Publishing Open Data and Services for the Flemish Research Information Space

Christophe Debruyne[1], Pieter De Leenheer[2,3], Peter Spyns[4], Geert van Grootel[4], and Stijn Christiaens[3]

[1] STARLab, Vrije Universiteit Brussel, Brussels, Belgium
[2] Business Web & Media, VU University Amsterdam, Amsterdam, The Netherlands
[3] Collibra nv/sa, Brussels, Belgium
[4] Flemish Dept. of Economy, Science, and Innovation, Brussels, Belgium

Abstract. The Flemish public administration aims to integrate and publish all research information on a portal. Information is currently stored according to the CERIF standard modeled in (E)ER and aimed at extensibility. Solutions exist to easily publish data from databases in RDF, but ontologies need to be constructed to render those meaningful. In order to publish their data, the public administration and other stakeholders first need to agree on a shared understanding of what exactly is captured and stored in that format. In this paper, we show how the use of the Business Semantics Management method and tool contributed in achieving that aim.

Keywords: ontology development, methodology, social process, business semantics management, fact-orientation, natural language.

1 Introduction: An Innovation Information Portal for Flanders

For a country or region in the current knowledge economy, it is crucial to have a good overview of its science and technology base to develop an appropriate policy mix of measures to support and stimulate research and innovation. Also companies, research institutions and individual researchers can profit from the information maintained in such a portal. EWI[1] thus decided to launch the Flanders Research Information Space program (FRIS) to create a virtual research information space covering all Flemish players in the field of economy, science and innovation. The current version of this portal[2] contains, for instance, mash-ups of data on key entities (such as person, organization, and project; and their relationships) on a geographical map. Another aim of FRIS is to reduce the current administrative burden for universities as they are confronted with repeatedly reporting the same information in different formats to various institutions. Indeed, if all information would be centralized and accessible in a uniform way, creating services for such reports would greatly facilitate the reporting process. Before data can be centralized, this initiative faces two problems: 1) capturing the semantics of the domain in an ontology and 2) appropriately annotate or commit the heterogeneous data sources to that ontology.

[1] The Department of Economy, Science and Innovation of the Flemish Government `http://www.ewi-vlaanderen.be/`

[2] `http://www.researchportal.be/`

O. De Troyer et al. (Eds.): ER 2011 Workshops, LNCS 6999, pp. 389–394, 2011.

As we will explain in Section 2, integrating all information and reducing the administrative burden faces some problems for which appropriate data governance methods and tools are needed. Such method and tool is presented in Section 3 and we end this paper with a conclusion in Section 4.

2 Problem: Heterogeneous Information Sources

Universities receiving funding from the Flemish government are asked to regularly report the same information to different organizations (local and international). As there is little alignment between those reports, universities are confronted with repeatedly sending the same information in other formats, other structures or according to different classifications not always compatible with each other[3]. This creates a heavy administrative burden on those knowledge institutions. Universities furthermore store their information in *autonomously developed* information systems, adding to the complexity of the problem.

As the EU also wants to track all research information in Europe, they ask all universities to report using the Common European Research Information Format (CERIF) [4], a recommendation to EU-members for the storage and exchange of current research information. While the CERIF model, created with Entity-Relationship (ER) diagrams, 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* the modeled domain facts to domain experts and end users. The learning curve for the domain experts to understand the ER model and translate it back to the conceptual level is quite steep. **Fig. 1** shows some CERIF entities, their attributes and relationships.

To populate the FRIS portal with all information provided by the delivered CERIF files and other heterogeneous sources, (i) a consensus amongst the involved parties on a common conceptual model for CERIF and the different classifications is needed (taking into account the non-technical expertise of most domain experts), (ii) an easy, repeatable process for validating and integrating the data form those sources and finally (iii) using that shared understanding to publish that information as in a generic way on the Web on which third parties can develop services (commercial or not, e.g. to produce the different reports) as demonstrated by other Linked Data initiatives.

3 Approach: Business Semantics Management

In order to overcome the above-mentioned difficulties, all these classifications and models need to be actualized and homogenized on a conceptual level, first within Flanders, later with more general and international classifications. The Business Semantics Management (BSM) [2] methodology was adopted to capture the domain knowledge inside CERIF and the different classifications. BSM adopts the Semantics of Business Vocabulary and Business Rules (SBVR) [1] to capture concepts and their relationships in facts. SBVR is a *fact-oriented* modeling approach. Fact-oriented modeling is a method for analyzing and creating conceptual schemas for information

[3] Different classifications are used within Flanders: IWETO discipline codes, IWETO science domains, VLIR scientific disciplines, IWETO application domains, SOOI (based on the IWI-Web of Science codes), NABS (used for budgeting) and FOS (Fields of Science), etc.

systems starting from (usually binary) relationships expressed as part of human-to-system communication. Using concepts and a language people are intended to readily understand, fact-oriented modeling helps ensuring the quality of a database application without caring about any implementation details of the database, including e.g. the grouping itself of linguistic concepts into records, relations, … In fact-oriented approaches, every concept plays roles with other concepts, and those roles may be constrained. It is those constraints that allow the implementer of a database (or in fact an algorithm) to determine whether some linguistic concept becomes an entity or an attribute, or whether a role turns out to be an attribute relationship or not. This is different from other approaches such as (E)ER and UML, where these decisions are made at design time.

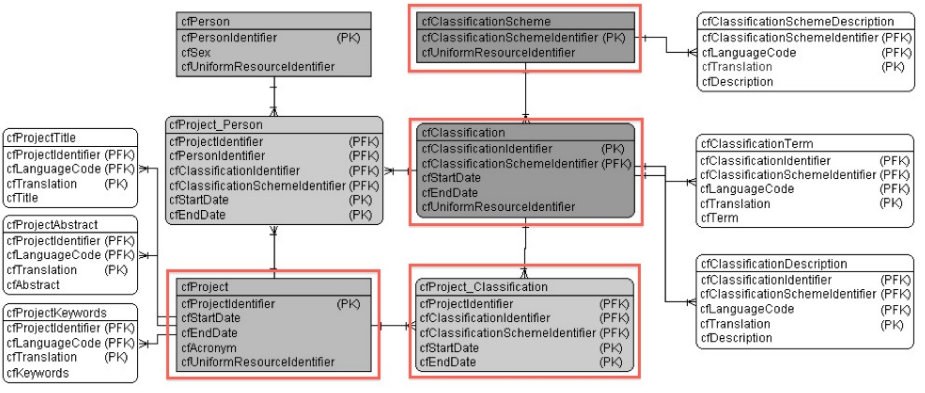

Fig. 1. The CERIF entity cfProject and its relationship with the entity cfProject_Classification (linked by the two identifiers of the linked entities). A CERIF relationship is always semantically enriched by a time-stamped classification reference. The classification record is maintained in a separate entity (cfClassification) and allows for multilingual features. Additionally, each classification record or instance requires an assignment to a classification scheme (cfClassificationSchemeIdentifier).

Business semantics management is the set of activities (depicted in **Fig. 2**) to bring business stakeholders together to collaboratively realize 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.

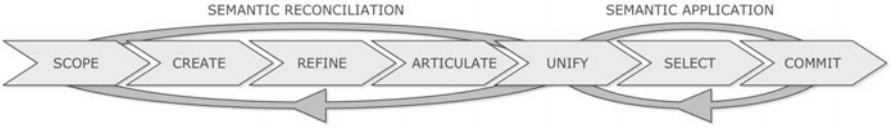

Fig. 2. Business Semantics Management overview: semantic reconciliation and application

The first cycle, *semantic reconciliation*, is supported by the Business Semantics Glossary (BSG) shown in **Fig. 3**. This figure shows a screenshot of the term “Project” (within the “Project” vocabulary of “CERIF” speech community that is part of the

"FRIS" semantic community). The software is currently deployed at EWI for managing business semantics of CERIF terms. A term (here "Project") can be defined using one or more attributes such as definitions, examples, fact types, rules sets, categorization schemas (partly shown in taxonomy), and finally milestones for the lifecycle. "Project" in this case is a subtype of "Thing" and has two subtypes: "large academic project" and "small industrial project". Re governance: in the top-right corner is indicated which member in the community (here "Pieter De Leenheer") carries the role of "steward", who is ultimately accountable for this term. The status "candidate" indicates that the term is not yet fully articulated: in this case "Project" only 37.5%. This percentage is automatically calculated based on the articulation tasks that have to be performed according to the business semantics management methodology. Tasks are related to defining attributes and are distributed among stakeholders and orchestrated using workflows.

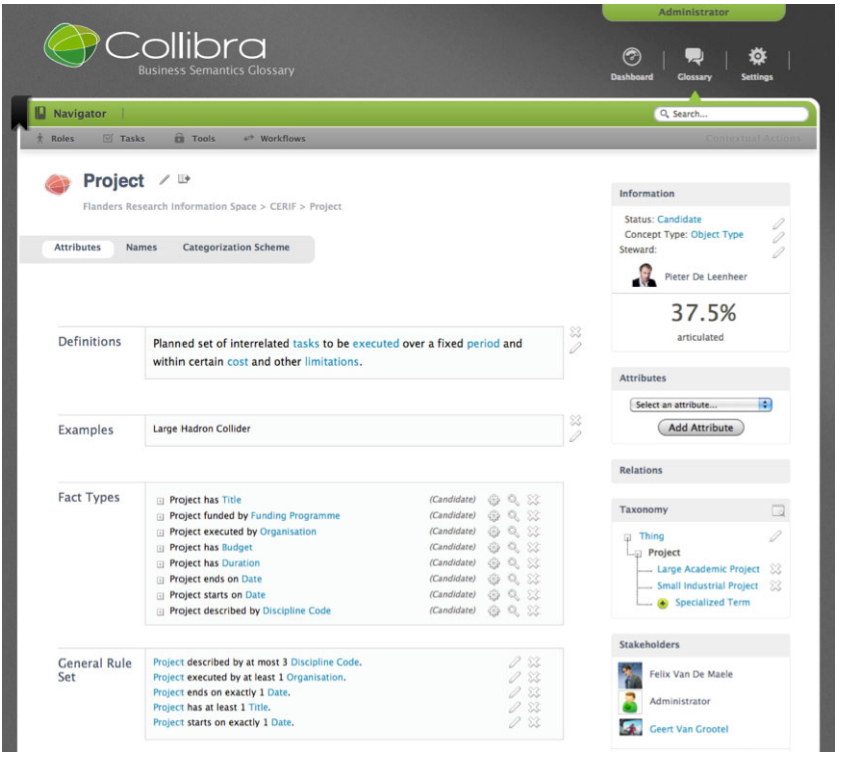

Fig. 3. Screenshot of Collibra's BSG supporting the semantic reconciliation process of the BSM methodology by providing domain experts means to enter simple key facts in natural language, natural language definitions on facts and terms in those facts as well as constraints.

Applying BSM results in a community driven (e.g. representing the different classifications and models mentioned earlier), iteratively developed shared and agreed upon conceptual model in SVBR. This model then is automatically converted in a CERIF-based ER model and RDFS/OWL for Web publishing. **Fig. 4** shows a part of the generated OWL from the concept depicted in the previous figure. In this figure,

we see that `Project` is a `Class` and all instances of that class are also instances of entities with at least one value for the property `ProjectHasTitle`, one of the rules expressed in SBVR in **Fig. 3** (see general rule sets).

Fig. 4. Screenshot of the OWL around Project generated by BSG. In this picture, we see that `Project` is a `Class` and all instances of that class are also instances of entities with at least one value for the property `ProjectHasTitle`.

The contents of the databases to be annotated can be published with off-the-shelve solutions such as D2R Server[4]. D2R Server generates an RDF a mapping for transforming the content of a database into RDF triples. This mapping – also described in RDF – contains a "skeleton" RDFS of classes and properties that are based on the database schema. **Fig. 5** below depicts a part of the generated mapping file around the table containing information around projects.

```
@prefix map: <file:///.../OSCB/d2r-server-0.7/map.n3#> .
@prefix vocab: <http://192.168.0.136:5432/vocab/resource/> .
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
...
map:CFPROJ a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:uriPattern "CFPROJ/@@CFPROJ.CFPROJID|urlencode@@";
        d2rq:class vocab:CFPROJ;
        d2rq:classDefinitionLabel "EWI.CFPROJ";
...
```

Fig. 5. Part of the generated mapping file by D2R server, it maps the table `CFProj` to the generated `CFPROJ` RDFS class. It uses the primary key to generate a unique ID and the class definition label is taken from the table's name.

Even though classes and properties are generated and populated with instances, these RDF triples are not semantic as they stem from one particular information system('s database schema). That RDFS is then aligned with the generated RDFS/OWL classes and properties generated from the BSM ontology. The commitments described in the previous section are used as a guideline to create this alignment. **Fig. 6** below shows the changes (highlighted) made on the generated mapping file with the ontology. The ontology can then be used to access the data.

[4] `http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/`

```
@prefix ont: <file:///.../Project.rdf#> .
...
map:CFPROJ a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:uriPattern "CFPROJ/@@CFPROJ.CFPROJID|urlencode@@";
        d2rq:class ont:Project;
        d2rq:classDefinitionLabel "Project";
...
```

Fig. 6. Modified mapping file with the ontology exported from BSG. An extra namespace (for the exported ontology) is added and the generated classes and properties are appropriately annotated with that ontology.

4 Conclusion: Inclusion of the Method and Tool in the Portal's Architecture

This paper presented a case of applying Business Semantics Management (BSM) in a Flemish public administration for the creation of an innovation information portal. The purpose of this portal is to integrate and provide a uniform access mechanism to all research information in Flanders as RDF, allowing third parties to create services around that data (e.g. reporting) and removing some of the administrative burden of universities. Business Semantics Management method and tools helped in constructing an ontology and was well received by the users.

Publication of data in relational databases became fairly easy with solutions such as D2R server. The triples generated by such tools are rendered "meaningful" by exporting the ontology into an implementation in RDFS/OWL and use these to annotate the instances, since the ontology is the result of collaboration and meaning agreements between stakeholders representing autonomously developed information systems. Future work in that area consists of developing a flexible layer between the generated RDF triples from existing tools and the generated ontology from the Business Semantics Glossary.

References

[1] OMG SBVR, version 1.0, http://www.omg.org/spec/SBVR/1.0/
[2] De Leenheer, P., Christiaens, S., Meersman, R.: Business semantics management: A case study for competency-centric HRM. Computers in Industry 61(8), 760–775 (2010)
[3] Halpin, T.: Information Modeling and Relational Databases. Morgan Kaufmann, San Francisco (2008)
[4] Jörg, B.: CERIF: The common European research information format model. Data Science Journal (9), 24–31 (2010)
[5] Jörg, B., van Grootel, G., Jeffery, K.: Cerif2008xml - 1.1 data exchange format specification. Technical report, euroCRIS (2010)
[6] Spyns, P., Tang, Y., Meersman, R.: An ontology engineering methodology for DOGMA. Applied Ontology 3(1-2), 13–39 (2008)
[7] Spyns, P., van Grootel, G., Jörg, B., Christiaens, S.: Realising a Flemish government innovation information portal with business semantics management. In: Stempfhuber, M., Thidemann, N. (eds.) Connecting Science with Society. The Role of Research Information in a Knowledge-Based Society: University of Aalborg, Aalborg University Press (2010)