Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

Say, "S" (as) = semantics – and mean it! Path to semantically interoperable digital research services

Suvi Remes[a] , Miika Alonen[b] , Patrik Maltusch[c], Mikael af Hällström[d], Stina Westman[e]*

[a]*CSC – IT Center for Science, Keilaranta 14, Espoo FIN-02101, Finland*
[b]*CSC – IT Center for Science, Keilaranta 14, Espoo FIN-02101, Finland*
[c]*Aalto University, PO BOX 11000, Aalto FIN-00076, Finland*
[d]*Finnish Tax Administration, PO BOX 325, Vero FIN-00052, Finland*
[e]*CSC – IT Center for Science, Keilaranta 14, Espoo FIN-02101, Finland*

Abstract

The more we invest in open science and research, the more we need to ensure that metadata enabling discovering and digital preservation of research material is of high-quality and semantically coherent. Still, interoperability of information systems and the lack of shared semantics, both between humans and machines, is an internationally recognised issue.

In Finland we are in the process of implementing information systems and harmonising the legacy data models in the way that it makes use of the shared semantics, standards and other best practices according to the common architectural vision. This basic infrastructure for information management is built by combining terminological theory, linked data and adaptable data modelling practices. The idea of the Semantic Interoperability Model and new tools, IOW – Interoperability Workbench, supporting it are presented in the context of research and science in Finland, but the vision of the linked information components is generic.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Keywords: semantic interoperability; terminological theory; linked data; metadata modelling; information architecture

* Corresponding author. Tel.: +358 50 320 4792; fax: +358-(0)9-457-2302.
E-mail address: suvi.remes@csc.fi

Peer-review under responsibility of the Organizing Committee of CRIS2016.

1. Introduction

Many efforts have been made to define best practices and guidelines for interoperability in the field of research. In Finland, a national data model has been defined to support data storage, data collection and reporting activities in the field of higher education and research administration. The XDW model[1] is designed to support institutions in implementing a data warehouse. Data in the data warehouse can be any functional data, ranging from students' grades to financial indicators or research publications. Mapping from the XDW model to the CERIF model[2], developed by the euroCRIS community, was studied on the conceptual level, but formal mapping is not supported in the current XDW implementation.

The loosely CERIF compatible XDW model itself has been used as a starting point in data warehouse implementations. This kind of a database approach to information management is typical, but architecture based on traditional databases which have their own data models require lot of harmonisation and high cost expert labor work to move the data from one information system to another. Currently more focus is needed in support for more adaptable data modelling and invocation of linked data practices, and on development of interfaces and data transfers between information systems — all including semantics.

The new approach (Fig. 1) to data management and interoperability we describe in this paper is a mix of existing practices, such as utilising core vocabularies as proposed by Charalabidis et al. (2010)[3] and documenting the use of such data vocabularies by defining application profiles as first defined by Heery and Patel (2000)[4]. Semantic technologies and linked data are used to describe machine-readable terminology, core vocabularies and application profiles. Expectations for the machine-readable application profiles as argued by Diane I. Hillmann and Jon Phipps (2007)[5] are realized by using shape expressions as described by Prud'hommeaux et al. (2014)[6]. Framework forms a structured, common architecture for connecting conceptual modelling of business, services and processes to defining and maintaining controlled terminology and further to constructing data models for information systems. It also offers Finnish agents operating in the field of research and science a new way to create and maintain linkages to relevant existing international work and resources.

2. Framework for information architecture

2.1. Concept modelling using terminological method

The modelling process of any information system should start from conceptual modelling. This is generally accepted statement, but according to our experience the actual meaning of it seems to vary. Moreover, what is usually missed at this stage is a systematic and formalised method for concept defining. It is not that we do not have them, it is that we have not fully recognised the value of terminological theory[7]. This originally humanistic approach, typically used in cross-human communication, argues that concepts within a subject field are interrelated and form concept systems. And we go on arguing that this shared understanding, the concepts we use in business and operations of agencies, should form the solid foundation also for semantics of data models used in information systems. The point is that these terminologies, resulting from terminological concept modelling, are domain specific, aimed at concept clarification, mutual understanding of concepts and consistent use of terms. They do not specify database or information system specific information or their mutual relationships, resulting from concept modelling[8]. And further, these terminological concept systems are adapted to the language and cultural realm they were created in. In multilingual terminologies and communication situations careful attention should be paid to glosses.

The controlled, methodologically intact terminology should be openly available both in human and machine readable formats. In Finland, terminologies are currently published mainly in printed or digital booklets or in term banks. These medias do not support machine readable formats, such as SKOS[9], required to apply terminologies as the cornerstone of semantics in information system development according to the architectural vision described in this paper. In Finland, however, situation is good as we already have a suitable service available: SKOS vocabularies and

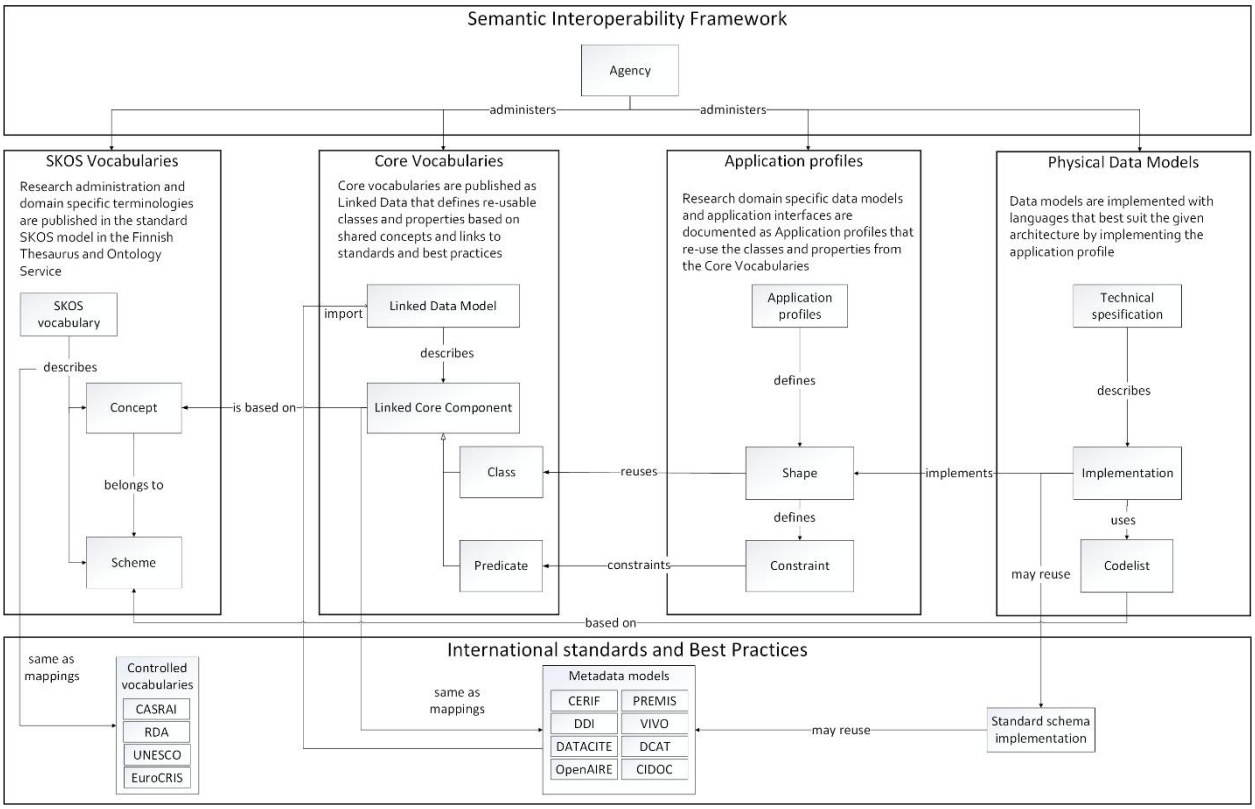

Fig. 1. Illustration of the semantic interoperability framework.

classifications will be published in the *Finnish Ontology Service*[10], maintained by the National Library of Finland. Finto service is originally designed to support publication of thesaurus ontologies needed in e.g. indexing material for searching and browsing purposes[11]. It is already possible to publish also terminologies as machine readable terminological ontologies in the Finto service, but it is worth noticing that conversion into SKOS format causes information loss compared to the amount of linguistic information terminology usually provides. Using SKOS format the information of concept definition and additional description, preferred term and alternative terms and, if available, their language versions in selected languages are provided.

With the support of Ministry of Education and Culture an inter-institutional working group of research domain experts is building a national terminology for research administration in Finland. Some widely recognised international sources (see Fig. 1.) are used as references. Formal linkages to the selected sources will be created later once the terminology is ready for publishing. However, the Finnish Open Science and Research Initiative (ATT)[12], also set out by Ministry of Education and Culture, already took advantage of the first terminology draft as a description of the operational environment when planning national level services for researchers and research institutions enabling more open science and research.

2.2. How to connect terminology and logical data models

Terminology work, however, takes no role in defining logical data structures, such as classes, properties and part-of relations. For this purpose core vocabularies are needed to connect the concepts and logical data models together and pass the shared semantics to every implementation that reuses its components. Core vocabulary, actually a linked

data vocabulary which we have named as *Semantic Interoperability Model* to emphasise its intermediary role, identifies the reusable information components with URIs (Uniform Resource Identifiers) and allows them to be used when creating application profiles for specific use cases. In the research context in Finland, a few central components were defined at the time of creating the enterprise architecture for open science and research[13]. Further work includes not only generation a few more components but also formal mappings with selected international metadata models.

When creating application profiles by reusing the elements from the core vocabularies, the semantics from the research domain experts built terminology follows. It is possible further define or focus the definition of the concept used in the application profile but not in a way that this changes the originally agreed meaning. Application profiles are constructed and documented in human and machine-readable formats following the guidelines by the Dublin Core Metadata Initiative (DCMI)[14] and by using formal constraints and restrictions as classified by Boch et al. (2015)[15]. Machine readable application profiles reuses the descriptions and mappings defined in the core vocabularies, and formal constraints make it possible to transform the reusable components into different standard technical formats, such as XML or JSON schemas — database scripts as with the national XDW model are also possible to make available in future. All this ensures the validity of the data by defining constraints and validation rules. Workflow for data modelling is visualised in the Figure 2.

This approach makes possible a clear distinction between concepts and logical data elements and further logical models in specific implementations and therefore clarifies the management responsibilities of the different resources. The success of this kind of distribution of work, however, requires commitment to collaboration.

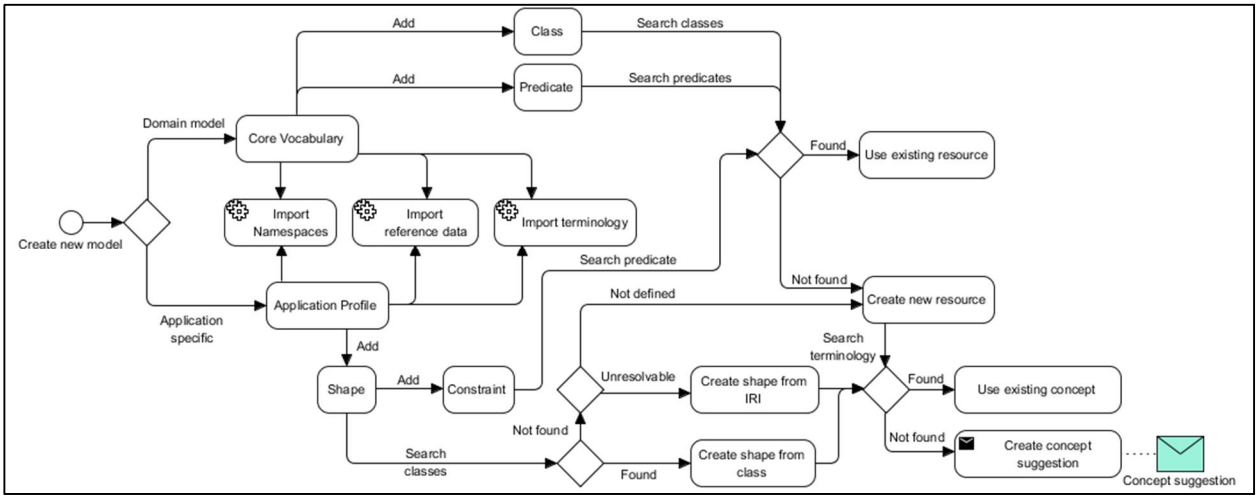

Fig. 2. Simplified workflow for data modelling.

3. New tool(s) to support collaborate data modelling and reuse of resources

One of the goals of our work has been to build a prototype for collaborative online tool for creating linked data vocabularies and application profiles — instead of documenting terminology and data model descriptions, as usual, in e.g. separate wiki pages or spreadsheet. The aim is to create “one-shop stop” for vocabularies and logical data models and allow formal linkages between them. There already exist data governance tools and some tools for developing and maintaining terminologies and this new tool is planned to be used in conjunction with them. The main issue with existing tools is that they are usually planned to be used within a single organisation. The created resources are not efficiently shared, nor identified with machine-readable identifiers or interlinked. Even if they are made publicly and openly available, the necessary information is dispersed in various web locations. For example, the

linkage to the source and information about updates made to the resources are almost impossible to manage. Enterprise architecture tools and various schema editors are, however, still needed to be used for data management purposes and more detailed technical specifications.

Our development work has proceeded and we are proud to present the solution which we call: *IOW - Interoperability Workbench*[16]. IOW consists of a set of tools designed to be used in developing, documenting and publishing re-usable core components and application profiles. The service is envisioned as a one-stop service for openly published data descriptions and vocabularies. These semantically coherent data models help manage information exchange operations (integration, interface, service bus) as well as database and software development. The system and its features are planned in a way that enables a content specialist – not always IT-specialist – to build application profiles and link the models online to Linked Data and SKOS vocabularies published e.g. in the Finnish Ontology Service (Finto). In the future, the content of these core vocabularies and application profiles will be jointly produced by experts associated to each subject area, e.g. research or education. IOW is developed as open source software and as a modular implementation which enables further development, as linkages e.g. to existing code services. The information descriptions published in the IOW may be accessed and browsed freely.

At the moment, we are applying the described modelling practices and the new IOW tool in publishing work done in the national open science framework. This includes the previously mentioned core components, but following the principles described in this paper we are also defining a *Research Data Catalog Vocabulary*[17] – utilizing the IOW tool. The model, still in draft form at the time of writing, is based on the European Commission's DCAT-AP specification and is adapted to suit the field of science and research. The linkages to international vocabularies seem to be best supported using the new tool. The model will also be formally linked with the national research terminology as soon as the terminology comes available in machine readable format. Once completed, the *Research Data Catalog Vocabulary* is planned as the master data model for national research data services. This will enable researchers, research groups and organisations to easily publish metadata on their datasets and offer them for wider use.

4. Conclusion

In this paper we describe an approach that forms a structured, common architecture to information management by connecting terminological work and data modelling and enabling reuse of these data descriptions in a semantically interoperable and sustainable fashion. The method, based many ways on the international standards, ensures that shared definitions are applied in a systematic way in IT system development. To support efficient implementation of the method a one-stop-service for openly published data descriptions and vocabularies has been developed and is currently tested with first real pilots. Aims are to reduce the amount of time and recourses needed for data modelling for digital research services and at the same time ensure the quality of used descriptions by using terminological methods.

Open and reusable terminologies and data specifications suit well into the open science mindset. As stated in the *Open Science and Research Handbook*[18], by making resources and materials more widely available we can also increase the potential for innovation. The closer the used terminology in IT systems supporting researchers is to common research parlance, the greater is the possibility that the services based on these specifications are both semantically interoperable and, even more important, understandable to their users.

The Semantic Interoperability Model and IOW are originally developed by Ministry of Education and Culture for supporting the needs of institutions of higher education and research, but the work will in future form the basis for a wider National Metadata Service. This same setup will allow also public sector organisations to harmonise their own vocabularies and internal data models and, by doing so, make the most of the National Service Architecture[19]. We believe that this approach to interoperability and semantics of information and digital services has potential also in the context of international cooperation.

References

1. RAKETTI-XDW. XDW Model. Retrieved from *http://tietomalli.csc.fi/*
2. euroCRIS. CERIF. Retrieved from *http://eurocris.org/cerif/main-features-cerif*
3. Charalabidis Y (ed.). *Interoperability in Digital Public Services and Administration: Bridging E-Government and E-Business*. Hershey PA: Information Science Reference Doi 10.4018/978-1-61520-887-6. 2010.
4. Heery R, Manjula P. Application Profiles: Mixing and matching metadata schemas. *Ariadne 2000:25*.
5. Hillmann DI, Phipps J. *Application Profiles: Exposing and Enforcing Metadata Quality*. DC-2007 Conference Proceedings.
6. Prud'hommeaux E, Labra Gayo JE, Solbrig H. *Shape expressions: an RDF validation and transformation language*. Proceedings of the 10th International Conference on Semantic Systems. ACM, 2014.
7. International Organization for Standardization. *Terminology work -- Principles and methods* (ISO 704: 2009).
8. Madsen BN & Thomsen HE. Concept modeling vs. data modeling in practice. In Steurs F, Kockaert H, editors. *Handbook of Terminology*. *Vol. 1*. Amsterdam: John Benjamins Publishing Company, 2015. p. 250-275.
9. World Wide Web Consortium (W3C). SKOS - Simple Knowledge Organization System. Retrieved from *https://www.w3.org/2004/02/skos/*
10. National Library of Finland. Finnish Ontology Service (FINTO). Retrieved from *http://finto.fi/fi/*
11. Suominen O, Pessala S, Tuominen J, Lappalainen M, Nykyri S, Ylikotila H, Frosterus M, Hyvönen E. Deploying National Ontology Services: From ONKI to Finto. In Proceedings of the ISWC 2014 Industry track. Retrieved from *http://seco.cs.aalto.fi/publications/2014/suominen-et-al-deploying-onki-finto-2014.pdf*
12. Open Science and Research Initiative (ATT). Retrieved from *http://openscience.fi/*
13. Open Science and Research Initiative (ATT). Avoimen tieteen ja tutkimuksen viitearkkitehtuuri. Retrieved from *https://avointiede.fi/viitearkkitehtuuri (*English version coming soon)
14. Dublin Core Metadata Initiative (DCMI). Guidelines for Dublin Core Application Profiles. Retrieved from *http://dublincore.org/documents/profile-guidelines/*
15. Bosch T, Nolle A, Acar E, Eckert K. *RDF Validation Requirements - Evaluation and Logical Underpinning*. Computing Research Repository (CoRR), abs/1501.03933. 2015
16. IT – Center for Science (CSC). Interoperability Workbench (IOW). Retrieved from *http://iow.csc.fi/#/*
17. Open Science and Research Initiative (ATT). Research Data Catalog Vocabulary (draft). Retrieved from *http://iow.csc.fi/#/model?urn=http:%2F%2Fiow.csc.fi%2Fns%2Fatt*
18. Open Science and Research Initiative (ATT). Open Science and Research Handbook. Retrieved from *http://openscience.fi/handbook*
19. Ministry of Finance in Finland. National Architecture for Digital Services. Retrieved from *http://vm.fi/en/national-architecture-for-digital-services*