Modeling the Context of Scientific Information: Mapping VIVO and CERIF

Leonardo Lezcano[1], Brigitte Jörg[2], and Miguel-Angel Sicilia[1]

[1] Information Engineering Research Unit, Computer Science Department,
University of Alcalá, Polytechnic Building, Ctra. De Barcelona km. 33.6
28871 Alcalá de Henares (Madrid), Spain
{leonardo.lezcano,msicilia}@uah.es
[2] DFKI GmbH, Language Technology Lab
Projektbüro Berlin, Alt-Moabit 91c
10559 Berlin, Germany
brigitte.joerg@dfki.de

Abstract. Institutional repositories (IR) and Current Research Information Systems (CRIS) among other kinds of systems store and manage information on the context in which research activity takes place. Several models, standards and ontologies have been proposed to date as a solution to give coherent semantics to research information. These present a large degree of overlap but also present very different approaches to modeling. This paper presents a contrast of two of the more widespread models, the VIVO ontology and the CERIF standards, and provides directions for mapping them in a way that enables clients to integrate data coming from heterogeneous sources. The majority of mapping problems have risen from the representation of VIVO sub-hierarchies in CERIF as well as from the representation of CERIF attributes in VIVO.

Keywords: CERIF, VIVO, CRIS, research information, scientific information, ontologies, knowledge representation, mapping.

1 Introduction

Traditionally, most of research has been curiosity-led, discipline-oriented, and motivated and executed by a small group of individuals following hypothesis, experiment or proven method. Nevertheless, the complex problems that science is facing nowadays require large teams with each member having a specialized contribution to the whole. These collaborative teams are often geographically dispersed and belong to different disciplines. Gibbons et al. refer to the fact that science has been shifting from discipline-oriented to cross-disciplinary research as Mode Two [1].

Increased knowledge, the paradigm shift, recognition of economic stimulus and collaborative interdisciplinary science lead inexorably to the need for systems to assist researchers, administrators, strategists, opinion-formers, entrepreneurs and also the general public [2]. Current Research Information Systems (CRIS) are expected to provide such scientific information. In order to support decision-making and

M. Bajec and J. Eder (Eds.): CAiSE 2012 Workshops, LNBIP 112, pp. 123–129, 2012.

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 [3] provides examples that could serve as a point of departure to develop an upper ontology for research methods and tools.

CERIF [4] 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 has been released as an EC Recommendation to European Member States in 2000.

On the other hand, the VIVO project [5] is 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.

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. Conclusions are finally presented in Section 4.

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[1]. 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 between 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 [6]. The CERIF model is conceptually structured into entity types and features. Among the types, *core*, *result*, *link*, and *2nd level entities* are distinguished. *Multilingualism* and *semantics* are considered as features. Further details can be found in [7]. A mapping between the CERIF part related to published results of scientific research and the MARC 21 bibliographic standard is studied in [8], and a CERIF data model extension for the evaluation of scientific research results is proposed by Ivanovic et al. in [9].

2.2 The VIVO Ontology

All data in VIVO are represented as RDF statements using classes and properties from OWL ontologies. These ontologies specify the types of resources described in VIVO and

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

their relationships. The VIVO core ontology[2] models the people, organizations, and activities involved in scientific research. According to the linked data initiative, the VIVO core ontology extends existing ontologies such as the Friend-of-a-Friend (FOAF) ontology[3], which provides the basis for describing persons and organizations, and the Bibliographic Ontology (BIBO)[4]. A comprehensive list of the ontologies integrated into the VIVO Ontology can be found in VIVO Project Wiki[5]. A description of VIVO ontology design principles like remaining independent of specific domains and representing temporal relationships is also available in the VIVO Project Wiki.

3 Mapping CERIF and VIVO

This document is intended to provide mapping recommendations for the elements of the CERIF model described in the FDM specification document [7] to the VIVO 1.4 ontology. Both the entire CERIF model and VIVO allow for many more types of relationships and entities, however, it is expected that the approach required to create any mapping between such knowledge artifacts 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.

Table 1. CERIF model metrics

CERIF			
Entities	**Attributes**	**Link Entities**	**Language Entities**
56	1766	120	61

Table 2. VIVO ontology metrics

VIVO		
Classes	**Datatype Properties**	**Object Properties**
209	94	218

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)[6] and therefore most of their attributes can be mapped as datatype properties between a given class in VIVO and a data literal and the rest. It should be noted that a minority of the attributes, like cfURI in the cfProj table, are mapped to an object property in VIVO like webpage, but the pointed object, in this case an instance of URLLink, plays a role of user-defined datatype rather than an instance of a complex object. At the same time, CERIF uses "2nd level entities" to represent complex objects, which are connected to base entities through link entities. These have been mapped to VIVO as shown in Section 1.3. From a conceptual perspective, "2nd level entities" in CERIF can be considered as the environment in which the base entities act and communicate, and produce results. The cfEvent table is currently classified by CERIF as a 2nd level entity while VIVO considers the Event class as an independent piece of information.

[2] http://vivoweb.org/download
[3] http://www.foaf-project.org/
[4] http://bibliontology.com/
[5] http://sourceforge.net/apps/mediawiki/vivo/index.php?title=Ontologies
[6] The Currency Code attribute (*cfCurrCode*) is an exception.

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

CERIF		VIVO	
Table	**Attribute**	**Property**	**Class**
cfProj	cfURI	webpage only URLLink	Project
	cfAcro	description only Literal	
	cfStartDate	dateTimeInterval only DateTimeInterval	
	cfEndDate		
cfOrgUnit	cfAcro	abbreviation only Literal	Organization
	cfURI	webpage only URLLink	
	cfHeadcount[7]	not modeled (Assuming that all members are in the system, it can be inferred by counting the number of Person instances which are related to a given Organization through the hasCurrentMember property)	
	cfTurn	not modeled (not even in PrivateCompany)	
cfPers	←----------→		Person
cfResPubl	←----------→		Document
cfResPat	←----------→		Patent
cfResProd	←----------→		CaseStudy, Dataset, etc.
cfFacil	←----------→		Facility
cfEquip	←----------→		Equipment
cfSrv	←----------→		Service

Result entities like cfResPubl, cfResPat and cfResProd can be mapped to specializations of the InformationResource VIVO class. Table 3 includes mapping examples. As in the first two cases, once the mapped classes and tables have been identified, the mapping granularity must be further increased by mapping CERIF attributes to VIVO properties. It should be noted that most of the Multiple Language CERIF features are not modeled in VIVO, for example, although VIVO allows for several abstracts, keywords and titles to be assigned to the same article, the language of such texts has not been defined.

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[8]. The VIVO ontology uses a sub-hierarchy to accomplish such specialization of concepts, e.g., *Human Study* is a subclass of *Research Project* which in turn is a subclass of the top classification *Project*. More examples are included in Table 4.

[7] A contact with the CERIF task group indicated that, in the next major CERIF release, measurement attributes such as headcount and turnover will not be supported anymore explicitly. The recommendation will rather be towards using the new and generic measurement entities for calculations, and other inferred data.

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

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

CERIF		VIVO	
Table	**Class Term**	**SubClass**	**Top Class**
cfProj_Class	Discipline Codes, Application Codes	Research Project, Human Study or Clinical Trial from **eagle-i**.	Project
cfPers_Class	Consultant, Lecturer, Research Fellow, etc	Faculty Member, Librarian, Non-Academic Staff, etc.	Person (**FOAF**)
cfOrgUnit_Class	Private non-profit, University College, etc.	Association, College, Consortium, etc.	Organization (**FOAF**)
cfResPubl_Class	Book, Review, etc.	BIBO and VIVO classes	Information Resource

3.3 CERIF Link Entities [link]

CERIF defines every relationship between two entities using a pair of records 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 mapping of the cfFraction attribute (Float) involves particular mechanisms depending on the semantics of each relation. For example, the *authorRank* property allows representing the cfFraction value related to the author-publication link, while the *description* property can be used to map the values of the cfFraction attribute in the person-project link. In both cases there are VIVO classes (i.e., Authorship and ResearcherRole) which are oriented to describe such relations and serve as domain classes for the mentioned properties (see Table 5).

Table 5. Examples of mappings between CERIF Link Entities and VIVO classes and properties

CERIF				VIVO		
Table	**cfId1**	**Semantic Stamp**	**cfId2**	**Domain**	**Object Property**	**Range**
cfProj_ResPubl	cfProjId	Originator	cfResPublId	Project	Information Product	Information Resource
cfProj_Fund	cfProjId	Funder	cfFundId	Project	hasFunding Vehicle	Agreement
cfProj_OrgUnit	cfProjId	Coordinator	cfOrgUnitId	Project	realizedRole	Role
cfProj_OrgUnit	cfProjId	Fract[0.2]	cfOrgUnitId	Project	not modeled	
cfPersName_Pers	cfPersId	Spelling Variant	cfPersId2	Person	firstName (datatype)	not defined
					middleName (datatype)	
					lastName (datatype)	
cfPers_ResPubl	cfPersId	Author	cfResPublId	Person	authorIn Authorship	Authorship
				Authorship	Information Resource InAuthorship	Information Resource

Table 5. *(Continued)*

cfPers_ResPubl	cfPersId	Author (percentage)	cfResPublId	Authorship	authorRank (datatype)	integer
cfPers_OrgUnit	cfPersId	Affiliation	cfOrgUnitId	Person	current MemberOf	Organization
cfPers_OrgUnit	cfPersId	Sub Affiliation	cfOrgUnitId	Person	current MemberOf	Organization
				Organization	hasSub Organization	Organization
cfPers_OrgUnit	cfPersId	Board-Member or TG-Leader	cfOrgUnitId	Person	hasLeaderRole	LeaderRole
				LeaderRole	roleContributesTo	Organization
				Person	currentlyHeadOf	Organization
cfProj_Pers	cfPersId	Coordinator	cfProjId	Person	hasOrganizerRole	OrganizerRole
				OrganizerRole	roleRealizedIn	Project
cfProj_Pers	cfPersId	Coordinator [fract=0.7]	cfProjId	OrganizerRole	description	Literal
cfProj_Pers	cfPersId	Participant	cfProjId	Person	hasResearcherRole	ResearcherRole
				ResearcherRole	roleRealizedIn	Project
cfProj_Pers	cfPersId	Participant [fract=0.3]	cfProjId	ResearcherRole	description	Literal
cfOrgUnit _PAddr	cfOrg UnitId	post-office-box	cfPAddrId	Organization	mailingAddress	Address
cfOrgUnit _EAddr	cfOrg UnitId	Email	cfEAddrId	Organization	email	not defined
cfOrgUnit _EAddr	cfOrg UnitId	Skype	cfEAddrId	not modeled		

4 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. As an approach to support interoperability and integration of the systems based in these models, this paper makes a comparison study between the CERIF relational database model and the VIVO ontology. 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.

Broadly analyzing the interoperability of both models reveals that VIVO does not support *multilingual* features as CERIF does. On the other hand, without considering the CERIF entities that are exclusively oriented to support the language features, the 209 VIVO classes provide a higher 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.

The detailed 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. Having 1700 attributes allows the CERIF model to represent very specific information like for example the Skype user assigned to a Person.

Acknowledgments. The work leading to these results has received funding from the European Commission under grant agreement nº 250525 corresponding to project VOA3R (Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment), http://voa3r.eu

References

1. Gibbons, M.: The new production of knowledge: the dynamics of science and research in contemporary societies. SAGE (1994)
2. Zimmerman, E.: CRIS-Cross: current research information systems at a cross-roads. In: Proceedings: Gaining Insight from Research Information, at the 6th International Conference on Current Research Information Systems, Kassel, Germany (2002)
3. Sicilia, M.-Á.: On Modeling Research Work for Describing and Filtering Scientific Information. In: Sánchez-Alonso, S., Athanasiadis, I.N. (eds.) MTSR 2010. CCIS, vol. 108, pp. 247–254. Springer, Heidelberg (2010)
4. Jörg, B.: CERIF: The common European research information format model. Data Science Journal 9, 24–31 (2010)
5. Krafft, D.B., Cappadona, N.A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B.J.: Vivo: Enabling national networking of scientists. In: Proceedings of the WebSci 2010: Extending the Frontiers of Society On-Line, Raleigh, US (2010)
6. Jörg, B., Ojars, K., Jeffery, K., van Grootel, G.: CERIF XML 2008 – 1.0 Data Exchange Format (2009)
7. Jörg, B., Jeffery, K., Dvořák, J., Houssos, N., Asserson, A., van Grootel, G.: CERIF 1.3 Full Data Model (FDM) - Introduction and Specification (2012)
8. Ivanovic, D., Surla, D., Konjovic, Z.: CERIF compatible data model based on format. The Electronic Library 29, 52–70 (2011)
9. 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)