Business Semantics Management Supports Government Innovation Information Portal

Geert Van Grootel[1], Peter Spyns[1,2], Stijn Christiaens[2,3], and Brigitte Jörg[4]

[1] Vlaamse overheid – Departement Economie, Wetenschap en Innovatie
Koning Albert II-laan 35, bus 10, B-1030 Brussel, Belgium
Geert.VanGrootel@ewi.vlaanderen.be
[2] Vrije Universiteit Brussel – STAR Lab
Pleinlaan 2, Gebouw G-10, B-1050 Brussel, Belgium
Peter.Spyns@vub.ac.be
[3] Collibra NV/SA
Brussel Business Base, Ransbeekstraat 230, B-1120 Brussel, Belgium
stijn@collibra.com
[4] Deutsches Forschungszentrum für Künstliche Intelligenz – Bereich Sprachtechnologie
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
brigitte.joerg@dfki.de

Abstract. The knowledge economy is one of the cornerstones of our society. Our economic prosperity and development is derived for a large part from technical knowledge. Knowledge unlocks innovation, which in turns spawns new products or services, thereby enabling further economic growth. Hence, an information system unlocking scientific technical knowledge is an important asset for government policy and strategic decisions by industry. In this paper it is explained how business semantics management and related tools are applied to realise the above mentioned endeavour.

1 Background

Needing a new system for managing research information, the Flemish department of Economy, Science and Innovation (EWI) launched the Flanders Research Information Space programme (FRIS)[1]. The term is used to refer both to the virtual environment of research information and the program that is being setup in order to create this research information space. The FRIS concept creates a virtual research information space covering Flemish players in the field of economy, science and innovation (see Fig. 1) – see e.g. [16] for some other use cases of a prototype in a related area. Within the space, research information can be stored and exchanged in a transparent and automated way. The FRIS program is centred on three strategic goals:

- accelerate the innovation value chain by efficient and fast access to research information for all relevant stakeholders;
- offer improved customer services (e-government);
- increase efficiency and effectiveness of the R&D policy.

[1] Flanders Research Information Space (FRIS): www.ewi-vlaanderen.be/fris

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 757–766, 2009.

The strategic goals will be achieved by a combination of service development and a managed change process. A first realisation is the FRIS research portal (www.researchportal.be) to exhibits current research information on projects, researchers and organisations of the Flemish universities.

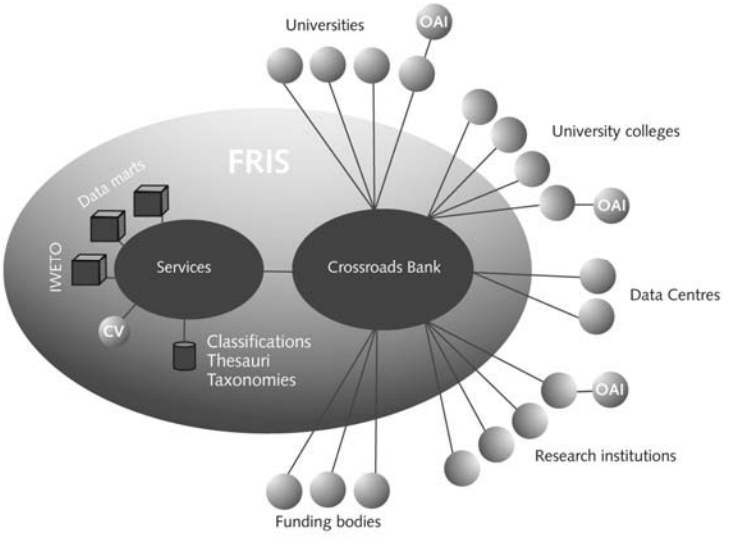

Fig. 1. The representation of the FRIS with data providers (universities, data centres supporting the Open Archives Initiative[2] …), utilities (classifications, …) and services (maintenance of scientific curriculum vitas (CVs), research portal called IWETO, …)

A key feature is that data can be immediately collected at the point of creation in the operational processes of data providers (e.g., universities, funding bodies …). E.g., the first hand information on a research project is already made available during the assessment process of an application for funding. Entering this information is part of the process – at the point of creation – and will be done by the applying organisation itself. The data are up-to-date and are supposedly more accurate (than second hand information entered by non or less related people elsewhere). Also, parallel data gathering processes (e.g., research administrators gathering data at regular intervals, but not related to assessment processes) can be eliminated, resulting in a lot of administrative work being spared.

2 Material

CERIF [7] is the standard of choice for the information interchange between the system that will be part of FRIS. CERIF that stands for Common European Research Information Format is a recommendation to EU member states for the storage and

[2] Open Archives Intitiative (OAI): http://www.openarchives.org/

exchange of current research information [8,9,10]. The custodianship of the CERIF standard has been transferred to euroCRIS[3], a non for profit organisation that maintains and develops CERIF and promotes the use of it for research information exchange.

CERIF is centred on a limited set of basic concepts:

- the business objects (entities) of the research and innovation domain: project, person, organisational unit, publication, event,…. – see Fig. 2;
- the relations through time that exist between these objects;
- support for multilingual text attributes;
- a separated semantics storage layer.

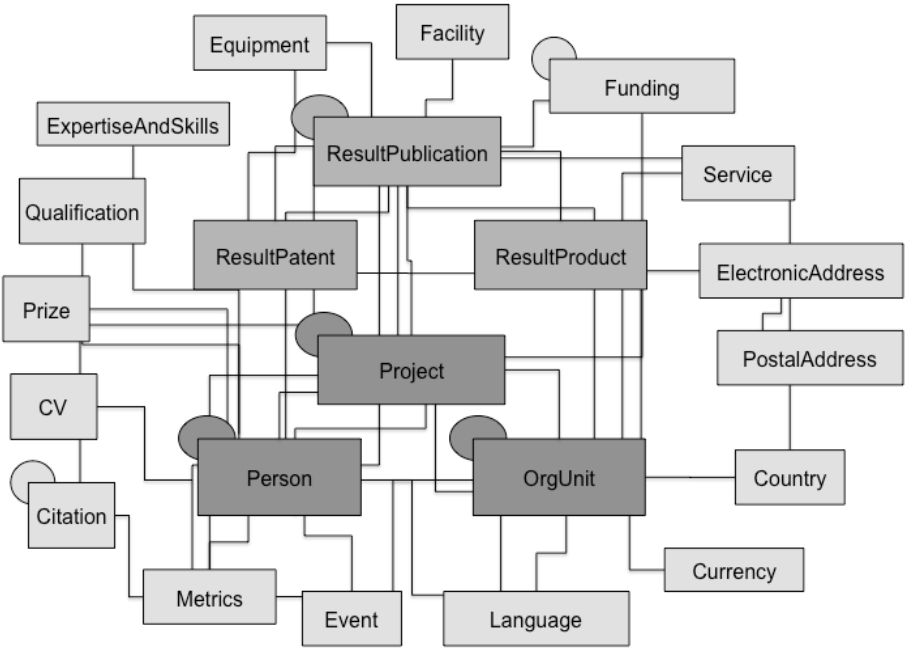

Fig. 2. CERIF entities and their relationships in abstract view, with Core entities (Project, Person, OrgUnit) in the lower centre, Result entities (Publication, Patent, Product) in the upper centre, and 2nd Level entities presented as a contextual environment around the Core and Result entities. The little circles with entities represent their recursive relationships.

In the next phase of the FRIS research portal we will realise the integration of the portal data with the publication repositories of the data suppliers. CERIF as the exchange format and as the basic mode for the database interface allows for searching in and browsing through the data including the formalised semantic links between the database objects. More concretely, this means information about a researcher with his research context: project, organisations, publications can be linked according to the

[3] euroCRIS home: http://www.eurocris.org/public/home/

formal CERIF semantics, and thus be presented in different ways: simple lists, collaboration graphs, competences maps[4], etc.

3 Work in progress

3.1 Why Business Semantics Management?

Traditionally CERIF has been modelled using Entity-Relationship (E-R) modelling techniques. While this technique is excellent for modelling activities of database designs, it is less efficient for communication with domain experts. The learning curve for the domain experts to understand the E-R model and translate it back to their conceptual knowledge is quite steep. Probably this difficulty constitutes one of the main reasons why the acceptance of CERIF is only slowly growing.

Also domain experts at EWI have been struggling for quite some time with the problem of how to express and explain adequately and flexibly conceptual models underlying the FRIS application to the other stakeholders involved (mainly non technical persons). Domain experts and stakeholders should not be bothered with how to think in a (new) formal language: adequately capturing and organising domain knowledge is a task sufficiently demanding as it is and mainly happens via the use of natural language.

Business Semantics Management (BSM) is the set of activities to bring stakeholders together to collaboratively realise the reconciliation of their heterogeneous metadata and consequently the application of the derived business semantics patterns to establish semantic alignment between the underlying data structures [2]. BSM relies heavily on fact oriented modelling, the basis of which has been described in [11] (ORM) and [20] (NIAM), and makes to a large extent use of natural language. ORM focuses more on verbalising the conceptual model for an easy validation process with domain experts, while NIAM puts more emphasis on using natural language as starting point and elicitation vehicle to build the conceptual model. Also, fact oriented modelling is much more flexible and sustainable than attribute-value modelling. BSM recognizes the need for different roles to distribute complexity and improve the work process. Domain experts work on semantic reconciliation and produce semantic patterns, as close to their natural language as possible, while application developers use these patterns as input for semantic application (see Fig. 4 below): to semantically align (or commit) their systems while providing feedback to the domain experts. Collibra's Information Enabler, a runtime semantics engine, can then apply these semantics by translating from existing legacy systems (e.g. the OAI Protocol for Metadata Harvesting[5]) to CERIF.

3.2 Some Difficulties with the Current E-R Representation of the CERIF Model

The latest CERIF 2008 – 1.0 release newly introduced the Semantic Layer for capturing formal contextual roles and types; by that, the semantics in the relationships between the CERIF entities is now being maintained flexibly and independently within

[4] As realized with: http://www.ist-world.org/
[5] http://www.openarchives.org/pmh/

the Semantic Layer, in natural language, based on the CERIF concepts, architecture, structure and style [10]. However, the fact that roles of the relations between entities and classifications for entities are not explicit in the E-R model, makes it non trivial for humans to read and understand the model. E.g., think of a Project (with a known identifier, associated classifications and a set of basic attributes) and a Person (also known by an identifier and a set of attributes) in the role of promoter with respect to a project. To express this relationship in CERIF the entities used are shown in Fig. 3.

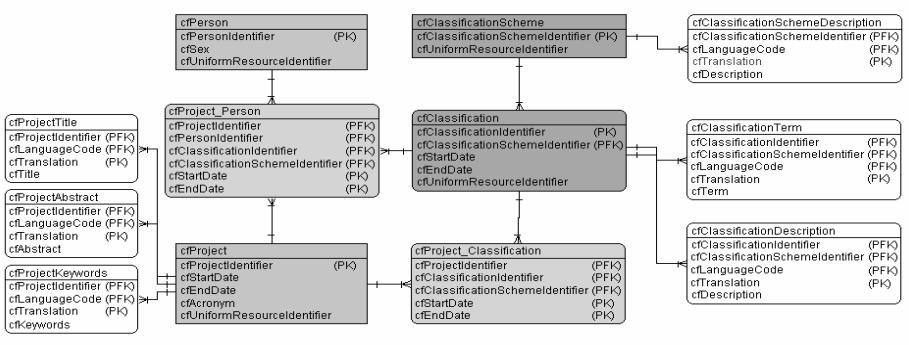

Fig. 3. Excerpt of some E-R model entities of Fig. 2 in physical view, with the CERIF Core entities cfPerson and cfProject, language-related entities cfProjectTitle, cfProjectAbstract, cfClassTerm, …, selected link entities cfProject_Person and cfProject_Classification, and entities from the Semantic Layer cfClassification and cfClassificationScheme

In CERIF, a time-constrained relationship, with a defined role is established in lin entities, i.e., the cfProject_Person link entity (see Fig. 3), or the classification (i.e. typification) of projects in the cfProject_Classification entitity via referenced IDs. The upper right set of entities represents the strength of the CERIF E-R model: the semantic layer. It is designed to store all possible roles of relations and all possible types or classifications associated with entities that have been established by IDs in link entities. In the example case the Classification Scheme used for cfProject_Person roles is called cfProjectPersonRoles and the cfClassification entity is the storage for the "P" value, while the language dependent associated entity cfClassificationTerm stores the human readable version of the value in a specified natural language: "promotor" in Dutch[6].

The association of specific classifications with instances uses a similar connection within the semantic layer, in this example via the entity cfProject_Classification. This allows for the association of multiple classifications belonging to different Classification Schemes with a project. For instance one scheme used for project types and two others for association with two different thesauri; disciplines or application domains.

Each of the business objects (CERIF entities) has a limited set of attributes, apart from the basic object identifier. Relationships between objects are achieved by means of linking entities consisting of (i) the identifiers of the objects taking part in the relation, (ii) a start and (iii) end date and (iv) a defined role of the relation (by ID reference to the

[6] Some entities are language dependent and allow for the storage of different language versions.

Semantic Layer). In essence, every instance of a linking entity is a time constrained binary relation. The instances of objects taking part in a linking relation to the same object (i.e. cfPerson_Person) represent recursive relationships. The roles of the relations are referred from the semantic layer of CERIF. This semantic layer physically consists of a set of interrelated entities enabling the storage of schemes and the instances representing all possible roles from all possible relations. The semantic layer is also used for the storage of classifications (i.e. types) associated with the objects (entities).

Much information in the research domain needs a representation in multiple languages. The support of multilingual features is very important in countries where several official languages are spoken and maintained. CERIF supports multiple language features for names, titles, descriptions, keywords, abstract and even for the semantics.

For a domain expert to understand how a set of business facts is expressed by using CERIF – including the content of the semantic layer both the insight in the CERIF E-R model and knowledge of the content (contextual semantics) is needed.

3.3 Business Semantics Management and CERIF

While the CERIF model allows for an almost unlimited flexibility on roles and classifications used with entities, the actual approach has shown its limitations when it comes to communicating on the modelled domain facts to domain experts and end users. In order to overcome these difficulties, EWI has decided to express the business facts in the domain concerned by the use of fact based modelling, thereby using the expertise and tools[7] of Collibra. Fig. 5. shows some binary fact types that can be derived from the original CERIF model (see Fig. 2). One can appreciate the clarity and understandability of the binary fact types compared to the E-R scheme (see Fig. 3).

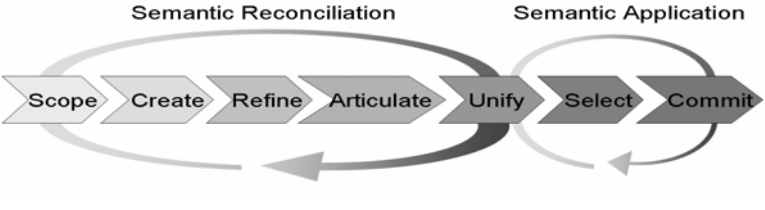

Fig. 4. Business Semantics Management (BSM) overview

Currently, the focus of applying BSM on CERIF has been semantic reconciliation: domain experts that collaboratively capture business semantics. This phase consists of five activities (see Fig. 4), which we briefly describe here[8] in the context of the work on CERIF. Note that some activities can be performed simultaneously:

1. Scope: defining the borders of the current iteration of BSM. For CERIF, which is already a well worked out initiative, we have used the core entities (see Fig. 2) as starting point for different iterations. Scoping helps in grounding discussions: one

[7] Available from http://www.collibra.com/trial
[8] For a more elaborate description on BSM, see [2].

can always refer back to the source documents (or other boundaries) to bring a difficult (and sometimes philosophizing) discussion back on track.

2. Create: generate fact types (lexons[9]) from the collected sources in the scoping activity. The focus lies on getting all the facts without leaving much room for discussion (divergence).
3. Refine: clean the collected facts by following some simple rules (a first step towards convergence). For instance, decide where typing is relevant by determining which concepts share differentia (e.g., the concept "Text" in Fig. 5). Another example is splitting a semantic pattern in smaller, more reusable patterns each describing some part of the model (e.g., Person pattern, Address pattern).
4. Articulate: create informal meaning descriptions as extra documentation. These descriptions can serve as anchoring points when stakeholders have used different terms for the same concepts (i.e., detecting synonyms). Where available, descriptions already existing can be used (e.g., the euroCRIS website on CERIF) to speed up the process and facilitate reuse.
5. Unify: collect the input from the various stakeholders and combine them in agreed upon and shared semantic patterns. This is an activity that leaves room for discussion when necessary. Any unresolved issues can be tackled here, based on the deliverables produced in the previous activities. In the EWI case, the activity was performed more or less simultaneously with the other activities as the relevant domain experts were participating.

The Collibra Studio supports many of the activities described above, thereby further assisting the domain experts in capturing their business semantics: (i) a fact editor allowing the domain expert to simply key in the facts in natural language; (ii) a visual editor (based on [19]) providing a graphical way of presenting and browsing through the collection of lexons (see Fig. 5); and (iii) a concept editor with a built-in browser for searching the web for already existing informal meaning descriptions (e.g., on the euroCRIS website, on Wikipedia, …).

3.4 Ontologising the CERIF Model

Having a standard conceptual schema, c.q. the CERIF model, (see Fig. 3) is not enough to guarantee flawless interoperability and information exchange. Intelligent agents have to be able to exchange "meaningful" messages while continuing to function autonomously (interoperability with local autonomy as opposed to integration with central control). Hence, there should not be any confusion about which data are labelled as belonging to which category or being associated with which label. Even in the ontology literature, authors do not always make a clear distinction between a global domain concept and a local conceptual schema term [17].

Next to the schema itself, agreement (or least univocity) should be achieved on the terminology of the schema: what is exactly meant by the names of the entities and relationships. Most importantly, interoperability requires a precise and formal definition

[9] Informally a lexon is a fact type that may hold for some domain, expressing that within a context the head term may plausibly have the tail term occur in role with it (and inversely the tail term maintains a co-role relation with the head term) [17]. Lexons operate on the linguistic level. We refer to [1] for a formalisation of DOGMA, including the notion of lexons.

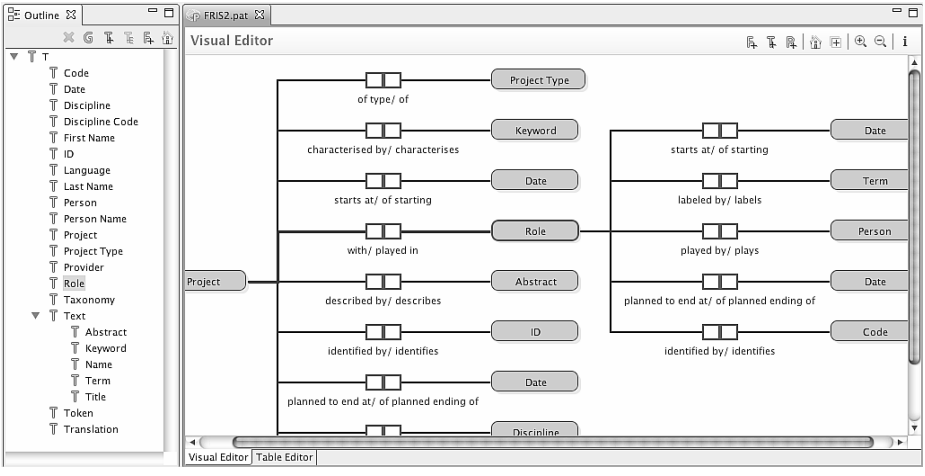

Fig. 5. The Collibra tool used during the modelling phase (semantic constraints yet to be added)

of the intended semantics of an ontology (see also [12,14] for more and other details in this discussion). Not only does this presuppose a definition in natural language, also a formalisation is needed to impose additional restrictions on the relationships between the entities. This separation corresponds to the DOGMA double articulation principle [17] for ontology modelling: the distinction between the domain conceptualisation (generic facts) and the application conceptualisation (application constraints) [18]. Verbalising the entire ontology (including the restrictions) ensures a good understanding by the domain experts and knowledge engineers of the various data providers involved. In particular, when multilinguality is required, as is here the case, it is extremely important to distinguish between a level of linguistic utterances and a language neutral conceptual definition of the labels and terms [16].

4 Discussion

An open question remains what to do with “complex concepts” (e.g., “cfProject_Person”). The question of naming conventions for complex concepts arises from the assumption that every concept refers to a piece of reality. Sometimes the meaning is felt to be too broad and some specialisation (expressed in natural language by a compound “project classification” or a circumscription “person who works on a project”) is wanted. Currently, we tend to reject “complex concepts”, albeit it more on philosophical grounds (“notions are not to be multiplied without necessity” = Occam’s razor). Practice should show if sufficient necessity is available.

5 Future Work

Over the coming months and years, the FRIS research portal will roll out a range of new services as part of the research information space. The possibilities are numerous: a white guide (who does what?), library of publications by a particular researcher

(digital library), a service for updating and reusing researchers' CVs and provision of information on patents, to name but a few – see Fig. 1.

In the future, more effort is to be spent on applying insights from formal ontologies to avoid modelling errors. Methods like OntoClean [4] or a more formal foundation of some often occurring relationships ([6] might be a good starting point) should be applied as a first "sanity check" on newly created ontologies.

6 Conclusion

The CERIF standard is, by its nature and status of standard, an ideal candidate to be "ontologised". For reasons of data quality and integrity as few ambiguities as possible concerning the meaning and use of domain terminology should occur. This requirement holds in particular in a context of a networked configuration, consisting of a limited number of data providers, a central portal and an unlimited number of potential data users (human and/or artificial) of various types of organisations (see Fig. 1). An ontology is exactly meant for this purpose.

The case presented in this paper, namely how to build a generic model for research information exchange in Flanders, shows that business semantics management are sufficiently easy to apply by non technical stakeholders and domain experts. In addition, it also provides an interesting case that illustrates how fact oriented modelling can be used as a robust basis for ontology modelling – an exercise already supported by appropriate tools by Collibra turning academic insights (mainly the DOGMA modelling methodology [1,2,11,13,14,18,19]) into industrial practice. In short, it is an ambitious endeavour and, for the Flemish government, it is an innovating and fresh ['fris' = 'fresh' in Dutch] approach.

References

1. De Leenheer, P., de Moor, A., Meersman, R.: Context Dependency Management in Ontology Engineering: a Formal Approach. In: Spaccapietra, S., Atzeni, P., Fages, F., Hacid, M.-S., Kifer, M., Mylopoulos, J., Pernici, B., Shvaiko, P., Trujillo, J., Zaihrayeu, I. (eds.) Journal on Data Semantics VIII. LNCS, vol. 4380, pp. 26–56. Springer, Heidelberg (2007)
2. De Leenheer, P., Christiaens, S., Meersman, R.: Business Semantics Management with DOGMA-MESS: a Case Study for Competency-centric HRM. Journal of Computers in Industry: Special Issue about Semantic Web Computing in Industry (in print, 2009)
3. Guarino, N.: Formal Ontologies and Information Systems. In: Guarino, N. (ed.) Proc. of FOIS 1998, pp. 3–15. IOS Press, Amsterdam (1998)
4. Guarino, N., Welty, C.: An Overview of OntoClean. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, International Handbook on Information Systems, pp. 151–172. Springer, Heidelberg (2004)
5. Guarino, N., Oberle, D., Staab, S.: An Introduction to Ontologies. In: Studer, R., Staab, S. (eds.) Handbook of Ontologies, International Handbooks on International Infomation Systems, 2nd edn. Springer, Heidelberg (fothcoming)
6. Guizzardi, G., Wagner, G., Herre, H.: On the Foudations of UML as an Ontology Representation Language. In: Motta, E., Shadbolt, N.R., Stutt, A., Gibbins, N. (eds.) EKAW 2004. LNCS (LNAI), vol. 3257, pp. 47–62. Springer, Heidelberg (2004)

7. Jörg, B.: The Common European Research Information Format Model (CERIF). In: Asserson, A. (ed.) CRISs for the European e-Infrastructure. Data Science Journal (in Print, 2009)
8. Jörg, B., Jeffery, K., Asserson, A., van Grootel, G.: CERIF2008 - 1.0 Full Data Model (FDM) - Model Introduction and Specification. euroCRIS (2009)
9. Jörg, B., Krast, O., Jeffery, K., van Grootel, G.: CERIF2008XML - 1.0 Data Exchange Format Specification, euroCRIS (2009)
10. Jörg, B., Jeffery, K., Asserson, A., van Grootel, G., Rasmussen, H., Price, A., Vestam, T., Karstensen Elbæk, M., Houssos, N., Voigt, R., Simons, E.J.: CERIF2008-1.0 Semantics, euroCRIS (2009)
11. Halpin, T.: Information Modeling and Relational Databases: from conceptual analysis to logical design. Morgan Kaufmann, San Francisco (2001)
12. Meersman, R.: The Use of Lexicons and Other Computer-Linguistic Tools. In: Zhang, Y., Rusinkiewicz, M., Kambayashi, Y. (eds.) Semantics, Design and Cooperation of Database Systems; The International Symposium on Cooperative Database Systems for Advanced Applications (CODAS 1999), pp. 1–14. Springer, Heidelberg (1999)
13. Meersman, R.: Ontologies and Databases: More than a Fleeting Resemblance. In: d'Atri, A., Missikoff, M. (eds.) OES/SEO 2001 Rome Workshop. Luiss Publications (2001)
14. Spyns, P., Meersman, R., Jarrar, M.: Data modelling versus Ontology engineering. In: Sheth, A., Meersman, R. (eds.) SIGMOD Record Special Issue, vol. 31 (4), pp. 12–17 (2002)
15. Spyns, P., Van Acker, S., Wynants, M., Jarrar, M., Lisovoy, A.: Using a novel ORM-based ontology modelling method to build an experimental Innovation router. In: Motta, E., Shadbolt, N.R., Stutt, A., Gibbins, N. (eds.) EKAW 2004. LNCS (LNAI), vol. 3257, pp. 82–98. Springer, Heidelberg (2004)
16. Spyns, P., De Bo, J.: Ontologies: a revamped cross-disciplinary buzzword or a truly promising interdisciplinary research topic? Linguistica Antverpiensia NS(3), 279–292 (2004)
17. Spyns, P.: Object Role Modelling for Ontology Engineering in the DOGMA framework. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM-WS 2005. LNCS, vol. 3762, pp. 710–719. Springer, Heidelberg (2005)
18. Spyns, P., Tang, Y., Meersman, R.: An ontology engineering methodology for DOGMA. Journal of Applied Ontology 1-2(3), 13–39 (2008)
19. Trog, D., Vereecken, J., Christiaens, S., De Leenheer, P., Meersman, R.: T-Lex: a Role-based Ontology Engineering Tool. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4278, pp. 1191–1200. Springer, Heidelberg (2006)
20. 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 1982). North-Holland, Amsterdam (1982)