraction* attribute (Float) has no direct mapping to VIVO currently; instead, VIVO stores author order via the *authorRank* property on an Authorship relating a person and a publication, while offering only a free-text *description* property on the *ResearcherRole* relating a person to a project. For clarity in semantic interpretation, we propose adding an explicit *vivo:fraction* datatype property to the VIVO ontology wherever there are VIVO classes (e.g., *Authorship*, *Role*, *Position*) to describe such relations (see Table 5). Instances of these intermediate classes, or "context nodes," also support relationships to time intervals for a relationship between two primary entities.

4 Semantic Integration of VIVO and CERIF data

Establishing mappings between the CERIF and VIVO data models will guide software developers in implementing systems that can import or export data in both formats, or translate from one to the other. Semantic Web technologies can also be applied to the mapping exercise for the goal of data integration: combining data from disparate CERIF and VIVO systems to answer particular queries without necessarily having to perform a wholesale conversion of data from one format to the other.

CERIF				VIVO		
Table	**cfd1**	**Semantic Stamp Example**	**cf1d2**	**Domain Class**	**Data Property**	**Range Datatype**
cfOrgUnit_Eaddr	cfOrgUnitId	Email	cfEAddrId	foaf:Organization	*vivo:email* (datatype property)	abc123@mydomain.com
cfOrgUnit_Eaddr	cfOrgUnitId	Skype	cfEAddreId	foaf:Organization	*vivo:skype* (yet-to-be-modeled datatype property)	abc123

Table	cf1d1	Semantic Stamp Example	cf1d2	Domain Class	Object Property	Range Class
cfProj_ResPubl	cfProjId	Originator	cfResPubId	vivo:Project	*vivo:informationProduct*	vivo:Information Resource

Table	cf1d1	Semantic Stamp Example	cf1d2	Domain Class	Object Property #1	Intermediate Class ("context node")	Object Property #2	Range Class
cfProj_Fund	cfProjId	Funder	cfFundId	vivo:Project	*vivo:has Funding Vehicle*	vivo:Grant	*vivo:grant AwardedBy*	foaf:Organization
cfProj_OrgUnit	cfProjId	Coordinator	cfOrgUnitId	vivo:Project	*vivo:realized Role*	vivo:OrganizerRole	*vivo:roleOf*	foaf:Agent
cfProj_OrgUnit	cfProjId	Fract[0.2]	cfOrgUnitId			OrganizerRole. *fraction (not yet modeled)*		
cfPers_ResPubl	cfPersId	Author	cfResPublId	foaf:Person	*vivo:autor InAuthorship*	vivo:Authorship	*vivo:linked Information Resource*	vivo:Information Resource
cfPers_ResPubl	cfPersonId	Author (percentage)	cfResPubId			Authorship.*fraction*		
cfPers_OrgUnit	cfPersId	Affiliation	cfOrgUnitId	foaf:Person	*vivo:person InPosition*	vivo:PrimaryPosition	*vivo:position InOrganization*	foaf:Organization #1
cfPers_OrgUnit	cfPersId	SubAffiliation	cfOrgUnitId		*vivo:person InPosition*	vivo:Position	*vivo:position InOrganization*	foaf:Organization #2
cfPers_OrgUnit	cfPersId	Board-Member or TG-Leader	cfOrgUnitId	foaf:Person	*vivo:has LeaderRole*	vivo:LeaderRole	*vivo:contributes To*	foaf:Organization
cfProj_Pers	cfPersId	Coordinator	cfProjId	foaf:Person	*vivo:has OrganizerRole*	vivo:OrganizerRole	*vivo:roleRealizedIn*	vivo:Project
cfProj_Pers	cfPersId	Coordinator [fract=0.7]	cfProjId			OrganizerRole. *fraction*		
cfProj_Pers	cfPersId	Participant	cfProjId	foafPerson	*vivo:has Researcher Role*	vivo:ReseacherRole	*vivo:roleRealizedIn*	vivo:Project
cfProj_Pers	cfPersId	Participant [fract=0.3]	cfProjId			ResearcherRole. *fraction*		
cfOrgUnit_Paddr	cfOrgUnitId	post-office-box	cfPAddrId	foaf:Organization	*vivo:mailing Address*	vivo:Address	*vivo:address1 (datatype)*	"P.O. Box 2012"
							vivo:addressCity (datatype)	"Midland"

Table	cf1d1	Semantic Stamp Example	cf1d2	Class	Datatype properties	Assertion	Class	Datatype properties
cfPersName_Pers	cfPersId	Spelling Variant	cfPersId2	foaf:Person #1	*foaf:firstName*	*owl:sameAs*	*foaf:Person #2*	*foaf:firstName*
					vivo:middle Name			*vivo:middleName*
					foaf:lastName			*foaf:lastName*

Table 5: Examples of mappings between CERIF Link Entities and VIVO classes and properties[14]*.*

The simple syntactic model of RDF, in which structured data is broken down into a set of simple triples or statements, makes the language well suited to tasks involving the combination of data from different sources. RDF and related semantic technologies provide features for linking data across the Web, for query of geographically distributed data stores, and for aggregation of data described using different data models in a common store.

[14] CERIF definition: "A coordinator is someone whose task is to see that work goes harmoniously."

4.1 Linked Data

The VIVO application stores its data natively as RDF statements in a semantic triple store, or collection of statements that refer to classes and properties in the VIVO ontology using URI identifiers. The statements also include URIs for *individuals*, or instances of classes. A person or organization or a publication, for example, will be represented by a URI. In VIVO, these URIs are designed not only to serve as unique identifiers to unambiguously distinguish the subject or object of a triple from resources with similar labels, but also to be dereferenceable. That is, requesting individual URIs should direct a user to meaningful data. Human users will be shown data on an HTML page rendered in the Web browser, while software agents may be directed to RDF triples about the URI if they indicate via HTTP content negotiation that they accept one or more of the RDF syntaxes.

Because the RDF triples encode links to URIs for other individuals which are in turn dereferenceable, a software agent may “crawl” this web of data the same way that a search engine indexer spiders HTML pages, following links to discover related information. By publishing RDF for mode of data traversal and discovery, VIVO applications aim to conform to the established “rules” of linked data originally described by [Berners-Lee, 06].

Although the use of common, well understood RDF predicates and ontology classes greatly promotes data reusability, the relative simplicity of the linked data technique means that data may be discovered and harvested without complete knowledge of its structure and without needing to formulate a query in a particular query language.

While the VIVO ontology was designed from the outset to support the publication of linked data, the exposure of linked data for systems based on CERIF is an emerging area of work. Though there is not yet a standardized, community-endorsed method of transforming CERIF data into RDF, experimental representations of CERIF’s base entities and semantic layer terms in the OWL and RDFS languages permit CERIF-based CRIS systems to publish their contents as linked data.

Many linked data sources so not store their data natively as RDF statements in triple stores. Instead a wrapper service such as D2RQ [Bizer, 04] translates semantic queries into SQL queries that are then issued against an underlying relational database. This technique may be applied to existing CERIF databases to offer linked data without otherwise modifying existing applications to support semantic technologies directly.

With both VIVO and CERIF-based CRIS systems publishing linked data, the entities described in each may be connected by URI reference even without complete transformation from one system’s data model to the other. For example, if a researcher described in a VIVO system with the URI:

http://vivo.example.edu/individual/n24

also appears in a CERIF-based CRIS system as, for example, a coauthor or collaborator or in a previous academic position. The CRIS-backed linked data will generate a different URI representing the same individual:

http://cerif.example.edu/dataset/resource/persons/John-Smith

The OWL vocabulary provides a "sameAs" predicate by which these two URIs can be asserted to refer to the same individual, as in the following RDF triple:

```
http://vivo.example.edu/individual/n24
   owl:sameAs
http://cerif.example.edu/dataset/resource/persons/John-Smith .
```

If such a triple is added to the VIVO triple store, software clients requesting linked data for the first URI can automatically discover the existence of additional data in the second CERIF-based system and request additional CERIF RDF. While the second set of linked data may use different classes and properties, the client may apply any number of mapping rules between VIVO and CERIF to translate the data to the extent deemed necessary for its particular application. Though the problem of discovering and maintaining *sameAs* links is not trivial and may require sophisticated algorithms for entity disambiguation, this lightweight approach to data integration through linked data enable permits discovery and harvest from heterogeneous systems without exhaustive semantic conversion between formats.

4.2 SPARQL Query Rewriting

A similar but more powerful technique for integrating VIVO and CERIF data through client-side translation involves queries written in the W3C-recommended language for querying RDF data, SPARQL[15,16]. In addition to publishing RDF as linked data, several VIVO sites have chosen to deploy SPARQL endpoints to make their semantic data stores directly queryable using this language. The installation of D2R servers or similar SPARQL-to-SQL rewrite engines will enable the addition of SPARQL endpoints to existing relational database-backed CRIS systems.

Queries written for one type of data may be translated for the other by using a mapping specification. While it is anticipated that software tools will be developed to perform this step automatically, certain types of edge cases are likely to require manual decisions about how to reformulate queries.

For example, the query below to retrieve projects and their related web sites from CERIF RDF

```
SELECT ?projectName ?projectPage WHERE {
   ?project a cerif:Project .
   project cerif:uri ?projectPage
}
```

may be written to retrieve similar results from a VIVO SPARQL endpoint:

[15] http://www.w3.org/TR/rdf-sparql-protocol/

[16] http://www.w3.org/TR/rdf-sparql-query/

```
SELECT ?projectName ?projectPage WHERE {
    ?project a vivo:Project .
    ?project vivo:webpage ?webpage .
    ?webpage vivo:linkURI ?projectPage
}
```

Note that the RDF structure in the VIVO example is slightly more complex: instead of linking a project directly to its webpage URI with a single triple, the VIVO RDF introduces a new resource to aggregate data about a webpage link. In VIVO, webpage URIs are typically linked with human-readable labels or "anchor text" to display in browsers. Thus, a query originally designed for VIVO links might contain an additional variable:

```
SELECT ?projectName ?projectPage ?projectPageLabel WHERE {
    ?project a vivo:Project .
    ?project vivo:webpage ?webpage .
    ?webpage vivo:linkURI ?projectPage
    ?webpage vivo:linkAnchorText ?projectPageLabel
}
```

Because there is currently no direct equivalent of anchor text in CERIF, there are several options for rewriting the above query, such as:

(1) Drop the pattern involving linkAnchorText, remove ?projectPageLabel from the SPARQL result set and allow application code to handle the lack of data.

(2) Keep projectPageLabel in the result set but assign it the value of the URI string.

(3) Keep projectPageLabel in the result set assign it the value of the project's label or some concatenated string such as "Project X webpage."

The best strategy will depend on the client consuming the data and its sophistication when dealing with missing data or the extent to which it can be modified to do so. In cases where query rewriting is performed without the client's direct knowledge—through a SPARQL endpoint proxy, for example, that introduces the modifications—solutions like (2) or (3) will likely be required.

4.3 Data Integration Using Ontology Axioms

Semantic Web technologies also make it possible to integrate data through the use of *reasoning*. A reasoner can act on axioms declaring, for example, that the vivo:Project class is equivalent to cerif:Project. With such an axiom, anything described as a project in a CERIF-based CRIS can be automatically inferred to be a VIVO project. While some Semantic Web reasoners, such as Pellet[17] return these entailed inferences only in response to a particular query, others such as OWLIM[18] use a *materialization*

[17] http://clarkparsia.com/pellet

[18] http://www.ontotext.com/owlim

approach where inferred triples are precomputed and added into a triple store as the original asserted data are entered or updated. However they are supplied, inferences have the potential to aid data integration in a number of ways.

4.4 Maintenance of terminologies

The OWL based on Description Logic[19] has a rich set of constructs for defining membership in a class. For example, a class FundingOrganization may be defined in OWL's logical language as the particular subset of all organizations whose members make some (that is, at least one) grant award. In the current version of the VIVO ontology, an *awardsGrant* property is referenced in the definition of FundingOrganization. With such a definition, a reasoner can automatically infer an individual organization's membership in the class. Given the following triples,

example:individual187 rdf:type foaf:Organization.
example:individual187 vivo:awardsGrant example:individual2043.

a reasoner can supply the additional triple:

example:individual187 rdf:type vivo:FundingOrganization.

Reasoners can also infer relationships between classes. For example, a new class TrainingFundingOrganization might be defined as the subset of organizations that have an *awardsGrant* property relating to some training grant. With the appropriate definition, a reasoner can also infer

example:TrainingFundingOrganization
rdfs:subClassOf
vivo:FundingOrganization.

Though this example may seem obvious, the problem of maintaining class hierarchies and equivalences becomes non-trivial for larger vocabularies, where reasoners are able to infer relationships between classes that might be overlooked by a human editor. Adding or enhancing logical OWL definitions for terms in CERIF's semantic vocabularies and for VIVO's ontologies has the potential to reduce the manual effort needed to create and maintain ontology mappings, provided that a base set of mapped properties can be agreed upon for use in constructing the logical definitions.

4.5 Temporal concerns

Both the CERIF and VIVO data models introduce some structural complexity in order to permit the aggregation of data that change over time. For example, a researcher may hold different employment positions over the course of his or her career. CERIF

[19] http://www.w3.org/TR/owl-guide/

and VIVO are both designed to store such positions as distinct entities with values for start and end dates.

For practical query purposes, it is often useful to identify persons who currently hold a certain type of position by making them members of a particular ontology class, such as vivo's *Postdoc*. Writing a completely accurate logical definition for Postdoc, however, is difficult given the limitations of the OWL language. Because the data include positions with end dates in the past, a definition of Postdoc as any person holding a postdoctoral research position means that the class will include anyone even known to have held a postdoctoral position in the past, which is typically not useful when querying data. To conform to typical expectations, the reasoner would need to be exposed only to triples about current data, and this segmentation of data would need to be maintained by a separate piece of software.

4.6 Rules to simplify the RDF structure

In other cases, complexity results from a need to store data that is important only in certain applications. For example, VIVO relates authors and their written works not through a simple direct RDF property but through an intervening "context node" of type Authorship that in turn links not only to the authored work but to an optional integer value denoting the rank order of the author in the publication's byline. Author order is vital to retain in order to reproduce appropriate listings and citation strings. But other Semantic Web applications may choose to ignore such data and may be designed discover simple common linked data predicates such as Dublin Core's *creator* property.

Rule languages such as the Semantic Web Rule Language (SWRL)[20] offer the ability to infer these direct properties that bridge or "shortcut" across more complex data. Producing inferences in the other direction – introducing new individuals and a more complex set of relationships from a simple one – can be more challenging. Certain reasoners do offer rule languages that would support mappings of the type above where VIVO introduces an additional node between a project and its webpage URI.

4.7 Discussion

There are some challenges to be addressed in implementing such a strategy, especially in the deployment of reasoners that can operate efficiently over large bodies of data. In addition, care must be taken to prevent seemingly trivial mapping axioms from introducing unexpected side effects that increase the complexity of other related reasoning tasks. The great advantage, however, of a reasoning-based mapping approach is the ability to issue a single query – using either CERIF or VIVO terms – and retrieve data originally described using a different standard. A client querying a SPARQL endpoint including the requisite inferences is thus relieved of the burden of having to perform its own query translation or conversion of retrieved data.

[20] http://www.w3.org/Submission/SWRL/

5 Conclusions

Information models and knowledge artifacts have been designed and improved in the last decade to represent the research domain. In particular, CERIF and VIVO have been widely adopted for such purpose. This paper makes a comparison study between the CERIF relational database model and the VIVO ontology to support interoperability and integration of the systems based in these models. A challenging task during the study has been that of properly mapping the information semantics represented in the CERIF Semantic Layer to the VIVO semantics supported by OWL.

Without considering the CERIF entities that are exclusively oriented to support the language features, an analysis of both models' interoperability reveals that the 209 VIVO classes provide a finer classification granularity than the 56 CERIF entities. Similar conclusions are reached when comparing the 218 Object Properties and sub-properties in the VIVO ontology with the 120 link entities in CERIF (see Table 1 and 2). While offering more classes and relationships improves the semantics and accuracy of the research knowledge representation, it should be noted that maintainability and integration feasibility may be jeopardized. In order to increase the semantics associated to entities while preserving the simplicity of the model, CERIF use a controlled vocabulary to describe entities and relationships (see Section 3.2). Nevertheless, it is a flat classification method that does not support attributes inheritance.

When comparing the amounts of CERIF attributes and VIVO properties, the 1700 CERIF attributes supports the explicit representation of very specific information such as the *Skype* user assigned to a *Person*. Therefore, and in contrast with classification support, the granularity of the entities detailed by CERIF is usually finer than the one provided by the properties describing VIVO instances. Again, this may have an impact on interoperability, so equilibrium must be preserved.

In spite of the differences, the mapping recommendations in Section 3 show that the most significant research information can be successfully converted from one representation to the other and vice versa. In fact, the three main entities in the CERIF model (i.e., *Person*, *Project* and *OrganizationUnit*) and their attributes can be straightforwardly mapped to three classes that also play an essential role in the VIVO ontology (i.e., *Person*, *Project* and *Organization*). At the same time, the study has found some particular cases where modeling at one side does not support a given piece of information from the other side. Maybe the best example is that the current VIVO version does not yet provide special support for *multilingualism* capacities while CERIF does. Application support for using RDF language tags with VIVO data has been introduced with the VIVO 1.5 release (summer 2012), and multilingual features are set to continue to expand in subsequent releases.

Acknowledgements

The work leading to these results has received funding from the European Commission under grant agreement nº 283770 corresponding to FP7 project agINFRA (A data infrastructure to support agricultural scientific communities Promoting data sharing and development of trust in agricultural sciences)[21], and from

[21] http://aginfra.eu/

the U.S. National Institutes of Health, U24 RR029822, "VIVO: Enabling National Networking of Scientists," http://vivoweb.org, and CTSA 10-001:100928SB23, project 00921-0001, www.ctsaconnect.org.

References

[Berners-Lee, 09] Berners-Lee, T.: Linked Data-The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22, 2009.

[Berners-Lee, 06] Berners-Lee, Tim: “Linked Data – Design Issues.” http://www.w3.org/DesignIssues/LinkedData.html (retrieved 13 Dec 2012).

[Bizer, 04] Bizer, C., Seaborne, A.: D2RQ – treating non-RDF databases as virtual RDF graphs. In Proceedings of the 3rd International Semantic Web Conference (ISWC 2004).

[Gibbons, 94] Gibbons, M.: The new production of knowledge: the dynamics of science and research in contemporary societies. SAGE, 1994.

[Ivanovic, 10] Ivanovic, D., Surla, D., Racković, M.: A CERIF data model extension for evaluation and quantitative expression of scientific research results. Scientometrics. 86, 155-172, 2010.

[Ivanovic, 11] Ivanovic, D., Surla, D., Konjovic, Z.: CERIF compatible data model based on MARC 21 format. The Electronic Library. 29, 52-70, 2011.

[Jörg, 09] Jörg, B., Ojars, K., Jeffery, K., van Grootel, G.: CERIF XML 2008 – 1.0 Data Exchange Format, 2009.

[Jörg, 10] Jörg, B.: CERIF: The common European research information format model. Data Science Journal. 9, 24-31, 2010.

[Jörg, 12] Jörg, B., Jeffery, K., Dvořák, J., Houssos, N., Asserson, A., van Grootel, G.: CERIF 1.3 Full Data Model (FDM) - Introduction and Specification, 2012.

[Krafft, 10] Krafft, D.B., Cappadona, N.A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B.J.: Vivo: Enabling national networking of scientists. Proceedings of the WebSci10: Extending the Frontiers of Society On-Line. , Raleigh, USA, 2010.

[Sicilia, 10] Sicilia, M.-Á.: On Modeling Research Work for Describing and Filtering Scientific Information. In: Sánchez-Alonso, S. and Athanasiadis, I.N. (eds.) Metadata and Semantic Research. pp. 247-254. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[Toral, 2013] Toral, S., Bessis, N., Martínez-Torres, M.R. : External collaboration patterns of research institutions using shared publications in the Web of Science. Program: electronic library and information systems, 47, 2, 170 – 187, 2013.

[Wang, 2013] Wang, M.: Supporting the research process through expanded library data services. Program: electronic library and information systems, 47, 3, 282 – 303, 2013

[Zimmerman, 02] Zimmerman, E.: CRIS-Cross: current research information systems at a crossroads. In Proceedings: Gaining Insight from Research Information, at the 6th International Conference on Current Research Information Systems. Kassel, Germany, 2002.