Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 106 (2017) 335 – 342

Procedia Computer Science

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

Governance of research information and classifications, key assets to interoperability of CRIS systems in inter-organizational contexts

Sadia Vancauwenbergh[a,b]*

[a]*Centre for Research & Development Monitoring (ECOOM), Division ECOOM-UHasselt, 3500 Hasselt (Belgium)*
[b]*Hasselt University, Research Coordination Office, Martelarenlaan 42, 3500 Hasselt (Belgium)*

Abstract

Research information systems are often described as important tools to enhance open innovation, by directly gathering and unlocking information on scientific and technological research to stakeholders. Current Research Information Systems (CRIS) systems are mostly developed within a single organization with a context-specific vocabulary and generally use the Common European Research Information Format (CERIF) for data storage. Although CERIF allows for almost unlimited possibilities to model research information, it has - like any standard - limitations when it comes down to communication to end-users in terms of semantics. However, if one wants to exchange the information kept within CRIS systems, it is essential to include research information and classification governance as this is key to truly comparable and thus meaningful information. In this paper, we elaborate on research information and classification governance as a prerequisite for establishing true semantic interoperability of CRIS systems in inter-organizational contexts.

Peer-review under responsibility of the Organizing Committee of CRIS2016

Keywords: Research information systems; interoperability; classification governance; business semantics management; inter-organizational contexts

* Corresponding author. Tel.: +32-11-269109
E-mail address: sadia.vancauwenbergh@uhasselt.be

Peer-review under responsibility of the Organizing Committee of CRIS2016
doi:10.1016/j.procs.2017.03.033

1. Introduction

Over the past decade, an increasing need for reusing and disseminating research information to key stakeholders has emerged. As such, researchers need to report on a regular basis on their research findings to funding providers and their host institutions, which on their turn need to report to governmental organizations. Mostly, each report has different requirements with regards to the formats and classifications used, thereby increasing the administrative burden put on the research community even more[1]. These research reports are not only used as justification to the budgets used by the researchers, but are also an important instrument in the foundation of new research policies. In order to facilitate an efficient manner of storing and exchanging data for these purposes, many organizations have developed or adopted a CRIS system, which in most cases comes down to a closed-world information system. These systems are characterized by a data model that represents the structure and integrity specification of the data of only the applications used by that specific organization[2]. As a consequence, the terminology used in the data models has an organization-specific semantic meaning. When such closed-world information systems are requested to exchange research information, an important semantic interoperability issue arises as the systems mostly use a different terminology for a similar concept or, alternatively use a similar terminology for a different concept. In an attempt to overcome these issues, the European Union has recommended all universities to use the Common European Research Information Format (CERIF) as a standard for the storage and exchange of current research information[3].

2. CERIF, the standard for storage and exchange of research information

The CERIF standard uses Entity-Relationship (E-R) modelling techniques to capture research information in base entities (Person, Project and Organisation Unit) that are linked via time-stamped relations to other entities (Fig. 1).

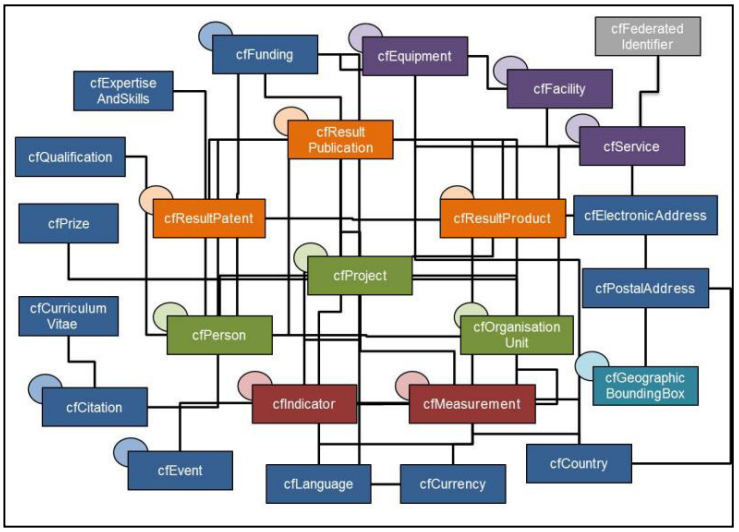

Figure 1: Overview of the CERIF 1.6 E-R model. Research information is modelled in base entities (green), result entities (orange), infrastructure entities (purple) and so-called second level entities (blue) that represent the contextual environment around the base and result entities. All entities are linked together via time-stamped relations, that are described via link entities or classification references.

These include result entities (ResultPatent, ResultProduct and ResultPublication), infrastructure entities (Facility, Equipment, Service) and so-called second level entities, presented as contextual environment around the base and result entities. For each of these entities, a wide variety of attributes is provided in CERIF to describe the accompanying metadata and to allow for multilingual features. In addition, the relationships used in CERIF are always semantically enriched by a time-stamped link entity or classification reference. In brief, the classification record is maintained in a separate entity and requires an assignment to a classification scheme, commonly referred to

as the CERIF Semantic Layer[4]. Altogether, CERIF 1.6 disposes over 293 entities and 1814 attributes, linked together via 665 possible relationships to store research information. Obviously, this allows for an almost unlimited flexibility to model research information. At the same time, this flexibility creates more complexity on the E-R model, making it less efficient for communicating the modeled domain knowledge to domain experts and end-users[2]. However, true semantic interoperability can only arise when the business experts use exactly the same terminology and corresponding meaning as in the CERIF scheme.
The actual practice shows that semantic mismatches in CERIF do occur, in particular at the level of the terminology and relations used. Examples of these mismatches are shown in Fig.2. In panel 2A, organizations use a different relation, i.e. an English or Dutch classification scheme, to denote exactly the same kind of funding received from funding organizations. Panel 2B illustrates the use of a different term from a single classification scheme, to indicate exactly the same kind of funding.

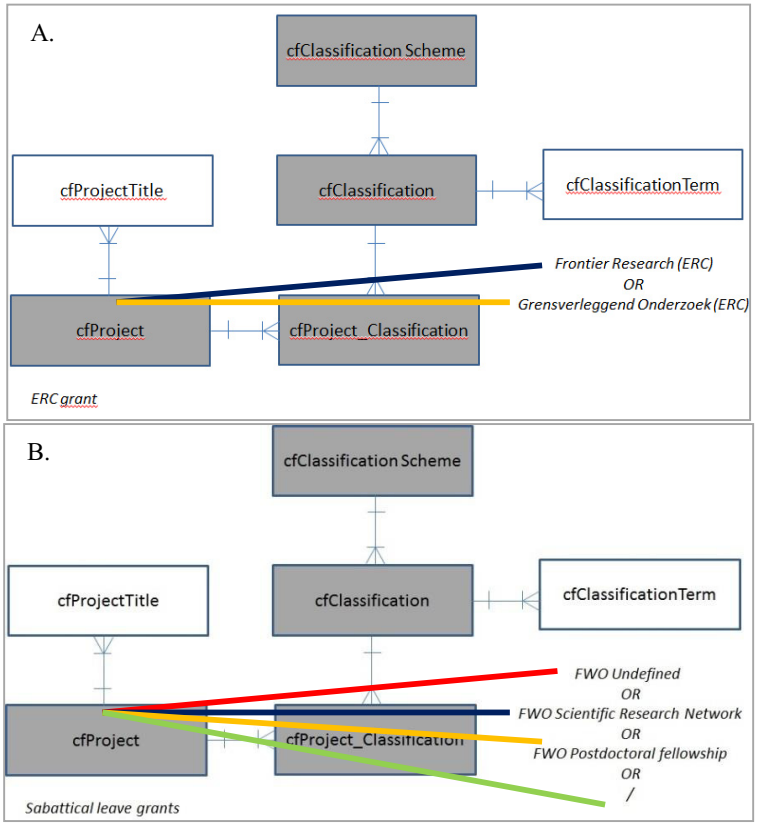

Figure 2: Mismatch at the relationship and terminology level. Panel 2A: Organization applications can use a different relation, i.e. English or Dutch, between a financial code concept and a project to denote exactly the same kind of funding. Panel 2B: Organization applications can use different financial code terminology to denote exactly the same kind of funding.

Both examples illustrate that the CERIF model indeed allows for great flexibility for modelling research information. However at the same time this flexibility can give rise to interoperability problems as heterogeneous representations can be modelled. Furthermore, misalignments can come from the fact that the CERIF semantics, which is modelled via the E-R scheme, is not easily understandable for domain experts leading to problems in translating it back to the CERIF conceptual level[2]. In order to overcome these issues, the business semantics management (BSM) methodology should be followed when using the CERIF standard for data storage and exchange in CRIS systems (Fig. 3). Business semantics management encompasses a set of prescribed steps and processes that brings together domain experts that collaboratively agree upon a common semantic vocabulary to describe metadata in iterative semantic reconciliation cycles. Furthermore, BSM includes the application of the

derived semantic patterns in order to establish semantic alignment with the underlying data structures[5]. Obviously, the BSM cycle should preferentially be repeated in case changes occur in the semantic applications, thus allowing for validation and feedback between the cycles, also known as full-cycle BSM[6].

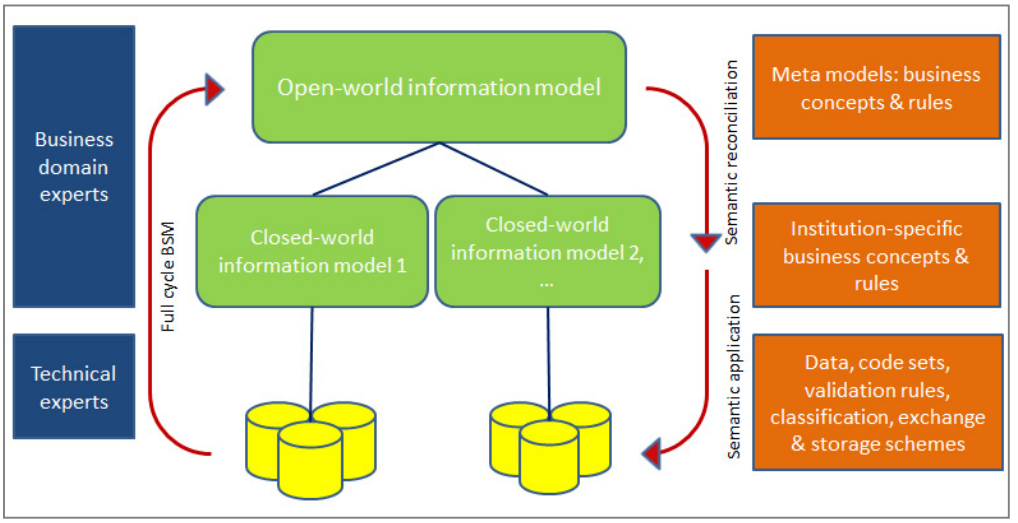

Figure 3: The Full-cycle Business Semantics Management methodology (BSM). BSM consists of iterative cycles of semantic reconciliation and the alignment of the resulting semantic pattern with the underlying data structures. When changes in the underlying applications act as starting point for new iterative cycles of semantic reconciliation, one can speak of full-cycle BSM.

Currently, many software packages exist that offer full-cycle BSM solutions. As these mostly have integrated solutions for communicating the semantic knowledge back and forward with applications, we will focus here on the implementation of semantic reconciliation of research information in an inter-organizational context. This process is an integral part of research information and classification governance. These disciplines comprise the specification of decision rights and an accountability framework that encourages desirable behavior in the creation, storage, use, archival and deletion of research information or alternatively classification systems and the inclusion of processes, roles and standards that ensure the correct use thereof[7]. In this paper, the implementation of research information and classification governance in an inter-organizational context is described. Nevertheless, their inclusion is recommended even within a single organization in order to efficiently manage and process research information.

3. Research information and classification governance

Closed-world research information systems mostly comprise two layers, i.e. an information and data layer. The information layer typically comprises the information model and business rules accompanying research information, while the data layer entails the actual storage of the research information following storage schemes and data exchange formats. Furthermore, the data layer contains classification schemes and related code sets as well as validation rules. In essence, this 2-layered structure can be sufficient for a closed CRIS-system in case domain experts strictly adhere to the semantics of the standards used and the information retrievers perfectly understand these. However, the actual practice has demonstrated that this is mostly not the case. This situation even becomes more problematic when one wants to create an open-world research information system in an inter-organizational context as each institution mostly uses their own specific vocabularies. Therefore, if one wants to create a truly semantic interoperable CRIS, an additional data governance layer needs to be added on top of the information and data layer. A key component of this governance layer is the creation of a conceptual meta-model that can be aligned to the various organization-specific information models. This can only be realized by in-depth discussion amongst

business domain experts on the concepts and the definitions used in the conceptual meta-model to describe the various institution-specific business objects (Fig. 4). The naming convention of these concepts and their accompanying semantic definitions should be formalized in business terms and glossaries. Furthermore, business rules should be agreed on amongst the business domain experts that define the precise use and constraints of specific pieces of information. Finally, a data governance structure should be installed. This structure should define the roles and responsibilities for every single piece of information kept within the CRIS system.

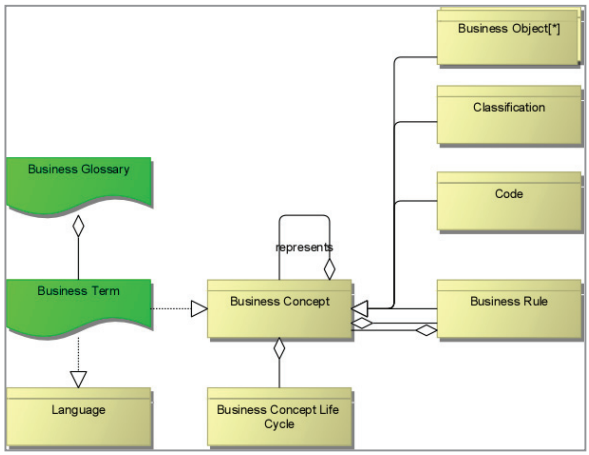

Figure 4: The alignment of a conceptual meta-model with institution-specific information models. Institution-specific information models in general include information on business objects (i.e. Project,…), their accompanying attributes (i.e. start date, end date,…) and related classification schemes and code sets. The conceptual meta-model arises when business domain experts discuss and agree on common business concepts, which for terms of ease are indicated with a naming convention (i.e. a business term) and explicitly defined semantics described in business glossaries.

4. Implementation of research information governance in the FRIS design

Although the use of data governance for describing research information tends to be quite easily accepted, hardly any CRIS system is currently implementing data governance thereby hampering true semantic interoperability. In Flanders, the inclusion of true semantic interoperability for research information systems is however one of the key elements in the reformation of the Flanders Research Information Space (FRIS)-portal. This portal , created in 2008 by the Department Economy, Science and Innovation, acts as an online research information space by making information on publicly funded research available to all stakeholders[8,9]. In 2014, the Flemish government decided to update the portal and its underlying design in order to get a more comprehensive and up-to-date dissemination of publicly funded research information from a wider variety of data providers for a broader range of use purposes. Obviously, this further stressed the need for true semantic interoperability, which prompted the Flemish government to adopt the design of its CRIS system from a 2-layered structure to a 3-tier architecture[10]. The topmost tier of this architecture comprises the data governance layer in which business experts govern the semantics of all concepts on research information and their corresponding classifications. The application layer involves the services and interfaces of the FRIS portal and the technological layer involves the actual data, the software and the application programming interface (API). By clearly separating these layers, yet maintaining formal relationships with the concepts within as well as across the layers, this allows for traceability across the layers when changes in one place occur. Most importantly, this ensures that business domain experts have an easier access to describe and manage their own research information without in-depth knowledge of the underlying information model.

However, the domain experts of the Flemish information providers use different terminologies for a similar concept or alternatively, use similar terminologies for a different concept within their, mostly autonomously developed information systems. Therefore, these institution-specific semantics cannot be simply aligned to the CERIF standard by information technologists. In order to reach true semantic interoperability, all business domain experts have to

agree upon a common ontology that renders the data provided meaningful. Yet, the domain experts consist of a large and heterogeneous group, ranging both in terms of the content owned as well as their affiliated institutions. In order to create semantic communities of topic-related business experts, a data governance tool named Collibra Data Governance Center (DGC) was used. DGC is an online software platform with a suite of data stewardship applications. The latter allows domain experts (in practice referred to as 'data stewards' to create meaning agreements on concepts in an easy accessible, collaborative and machine-readable manner[11].

Foundational to DGC is a fact-oriented operating framework and modelling methodology, based on Semantics of Business Vocabulary and Business Rules (SBVR) to capture concepts and their relationships in facts[5,12,13]. As such, all information providers can feed their own business models and classifications for describing research information together with the institution-specific semantics. Following the full cycle BSM method[5], the information suppliers are then virtually brought together in the data governance tool to discuss and agree on the conceptual FRIS metamodel together with the meaning of the concepts used. By running the process in iterative cycles, the method achieves to align business concepts, their semantics and corresponding representations between organizations and FRIS in a collaborative and dynamic manner (Fig. 4). Importantly, this model can be converted into a CERIF-based E-R model, that is implemented as optimized MySQL database tables and the accompanying ontology can be exported into an implementation in RDFS/OWL[2,8,9].

By introducing model driven formal semantic management at the business level, the FRIS 2.0 methodology tries to create a complete and managed modeling stack from business to the technical level. As the FRIS data governance platform is used to formally represent and relate all necessary representations of business concepts and the technical representation, this approach allows for a traceability across the complete FRIS 2.0 model stack, from the different business objects to their CERIF ER and CERIF XML representations and back. Importantly, this information can be extracted in a machine-readable manner.

5. Adoption of classification governance for research funding code schemes

Moreover, the FRIS 2.0 environment is even further strengthened as it includes research classification governance. As many information providers feed their information using or provided with research classifications, a perfect semantic understanding of these classifications is essential for meaningful information delivery and retrieval. Yet all too often, research classification governance is hardly considered in the construction of a CRIS system. This mostly comes from the fact that research classification are considered by many stakeholders as simple, static code systems used to classify research for reporting purposes. Obviously, these code systems should be used with full comprehension of the meaning of each code. However, in most cases research classifications lack semantic definitions and only contain codes and related terms. This directly opens the door to different interpretation possibilities at the side of the business domain experts, thereby causing dissimilar use on the intra- and sometimes even inter-organizational level. The problem becomes even more problematic when information technologists use such classified information to collect, feed and retrieve data as this leads to inconsistent research information within a CRIS system, its derived services and applications. In order to overcome this issue, the Flemish Government contracted the Centre for Research and Development Monitoring (ECOOM), which is a interuniversity consortium with participation of all Flemish universities that develops indicator systems for R&D and innovation monitoring. These ECOOM activities support the Flemish government in its ambition to perpetuate Flanders as a leading, innovative intensive region. In line with these objectives, ECOOM is also responsible for the development and dynamic management of classification systems for the monitoring of the Flanders' research project portfolio in terms of funding, publication and innovation output, and its assignment to scientific disciplines. In this paper, the adoption of classification governance for research funding code schemes is being exemplified.

At the start of the ECOOM-project, the FRIS 1.0 environment was already using a research funding code scheme, which will be further referred to as the Generation 1 FINcode scheme. This Generation 1 FINcode scheme, designed by the steering group of the Inventory of Scientific and Technological Research Flanders (IWETO), was characterized by a 3 levels deep hierarchical structure, where each level aggregated the funding origin to a different level of granularity. This hierarchy was also reflected in the corresponding code schemes, consisting of 4-digit code sets, thereby putting large constraints on the number of codes that could be used on the deepest level. For example, on the 3rd level, only 9 codes were left to denote all funding possibilities within this level. This forced the steering group to cluster different funding programs into a single code and to denote these with rather generic code terms (ex.

Code 3800 as representation of EU – out of framework – indefinite funding). Unfortunately, no explicit semantics was defined for each code, which resulted in different interpretations and thus usage of the codes in between research administrators belonging to different, or even the same university. Furthermore, the Generation 1 FINcode scheme was maintained decentral in spreadsheets at the side of the information providers, which over time even further nurtured the dissimilar use of the code sets by the information providers. Obviously, this put large constraints on the usefulness of the information on the FRIS 1.0 portal, represented by these codes.
In agreement with the decision of the Flemish government to revise the overall FRIS design to allow for more and better information dissemination, research classification governance was put in place to overcome these problems. A first task was the complete revision of the Generation 1.0 FIN code scheme design and its underlying funding model. The resulting Generation 2.0 FIN code scheme is characterized by the complete omission of hierarchy, which is also reflected in the auto numbering of the funding codes. The information residing in the former hierarchical levels is however recapitulated as a characteristic of a funding program, however much more characteristics have been added in order to facilitate research reporting and knowledge discovery. Furthermore, the funding codes are defined at a more granular level with an accompanying semantics, i.e. each code has a definition and descriptive examples in the DGC tool (Fig. 5).

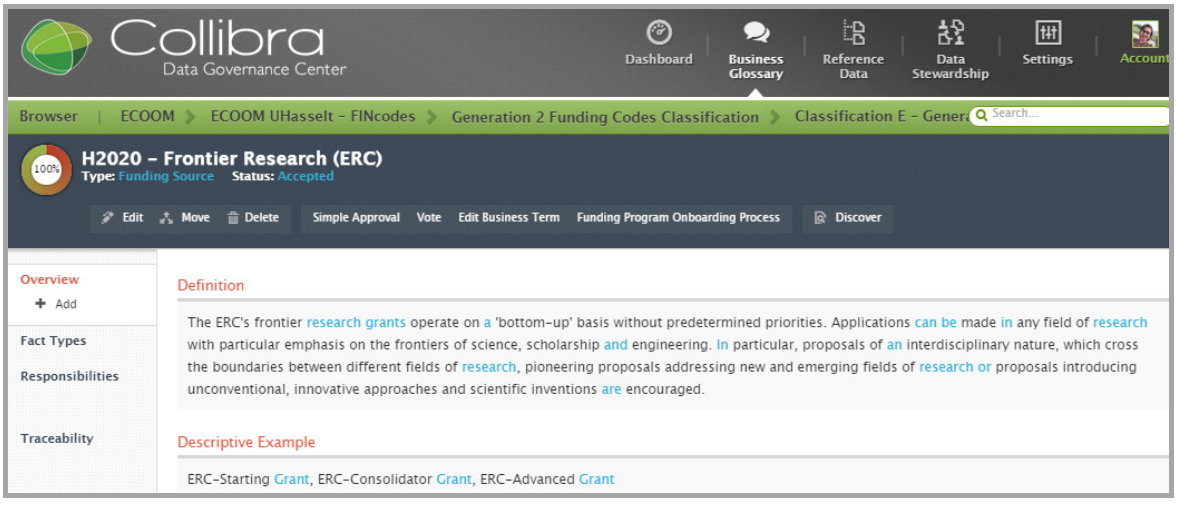

Figure 5: The Collibra® Data Governance Center as a governance tool for research funding classifications

The DGC tool not only allows for the central management of the code list as a single source of truth that can be consulted by all stakeholders, but also allows for a controlled and registered manner of applying for new funding codes or modifications of existing codes. To this end, workflows had to be designed and responsibilities had to assigned to each information provider, in order for joint decisions to be made by the research administrators (i.e. the business domain experts) that needed to be implemented by the information technologists of the different research universities. Furthermore, comments and replies are preserved by the tool, which ensures that stakeholders can have a perfect view on changes, the underlying reasons and the timing when these occurred. In addition, business rules are set that define the relations that should be used. Finally, the tool comprises a reference manager tool which can be used for the creation of concordance tables to existing (inter)national classifications. As all information residing in the tool is machine-readable, the information can be easily incorporated (and displayed) in the portal leading to a plethora of meaningful webservices that can be created around the data and thereby can lead on the long term even to the action of the FRIS 2.0 portal as a linked open data store[14].

6. Conclusion

Over the past decade, many organizations have started to adopt or build their own CRIS system using the CERIF standard. Although this standard is specifically designed for storage and exchange of research information, its inclusion does not fully guarantee the flawless exchange of semantically interoperable research information. This is due to the fact that research organizations do not strictly adhere to specific semantics used in the CERIF standard, while using the CERIF terminology. However, in order to get a true semantic interoperable CRIS system, especially in inter-organizational contexts, it is crucial to get a complete understanding of the information contained. This can only be accomplished by the inclusion of both research information and classification governance. Research information governance provides a structural framework that allows business domain experts to define a conceptual meta-model for research information and its alignment with institution-specific information models. Similarly classification governance, comprises the semantic delineation of research classifications and their accompanying concordance tables. As demonstrated in the paper, both disciplines are key assets to achieve true semantic interoperable research information systems that allow for the dynamic maintenance and retrieval of understandable, comparable research information for various purposes.

Acknowledgements

This work is part of the Classification Governance project carried out for the Expertise Centre for Research & Development Monitoring (ECOOM) in Flanders, which is supported by the Department of Economy, Science and Innovation, Flanders.

References

1. Peters, A., Lambrechts, L. De vereenvoudiging van onderzoeksverslaggeving, een analysetraject uitgevoerd door de Vlaamse universiteiten en hogescholen en de VLIR, in opdracht van de Vlaamse Overheid (EWI); 2011.
2. Debruyne, C., De Leenheer, P. Business Semantics as an interface between Enterprise Information Management and the Web of Data: a case study in the Flemish Public Administration. *In: eBiss;* 2012, **138:** 208-233.
3. EU Commission. CERIF: An EU recommendation to Member States, Commission Recommendation concerning the harmonisation within the Community of research and technological development databases. *In: Official Journal L;* 1991, **189:** 1-34, http://cordis.europa.eu/pub/cerif/docs/cerif1991.htm.
4. Jörg, B., Krast, O., Jeffery, K., Van Grootel, G. CERIF2008XML – 1.0 Data Exchange Format Specification, euroCRIS; 2009b.
5. De Leenheer, P., Christiaens, S., Meersman, R. Business semantics management with DOGMA-MESS: a case study for competency-centric HRM. *In: Computers In Industry*; 2010, **61(8):** 760-775.
6. De Leenheer, P., de Moor, A., Christiaens, S. Business semantics management at the Flemish Public administration; 2010.
7. Logan, D. What is information governance? And why is it so hard? *In: http://blogs.gartner.com/debra_logan/2010/01/11/what-is-information-governance-and-why-is-it-so-hard/*; 2010.
8. Van Grootel, G., Spyns, P., Christiaens, S., & Jörg, B. Business semantics management supports government innovation information portal. On the move to meaningful internet systems. *In: On the move to meaningful internet systems: OTM 2009 workshops. Lecture notes in computer science,* 2009, **5872**, 757–766.
9. Spyns, P., & Van Grootel, G. Realising the flanders research information space. *In On the move to meaningful internet systems: OTM 2011 workshops. Lecture notes in computer science;* 2011, **7046**, 138–141.
10. Vancauwenbergh, S., De Leenheer, P., Van Grootel, G. On research information and classification governance in an inter-organizational context: the Flanders Research Information Space. *In: Scientometrics, Special Issue on Grand Challenges in Data Integration for Research and Innovation Policy;* 2016, **106(3)**, DOI 10.1007/s11192-016-1912-7.
11. Plotkin, D. Data stewardship: An actionable guide to effective data management and data governance. 2014, ISBN:978-0-12-410389-4.
12. Verheijen, G., & Van Bekkum, J. NIAM, an information analysis method. *In: Proceedings of the IFIP TC-8 conference on comparative review of information system methodologies (CRIS 82);* 1982.
13. OMG SBVR In http://www.omg.org/spec/SBVR/
14. Dimou, A., De Vocht, L., Van Grootel, G., et al. Visualizing the information of a linked open data enabled research information system. *In: Procedia Computer Science*; 2014, **33**, 245–252.