Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 1563 – 1570

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Modeling the semantics of contextual and content-specific research metadata using ontology languages: issues on combining CERIF and OWL

Brigitte Jörg[a], Jaakko Lappalainen[b], Kostas Kastrantas[c]

[a]*German Research Center for Artificial Intelligence (DFKI GmbH), Alt Moabit 91c, 10559 Berlin, Germany*
brigitte.joerg@dfki.de
[b]*Computer Science Department, University of Alcalá Polytechnic Building 28871 Alcalá de Henares, Spain*
jkk.lapp@gmail.com
[c]*Greek Research and Technology Network (GRNET), 56 Messogeion Av., 115 27, Athens, Greece*
kkastrad@grnet.gr

Abstract

Current Research Information Systems (CRISs) enable the maintenance of information related to research activities of organizations and their members, including outputs or products from these activities. Such contextual information is of uttermost importance for the processing of datasets and with the retrieval of scientific documents, providing e.g. the key information on provenance and characteristics of research activities that are needed when searching for data or scholarly content. In the context of the expanding initiative of the Web of Linked Data, translating that information into semantic languages enables new ways of querying benefitting from the reuse of domain ontologies. In that direction, this paper reports on the engineering of an ontology-based version of the CERIF standard for CRISs using the OWL language and a proposed mapping to research datasets.

Keywords: CERIF; Research Information Systems; CRIS; ontologies; OWL; datasets

1. Introduction

CRISs (Current Research Information Systems) represent the research activities of organizations, following models that allow to expand e.g. to detailed accounts of organizational structures, researcher profiles, project outputs or received grants among other information entities (Jeffery & Asserson, 2008). These aspects may be considered as descriptions of the *context* in which the data and information of different kinds are produced. Research products can take a variety of forms; from raw data coming from observations to formal scholarly publications that summarize and report the main outcomes of the process. Existing standards provide suitable data models to represent the main research entities and their relationships for storage and exchange (Jörg et al., 2012a, b). They 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 (Jörg, Jeffery and van Grootel, 2011) but for techno-historical reasons they assume

doi:10.1016/j.procs.2012.04.171

the completeness of information within well-defined system boundaries. A more open-ended approach is therefore required to enable the linkage of entity descriptions with other systems that are curating research outputs. These include notably institutional repositories and open access systems, and more particularly, repositories curating datasets and constituting the basis for the sharing of scientific data, supporting experiment repetition and scientific data combinations and thus, facilitate the progress of science by leveraging research results to wider communities. However, advanced sharing of content specific data requires full semantic descriptions. This in turn entails, that the data about the production context is represented in an appropriate semantic form. While some of the representations of CRIS data in ontology languages have been proposed – including representations of datasets – the issues that have been raised for interlinking to be effective have still not been fully explored. This paper reports from a concrete experience in sharing research context information in CERIF[a] using Semantic Web languages that enable linking to semantic representations of research product metadata (bibliographic information and dataset descriptions). The key assumption is that it is possible (and likely) that contextual metadata are stored in a CRIS, while the more detailed metadata on research products (e.g. datasets) will be available in different systems. This requirement resulted in the turning to a Linked Data approach (Bizer, Heath and Berners-Lee, 2009) with support of a maximum flexibility for interlinking. The work presented here describes the general architecture for such a solution and the main mappings required to bridge context and content-specific information in a way that supports descriptions of methods of research production. One case about how a concrete dataset can be represented under the presented framework is also provided as an example illustrating the benefits and costs of moving to a more semantically rich representation covering all these aspects.

The rest of this paper is structured as follows. Section 2 provides a concise background on CRISs and the CERIF model as an underlying standard for CRIS data, as well as a brief overview of metadata for research datasets. Then, Section 3 describes an approach for bridging CRIS systems with research content repositories introducing a mapping between the relevant models. A real case is described in Section 4, illustrating how the description of a particular set of research data is semantically described. Finally, conclusions and outlook are provided in Section 5.

2. Background

To 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 modelling constructs: research entities, link entities and multilingual entities. Research entities such as person, project, organisation or publication have attributes such as identifier, acronym, gender, or ISBN; link entities account for relationships and their changes through time; multilingual entities allow for text recordings in multiple languages.

CERIF can be modeled in OWL using a number of conventions for the translation of CERIF Entity-Relationship expressions into elements of ontology languages. A straightforward translation represents base entities as OWL classes, e.g. the *cfOrganization* entity is translated into an *Organization* class. The same can be applied to result entities and other research entities. Further superclasses or subclasses such as the organizational types or roles in the relational CERIF are identified from and managed within the so-called CERIF Semantic Layer. The Semantic Layer is a flexible mechanism for e.g. the declaring of entity types and roles between CERIF entities. This is done (i) through a mechanism by which a *class* is used to declare at entity instance level the different types of a base entity organized in *class schemes*. For example, for a particular *cfOrganization*, there exists a *cfOrgUnit_Class* entity which allows to state, that a particular organization is for example of a sub-kind “Higher Education Institute”. A different approach is used for refining and specializing of link entities with roles (ii). Link entities in CERIF model the relations between entities; e.g. *cfPers_OrgUnit* represents relations between organization units and persons. A straightforward mapping to OWL is that of translating explicit link entity roles into object-type properties specifying as their domain and range the corresponding entities at both ends of the relation.

The above mentioned techniques (i) and (ii) produce a straightforward mapping from which a fragment is shown in figure 1. The main drawback of such a mapping in particular for the roles is, that all the potential relations are

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

defined at the level of the research entities and not as sub-classes like with entity types as such. This is evident when visualizing the relations using Protégé as shown in figure 1. However, this could be enriched by refining the CERIF listing of object-type properties as explicit subsumption classes. The outputs of research processes are modeled in result entities, namely publications, patents and the more generic "product". This latter entity can be used as a placeholder for all the different typologies of research outcomes not fitting in the publication and patent categories. Since CERIF aims at becoming a model with maximum flexibility for interoperability and extension, the specifics of different products are not part of the relational data model and thus syntax. Nonetheless, they could be included as subtypes related to the *cfResProd* entity in the CERIF Semantics, that is, in the vocabulary and therefore in an ontology. For example, a "dataset" typed through a given *ClassId* used in a *cfResProd_Class* would be enough to set apart those research products that can be considered primary observation data or secondary derived data.

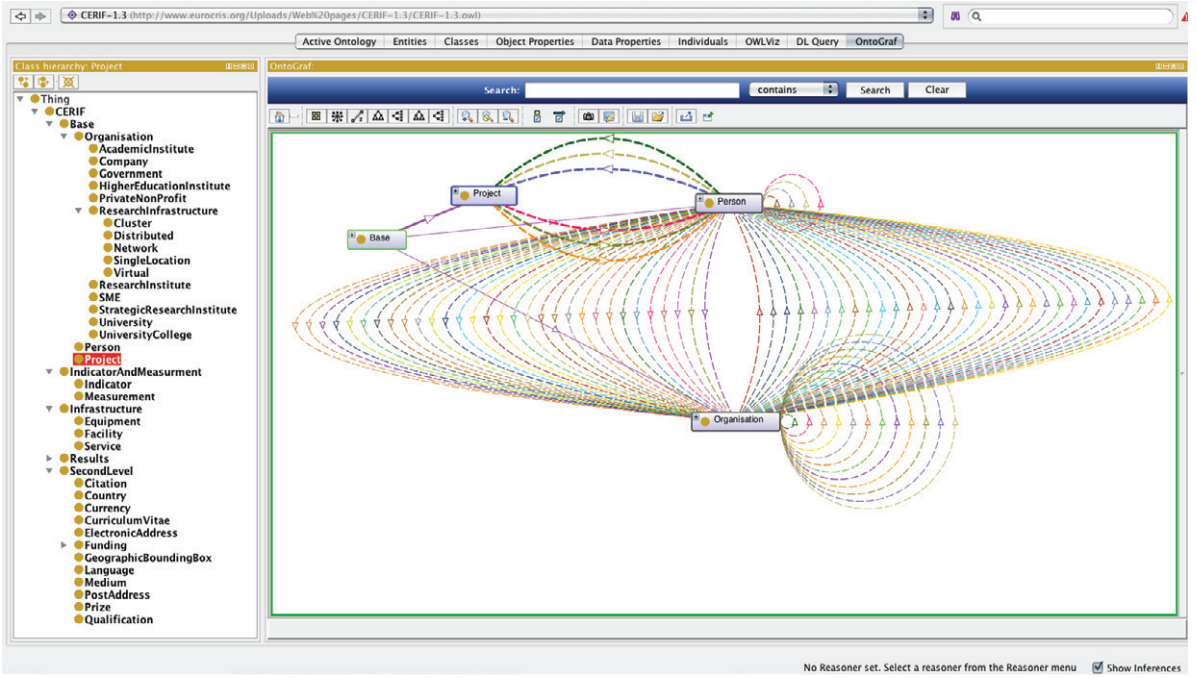

Figure 1. Screenshot of the Protégé tool showing part of the hierarchy of a CERIF-based ontology.

When considering research products, datasets are the focus of much attention nowadays for their importance in the transparency and reliability of research results. There are several metadata schemes dealing with the specifics of datasets. Some of them are extensions of existing general-purpose metadata schemes. Examples are the existing Dublin Core application profiles for datasets. The DCMI Type Vocabulary includes a Dataset term, and projects such as DataShare[b] have elaborated specific ways of using Dublin Core for data repositories. Other schemas are specific for the description of datasets, EML[c] being a relevant example. EML is a metadata specification implemented as a set of XML (Extensible Markup Language) modules enabling the documentation of ecological data in a modular and extensible way, where each module is defined to contribute essential information to describing the ecological data, as well as their recommended format. The Science Environment for Ecological Knowledge (SEEK) project[d] has developed and formalized critical aspects of EML in the Extensible Ontology for Observation (OBOE). The main approach followed was to extend EML ideas to allow for the semantic annotation of ecological data sets using ontologies. OBOE has also been developed within the Semantic Tools for Data Management

[b] DataShare Project: http://www.disc-uk.org/datashare.html
[c] Ecological Metadata Language (EML): http://knb.ecoinformatics.org/software/eml/
[d] Science Environment for Ecological Knowledge (SEEK) project: http://seek.ecoinformatics.org/

(SEMTOOLS) project[e] for describing a wide range of ecological datasets stored within the Knowledge Network for Biocomplexity (KNB), as well for extensions of ontology based data annotation and discovery within the MetaCat software infrastructure[f]. OBOE is an extensible set of ontologies represented in a formal ontology language (OWL-DL) that serves as a way to describe scientific observations and opens up the possibility of sharing, integrating and discovering all the datasets even though their contexts are different, because OBOE is not domain specific.

3. Combining CRIS data with scholarly content-specific metadata

This section describes the approach taken in the VOA3R[g] project for combining CRIS data and research output data. Following the assumptions that different kinds of data will be separately stored in different systems, a Linked Data (LD) approach is taken.

3.1. Overall Approach

The framework of VOA3R is based on the principles of LD, and as such it is relying on a combination of terminologies/ontologies exposed openly and independent, and furthermore on the exposure of datasets that make use of these terminologies/ontologies by means of linking. Figure 2 provides a simplified view of the framework used in VOA3R for the combination of ontologies/terminologies.

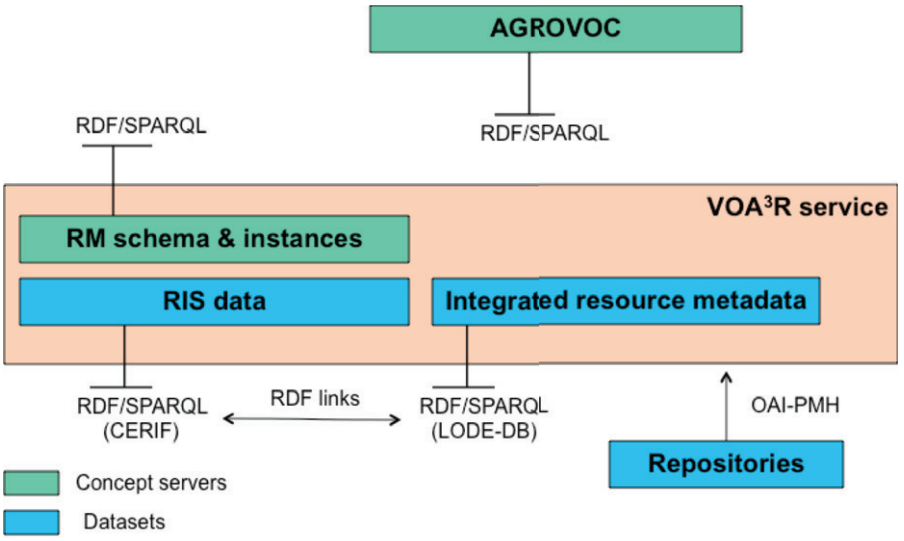

Figure 2. Overall architecture supporting ontologies and semantic representations of datasets.

Figure 2 shows how external terminologies are decoupled from the service. FAO AGROVOC[h] is depicted in the diagram because VOA3R is focusing on agriculture, but any terminology exposed as linked data could be used. The case of the research methods (RM) ontology described later in this document is special, as it will be curated from inside the service, given that existing ones do not cover it. However, functionally it is equivalent to other concept servers. The main architectural ideas are that RIS data is residing in a separate repository, which may be a CERIF-

[e] The SEMTOOLS project: https://semtools.ecoinformatics.org/
[f] The Knowledge Network for Biocomplexity (KNB): http://knb.ecoinformatics.org/knb/metacat
[g] Virtual Open Access Agriculture & Aquaculture Repository (VOA3R) project: http://voa3r.eu/
[h] AGROVOC 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

based relational database. Then, a separate integrated repository is storing the metadata (and eventually the contents themselves) for different kinds of research products. These in VOA3R are coming from institutional repositories via open harvesting standards and using different kinds of content-specific metadata, e.g. EML for datasets or MODS[i] for bibliographic entries. In any case, both databases are providing de-referenceable URIs for each of the entities exposed, enabling RDF linking. For example, a *cfPerson* in the RIS repository could be identified as the producer or creator of a particular dataset (exposed in the other repository) or as an author for a particular paper.

3.2. The Research Methods Ontology

Bridging contextual information regarding persons, organizations and projects with research output can be done through direct linking. However, a richer representation should consider research activities as an intermediate entity. Figure 3 depicts the main elements of the mappings required for bridging the two kinds of entities.

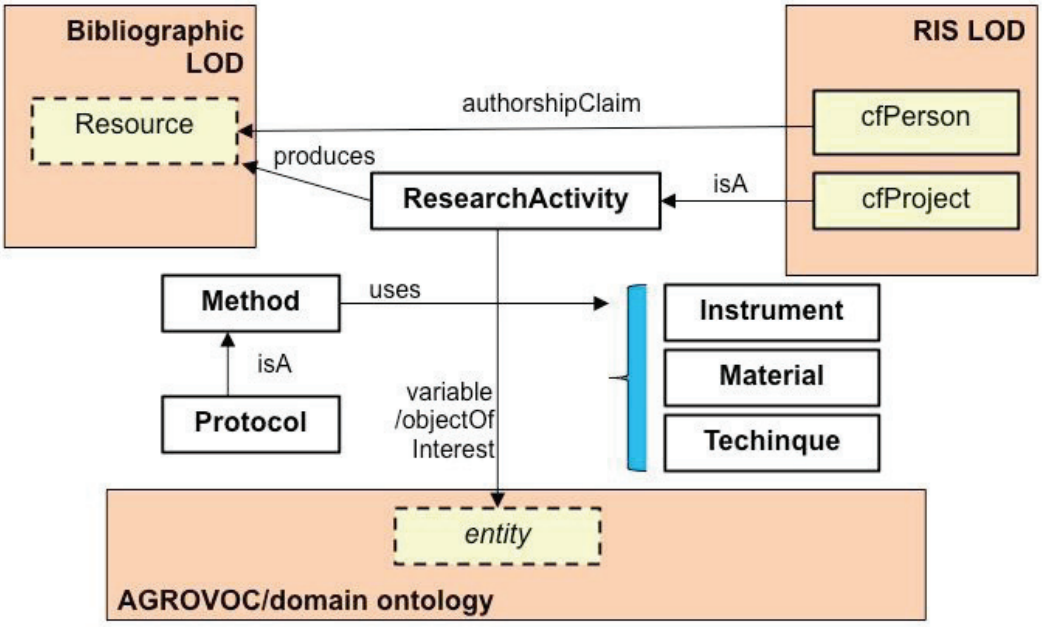

Figure 3. Main elements of the resulting ontological mapping

A small research methods (RM) ontology is used as an intermediate model. A generic concept *ResearchActivity*[j] is used to model a variety of potential activities. That concept can be related to CERIF as subsuming *cfProject*, which is the base entity representing an organized research activity. For bibliographic or other kinds of elements, a *produces* predicate is used as a generic means of mapping. Then, the description of research activities can be done independently from contextual and content-specific aspects, and elements related to research methods, protocols and others can be described as a part of the research activity. There are some additional mappings possible. For example, *Instrument*s used in research activities can be mapped to the CERIF *Equipment* entity if information for a concrete instrument used (not only indicating the type) is required. Research activities can be further specified as particular kinds of activities performed by researchers. However, instead of developing a complete model, we have decided to

[i] Metadata Object Description Schema (MODS): http://www.loc.gov/standards/mods/
[j] Research Activity is also a proposed name in the CASRAI vocabulary: http://dictionary.casrai.org/research-activity-profile-v0.9-draft/1.1.0

find some common semantic ground in existing ontologies. Concretely, we used one of the most well known and largest ontologies available: the common sense Cyc ontology (Lenat 1995).

The use of a RM ontology as a mediator between contextual and content-specific information can also be combined with the reuse of existing terminologies and/or domain ontologies exposed as LD for different purposes. For example, they can be used to describe the dependent and independent variables for a particular hypothesis contrast associated to a particular dataset. This kind of descriptions move a step ahead in providing computational support to automate data processing, e.g. by combining datasets describing the same observation but at different points in time.

4. A case combining contextual and content descriptions

As an illustration of the use of the mappings proposed, in this section an example is briefly sketched. The dataset considered is the study mentioned by San Gil, Vanderbilt and Harrington (2011) that in turn refers to a climate study published in Nature in 2002 (Doran et al., 2002). It showed how the average air temperature at the Lake Hoare weather station in East Antarctica dropped over the last 20 years of the last millennium.

The following are the main elements of the representation of contextual information regarding the dataset.

EML element	CERIF representation	Description
`creator`	*cfPerson* *cfResultProduct* *cfPerson_ResultProduct* *cfClass=creator* *cfStartDate/cfEndDate*	The role of creator or "owner" of the dataset requires a Person (or an Organization) entity and a corresponding link to the result or product representing the dataset. From within the link, a *classification as applied through the term* "creator" from e.g. the CERIF vocabulary is appropriate for the role of the dataset creation at a particular time.
`instrumentation`	*cfEquipment* *cfEquipment_Class* *cfClass=instrument* *cfResultProduct_Equipment* *cfClass=eg:FirstMeasurement* *cfStartDate/EndDate*	Instrumentation as the following can be mapped to an instance of the *Equipment* entity where "instrument" would be the classification type of the equipment itself. The explicit link with the dataset *Product* is then established, for which further roles or classes according to time ranges could be added: `<instrumentation>1993-1994 - 1999-2000: Campbell Scientific 207 temp/rh probe.</instrumentation>`

The described mappings are examples of potential links from a CRIS to a dataset repository. These could be resolved either by direct linking or via intermediate RM ontology entities if elements of the methodology appear in the method description of e.g. the EML dataset. In addition to the above, the dataset-specific information includes approximate location in `<boundingCoordinates>` and temporal extent in `<temporalCoverage>`. Where temporal coverage in the relational CERIF model is considered with all link entities by cfStartDate/cfEndDate attributes, the geographic bindings do currently refer to postal addresses of equipment through the CERIF link entity *cfPostAddress_GeographicBoundingBox*. The temporal aspects and the recently introduced geospacial aspects in the relational CERIF model (1.3) as well as the indicators and measurements have not yet been entirely considered in the presented CERIF ontology.

Assuming a translation of observation information into an ontology form following the OBOE ontology model (Madin et al., 2007), the dataset would require some additional semantic information about the *entity* being observed and for the *measurement* taken.

5. Conclusions and outlook

The VOA3R project has adopted an approach to reuse terminologies and domain ontologies based on the usage of externally curated and maintained services. This makes the service fully extensible in terms of incorporating other terminologies and ontologies since all of them are reused via linking. The AGROVOC ontology has proved to be sufficiently comprehensive and broad enough to cover the descriptions of the inspected sample resources, and it has been selected as the main means for terminology-based browsing, resource annotation and term extraction. However, other Knowledge Organization Systems (KOS) can be integrated in the system seamlessly due to the set of governing principles that require them to be externally available.

Bibliographic descriptions and production context information (organizations, persons, projects) have been approached by building on existing best practice, again in an attempt to make the service sustainable in the long term. In the case of bibliographic information, the LODE-DB recommendations provide the basis for the provision of semantically consistent data accounting for the wide heterogeneity of metadata coverage and quality, currently available in different institutional repositories and digital collections. In the case of research context information, VOA3R has adopted the strict, well-defined semantics in the CERIF standard, and has started work inside euroCRIS, to develop shared recommendations for exposing data using the LOD paradigm. Through collaboration between the LOD and CERIF task groups, a CERIF ontology is planned to be presented before the end of the year, taking into account existing approaches and related initiatives.

The only aspect of research work that has been considered to deserve separate treatment has been the description of the methods, protocols, instruments and materials. A core research methods ontology has been devised (Sicilia 2010) that allows for the description of key aspects of the scientific methods and can be combined with observation data to come up with rich semantic descriptions that set the basis for future more advanced functionality beyond annotation. That ontology has been defined at the level of the schema (t-box) only, and will be populated progressively at the instance level during the lifetime of the service either by the interested communities or by experts devoted specifically to the task. In the case of scholarly content that contains data, the OBOE ontology (Madin et al. 2007) has been linked to the overall schema allowing for computations, that set the basis for future more advanced scenarios including computing with semantically consistent information. Such kind of computations may in the future cover experiment repetition or reactive algorithms that test hypothesises against incoming data. These are currently possible to be done only with human intervention even in domains in which dataset description is a common and regular practice like in the environmental sciences (San Gil et al., 2011).

Acknowledgements

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. The coordinators of the VOA3R project provided examples and cases included in the paper that were originally developed for VOA3R deliverables. The research described in this paper had the support of EU FP7 project agINFRA (http://aginfra.eu/), contract number 283770.

References

1. C. Bizer, T. Heath and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*. (IJSWIS), 5(3): 1-22 (2009).
2. P.T. Doran, J.C. Priscu, W.B. Lyons, J.E. Walsh, A.G. Fountain, D.M. McKnight, D.L. Moorhead, R.A. Virginiaq, D.H. Wall, G.D. Clow, C.H. Fritsen, C.P. McKay, and A.N. Parsons. Antarctic climate cooling and terrestrial ecosystem response', *Nature*, Vol. 415, pp.517–519 (2002).
3. K. Jeffery and A. Asserson. Institutional Repositories and Current Research Information Systems. *New Review of Information Networking*, 14(2), 71-83, (2008).
4. B. Jörg, K. Jeffery, J. Dvorák, N. Houssos, A. Asserson, G. van Grootel, R. Gartner, M. Cox, H. Rasmussen, T. Vestdam, L. Strijbosch, A. Clements, V. Brasse, D. Zendulkova, T. Höllrigl, L. Valkovic, A. Engfer, M. Jägerhorn, M. Mahey, N. Brennan, M.-A. Sicilia, I. Ruiz-Rube, D. Baker, K. Evans, A. Price and M. Zielinski, CERIF 1.3 Full Data Model (FDM): Introduction and Specification (euroCRIS 2012).
5. B. Jörg, J. Dvorák, T. Vestdam, G. van Grootel, K. Jeffery and A. Clements, CERIF 1.3 XML Data Exchange Format Specification (euroCRIS 2012).
6. B. Jörg, K. Jeffery and G. van Grootel, Towards a Sharable Research Vocabulary (SRV) – A Model-driven Approach, in *Proc. Metadata & Semantics Research Conference*, Yasar University, Izmir, Turkey, Springer Berlin, Heidelberg, (2011).
7. D.B. Lenat. Cyc: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM, 38, 11 (1995), 33-38.
8. J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. An ontology for describing and synthesizing ecological observation data, *Ecological Informatics* 2: 279–296, (2007).
9. R. Rice. Dublin Core for Datasets: DISC-UK Approach. *Metadata for Scientific Datasets Workshop*, Berlin, 25 September, (2008).
10. I. San Gil, K. Vanderbilt and S. Harrington Examples of ecological data synthesis driven by rich metadata, and practical guidelines to use the Ecological Metadata Language specification to this end. *International Journal of Metadata, Semantics and Ontologies,* 6(1), pp. 46-55, (2011).
11. M.-A. Sicilia, On Modeling Research Work for Describing and Filtering Scientific Information, in *Proc. Metadata and Semantic Research – Communications in Computer and Information Science* (Springer 2010) pp. 247–254.