

Shared Vocabularies to Support the Creation of
Energy Urban Systems Models. *4th Workshop
organised by the EEB Data Models Community
ICT for Sustainable Places, 2013*

1. Session: Title of the session

1.1. Shared Vocabularies to Support the Creation of Energy Urban Systems Models

<i>L. Madrazo</i>	RAMON LLULL UNIVERSITY (ARC ENGINEERING AND ARCHITECTURE LA SALLE), SPAIN	madrazo@salleurl.edu
<i>G. Nemirovski</i>	ALBSTADT-SIGMARINGEN UNIVERSITY (FACULTY OF BUSINESS AND COMPUTER SCIENCE), GERMANY	nemirovskij@hs-albsig.de
<i>A. Sicilia</i>	RAMON LLULL UNIVERSITY (ARC ENGINEERING AND ARCHITECTURE LA SALLE), SPAIN	sicilia@salleurl.edu

Abstract

The problem of carbon emission reduction in urban areas cannot be constrained to a particular geographical area or scale, nor is it the concern of a particular discipline or expert: it is a systemic problem which involves multiple scales and domains and the collaboration of experts from various fields. The aim of models of urban energy systems is to identify the processes that determine the energy intensity in a specific urban area. Such models can help experts to understand the systems' behaviour and take measures to improve its performance. The application of semantic technologies can help to create urban energy models which integrate the knowledge from experts in various domains. The goal of the SEMANCO research project is to create a comprehensive framework –i.e. methods and tools– using semantic technologies which enable experts from different domains to devise and deploy urban energy models that help various stakeholders –planners, consultants, policy makers– to understand the complexity underlying carbon reduction in urban areas. A key component of the project is the Semantic Energy Information Framework (SEIF) which facilitates the link between the tools which are intrinsic to an energy model and the required data. This paper describes the process and results obtained in the development of this semantic framework. In particular, the paper discusses the creation of its underlying ontology, that is, the vocabulary shared by different domain experts which is necessary to access the contents of the different data sources required by an energy model. The configuration of the urban energy models and the access to the semantic data and the tools that characterise them take place through the SEMANCO integrated platform. Therefore, the current state of the development of this platform is also presented in the paper.

Key words

Semantic technologies, ontologies, urban energy systems, urban energy models

1 Urban energy systems and energy models

Urban energy systems have been defined as “the combined process of acquiring and using energy to satisfy the demands of a given urban area” (Keirstead and Shah, 2013, p.273), whereas an energy system model is “a formal system that represents the combined processes of acquiring and using energy to satisfy the energy service demands of a given urban area” (Keirstead et al., 2012, p.6). A model of an urban energy system fulfils two main purposes: to understand the current state of the system and to help to take decisions to influence its future evolution (Shah, 2013). An urban energy model is expected to provide answers to questions formulated by actors involved in the improvement of the urban energy system’s efficiency. For example, it should enable those actors to address questions such as how much energy is consumed in an urban area, what is that energy used for, how can that consumption be reduced and what are the connections between urban density and energy demand.

A model, according to the definition of Echenique (1972, p.164) is “a representation of a reality, in which the representation is made by the expression of certain relevant characteristics of the observed reality and where reality consists of the objects or systems that exist, have existed or may exist”. Such ‘representation’ is built with a set of abstractions that is, with the methods, data and tools that make the theoretical framework of the model. These capture the internal structure and the dynamics of a system as perceived by the observers. In the case of urban energy models, a multiplicity of these abstractions comes into play, in so far as there are multiple experts and knowledge domains involved in understanding how an urban energy system works. These include experts in energy supply and demand, in transportation networks, in building stock evaluation, in socioeconomic analysis and in environmental policy-making. The multiple models built from the particular point of view of the different observers need to be integrated to create urban energy models which span across various disciplines (Shah, 2013).

One inherent difficulty with urban energy models is the delimitation of the boundaries of the energy systems they represent. As Steinberg and Weisz (2013) have contended, the limits of an energy system can be established in two ways: adopting a ‘production’ perspective, by considering fixed geographical limits based on physical or administrative territorial divisions or, from a ‘consumption’ perspective, by establishing unfixed limits which take into account economic exchanges linked to energy use. As these authors argue, the answers to questions which can be informed by a model –for instance, how much energy a type of building consumes in a city –depend on the limits of the system. Urban energy assessments, therefore, need to include an explicit definition of the systems’ boundary since “arbitrary, or ill-defined, system boundaries defy the very purpose of urban energy assessments: to guide public and private sector policies and decisions and to allow comparability and credibility of the entire process” (Steinberg and Weisz, 2013, p.54).

Ultimately, the value of a model relies on the availability and reliability of the data with which the model operates. Energy related information is dispersed in numerous databases and open data sources and it might have different levels of quality. It is also continuously changing, since urban energy systems are dynamic entities in continuous transformation. Moreover, the information which is required by integrated urban energy models is heterogeneous since it is generated by different applications in various domains. The effectiveness of an energy model depends on having access to the data required for a particular purpose (for example, to compare alternative solutions to reduce energy consumption in an urban area) and on assuring the reliability of the data which is handled by the model, the input data as well as the output data.

2. Semantic technologies and urban energy models

The application of semantic technologies can help to overcome some of the difficulties which are intrinsic to the development of urban energy systems models, in particular those concerning the integration of multiple domains and the accessibility to the data. Ontologies can be used to create shared vocabularies which help

experts from different fields to establish relationships between certain objects of an urban energy system according to their knowledge and experience. An ontology, as formulated by Gruber (1992), stands for “a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents”. Considering this definition, an ontology can be thought of as collectively constructed knowledge that various experts have about an urban energy system. In fact, building a common vocabulary is itself, a knowledge construction process by which the knowledge that the different domain experts have on the issue at stake is made explicit and formal. At this point, there is a fundamental distinction to be made with previous concepts of urban energy models. An urban energy model supported by ontologies built by a group of experts is not just an abstraction of a complex system (e.g. an isomorphism of the system’s structure) but it stands for a way of thinking from multiple perspectives about a complex problem which is embodied in the ontology. In other words, a model is not a representation of a simplified reality, but a representation of a complex reality as conceptualised by experts and formalised in the ontology.

Ontologies can serve to foster communication between the semantically modelled data and the various software applications used by experts. The connections between tools and the data they handle can be captured by the ontologies. This way, when a tool is used within a particular energy model, the data which the tool needs as input can be retrieved via ontologies (in the case of SEMANCO, this function is fulfilled by the Semantic Energy Information Framework). This makes it possible to create multiple urban energy models of an urban energy system, each one with its own set of tools and associated data. This way, semantic technologies can facilitate the interoperability between the semantically modelled data and the variety of tools with which an urban energy model operates.

In the SEMANCO project, semantic technologies are used to create a comprehensive framework which supports the creation –collaboratively and over time– of urban energy systems models. These models represent the combined knowledge of the different experts involved in the evaluation and planning of the system. This framework includes procedures to build an ontology model (i.e. shared vocabularies) and a multiuser platform. The latter enables different users (planners, consultants, policy makers) to create urban energy models and to develop and assess different scenarios to improve the performance of the urban energy system.

3 Using ontologies to model experts’ knowledge

Ontology design is a process by which the knowledge that experts, from one or numerous domains, have is made explicit. In the case of energy urban systems, different experts –planners, consultants, policy makers– know about a particular part of the overall system. Their knowledge is determined by the tools and methods in their particular disciplines, by their experience, and by the information they have at any given moment.

Typically, the knowledge of experts arises as they are confronted with the solution to specific problems. To make this knowledge explicit so that it can be formalised as ontologies, a use case methodology has been applied in three cases studies: Manresa (Spain), Copenhagen (Denmark) and Newcastle (United Kingdom).

Within the SEMANCO project, a case study refers to the delimitation of research scope to a geographic location and to the factors that influence the problem of carbon reduction in a particular urban area. That is, to the stakeholders involved the planning issues at stake and the energy policy agenda (Madrado, 2012). A use case, on the other hand, is a framework which encapsulates data, tools and users and the interactions between them in to fulfil a specific goal within an urban energy system (for instance, reducing carbon emissions at the district level). A use case, therefore, stands for a pre-conceptualization of a model which represents an urban energy system, as thought by experts within a particular context (Figure 1).

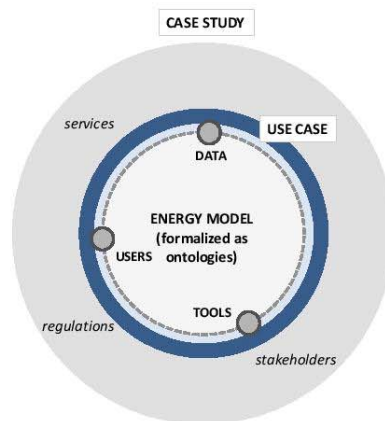


Figure 1. A use case as a pre-conceptualization of the energy model within the context of a case study

To solve the complex problem described by a use case, a series of discrete actions –called *activities*, in the language of the project– need to be undertaken (Figure 2).

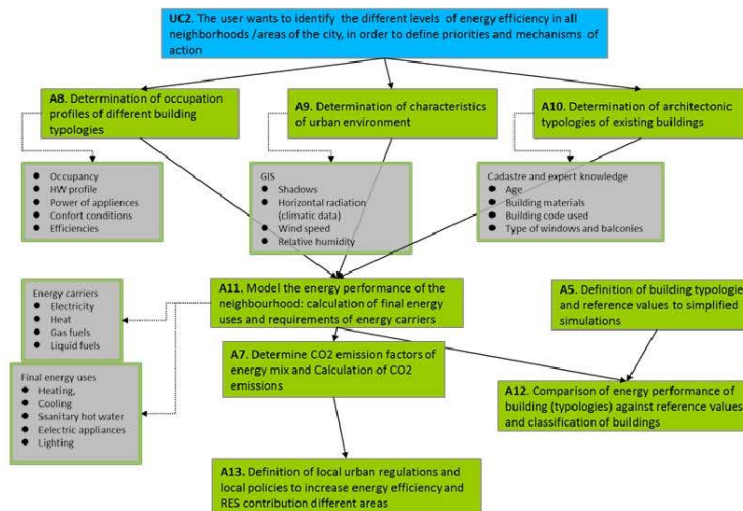


Figure 2. An example of a use case, its activities and the data associated to them.

Use cases and activities defined in this way give rise to a network by which the same activities can be shared by different use cases (Figure 3).

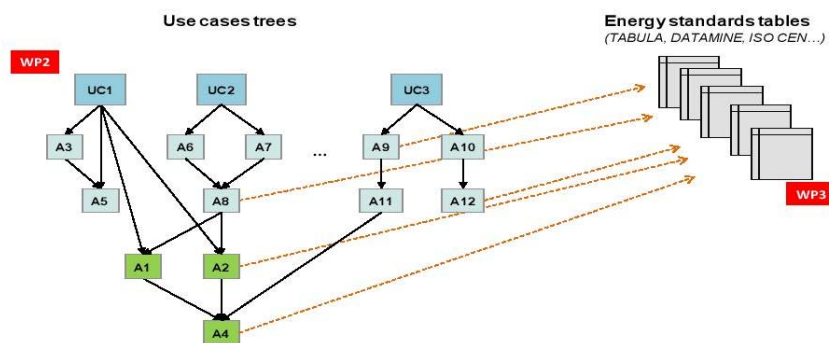


Figure3. Network of activities connected to different use cases

In SEMANCO, use cases and activities are defined by means of templates (Figures 4 and 5) which were specifically created for this purpose. The terms and units of measurement used in the templates are derived from international standards and/or established by the research community. The templates provide enough detail for experts to define a specific issue, while the use of terms based on standards assures that the contents can be transformed into the ontology. Therefore, use cases and activities defined by means of templates are the first step in the construction of a shared vocabulary which can then be formalised as an ontology.

Acronym	UC10
Goal	To calculate the energy consumption, CO2 emissions, costs and for socio-economic benefits of an urban plan for a new or existing development.
Super-use case	None
Sub-use case UC9	
Work process	Planning
Users	1. Municipal technical planners 2. Public companies providing social housing providers 3. Policy Makers
Actors	4. Neighbor's association or individual neighbours: this goal is important for them to know the environmental and socio-economic implications of the different possibilities in the district or environment, mainly in refurbishment projects. They use to ask these questions to the municipality 5. Mayor and municipal councillors: In order to evaluate CO2 emissions impact of different local regulations or taxes
Related national/local policy framework	6. Sustainable energy action plan (Covenant of Mayors) 7. Local urban regulations (PGOUM, PERI, PE in Spain) 8. Technical code of edification and national energy code (CTE, Calener in Spain)
Activities	9. A1.- Define different alternatives for urban planning and local regulations 10. A2.- Define systems and occupation (socio-economic) parameters for each alternative 11. A3. Determine the characteristics of the urban environment 12. A4. Determine the architectural characteristics of the buildings in the urban plans 13. A5. Model or measure the energy performance of the neighbourhood 14. A6. Calculate CO2 emissions and energy savings for each proposed intervention 15. A7. Calculate investment and maintenance costs for each proposed intervention

Figure 4. Template to define a use case

Acronym	A1		
Super-activity/use case	UC10		
Sub-activities	A2, A3, A4		
Goal	Define different alternatives for urban planning and local regulations		
Urban Scale	Micro-Meso		
Users	1. The municipality (councilors of urban planning, housing, environment and countryside...) (stakeholder) 2. Urban planners 3. Public company of social housing 4. Owner/promoter of the building (stakeholder) 5. Neighbor's association (stakeholder) 6. Consultants and technicians from Engineering and consultancy companies 7. Supply companies (i.e. supply company of district heating)		
Related national/local policy framework	<ul style="list-style-type: none"> Sustainable energy action plan (SEP from Covenant of Mayors) Local regulations National energy codes (Código Técnico and certificación energética in Spain, DECC 2012 and HECA in UK, and Heat Planning Act and Danish Planning regulation in Denmark) 		
Issues to be addressed	<ol style="list-style-type: none"> To define the comparison of different CO2 emissions scenarios of urban planning, according to local energy requirements acts and/or Plans, in order to select the most efficient urban planning alternative in next steps. To select a set of technologies, and local regulation in order to evaluate their CO2 impact <ul style="list-style-type: none"> To select different scenarios to evaluate the socio-economic impact of different measures To define alternative building performance levels in order to calculate scenarios of improvement of energy efficiency 		
Input Data			
Name	Description Domain Format		
Local regulations and requirements	Local regulations related to Energy Efficiency, RES, and CO2 emissions, as well as Local Urban regulation that can affect to de different proposals to implement	Energy efficiency Urban planning	Maps, and technical requirements
List of objectives from different users	List of scenarios of energy performance, energy supply, and/or urban planning	Energy efficiency Urban planning	Documents

Figure 5. Template to define an activity within a use

case

Activities templates include references to the data sources required to perform the activities, as well as specifications of the tools and the data required. Altogether, the information collected through the use case and activities templates, in each case study, provide the specifications required to develop the semantic energy framework and the tools associated to it (Figure 6).

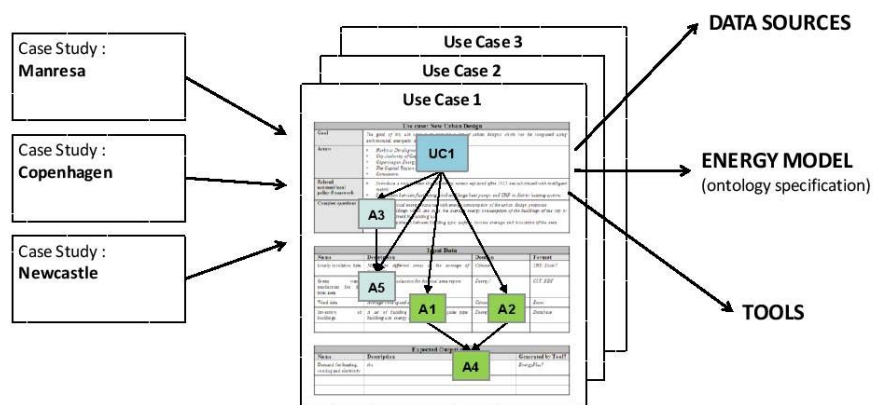


Figure 6. Use cases as links between case studies and the technological development of the project

4 Semantic Energy Information Framework (SEIF)

The Semantic Energy Information Framework (SEIF), developed in SEMANCO, is the nexus between the distributed data sources and the tools using the semantically modelled data (Figure 7). The access to the tools takes place via an integrated platform, which provides services for different types of user.

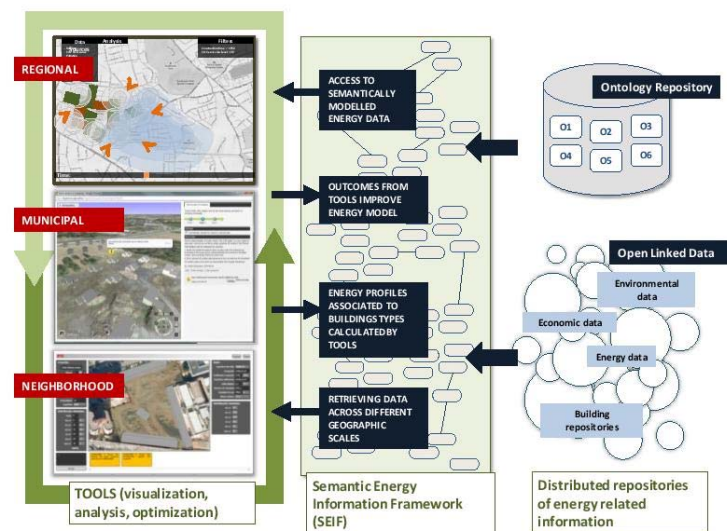


Figure 7. SEIF as a bridge between data and tools

The SEIF has three main goals:

- Integrating proprietary data which is presently off-line or/and heterogeneously structured into a consistent knowledge base, making the data accessible for information discovery and retrieval purposes.
- Providing a bridge between different domains (city planning and energy provision) and contents (consumption data, pollution sources, simulated energy profiles and benchmarks).
- Gathering outputs generated by the tools developed in the project –tools for design evaluation and energy simulation, visualisation and modelling at urban scale, and analysis and optimisation processes– in order to create a distributed knowledge base.

4.1 The ontology building process: creating a semantic energy model

The process of creating an ontology requires a methodological approach to avoid redundant work, to reduce design errors, and to be replicable in other contexts. Generic processes are described by Gruber (1995) and Uschold and King (1995) assuming that ontology design will follow the same process as software development: identification of the requirements, development, evaluation and documentation. This approach is further elaborated by Fernandes, Guizzardi and Guizzardi (2011). A survey of methodologies for ontological design can be found in Fernández-López (1999). However, these methodologies mostly focus on modelling the conceptualisation of a specific domain, rather than on the integration of data sources in ways that support querying using federated access. Besides, it can be argued that a methodology per se is not enough. Rather, it should be supported by design patterns, document templates, tools or platforms which guide developers along the process. Since no methodological approach takes into account the integration of data sources and their querying using federated access, it has been necessary to develop an ontology design process (Nemirovski, Nolle, Sicilia, Ballarini and Corrado, 2013).

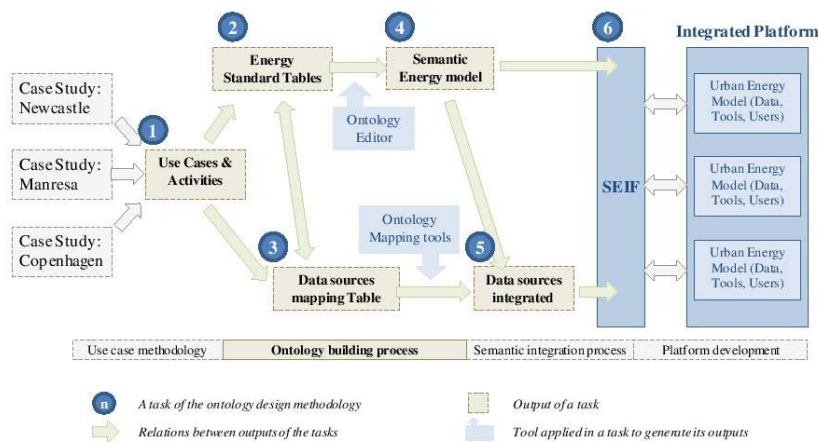


Figure 8. The processes and methods employed to build the SEIF

The methods and processes followed to create the SEIF are summarised in Figure 8. It starts with a description of use cases and activities –according to the use case methodology– from which energy standard tables containing the terms and definitions of the vocabulary which are then transformed into an ontology. In parallel, the data sources are identified and the contents mapped to the terms of the energy standard tables. Finally, the ontology is mapped to the data sources to transform them into Resource Description Framework (RDF) data. Both the semantic energy model (a model of the urban energy system represented as global ontology) and the RDF data sources make the SEIF.

The goal of the process outlined above is twofold: to design a semantic energy model as a formal ontology and to integrate data sources by reorganising them according to the ontology structure. The resulting semantic energy model is a formal global ontology embracing the terminology and relations needed to integrate the data sources and query them in a unified way. This way, the semantic integration process converts the data sources to RDF in accordance with the global ontology.

In the following sections the six main tasks involved in the ontology building process are explained and the outcomes achieved are described.

4.1.1 Vocabulary capture

The first task of the ontology design process is to capture the base terminology for the ontology, that is to say, to make the knowledge that domain experts have about the issues related to a use case explicit. By means of use cases, experts describe how actors, tools, and data relate to each other in order to fulfil a specific goal under a specific policy framework. The activities encompassed by a use case are described in form of requirements and competency questions following current approaches, such as the Neon methodology (Suárez-Figueroa et al., 2012). This way, the data sources required to carry out the activities are identified and briefly described.

The output of the process of vocabulary capture is 14 use cases and 44 activities defined through templates. The actors considered in the use cases encompass social housing providers, city councils, building owners and energy consultants. The policy frameworks considered are local urban regulations, Covenant of Mayors, national building codes, UK Fuel Poverty Strategy among others. The activities deal with a wide range of issues

examples include the identification of areas with high instances of fuel poverty the calculation of the potential of local solar gains, and the calculation of the CO₂ emissions of buildings and urban areas.

4.1.2 Building an initial vocabulary

In the second task, the use cases and activity specifications are analysed with the goal of defining an initial vocabulary. This is a categorised set of terms connected by simple relations such as subsumption (is) and aggregation (has). To build the initially vocabulary it is necessary to identify the data categories, to scrutinise the existing international standards for energy modelling and to create energy standard tables, which are a set of semantically structured terms, including objects, attributes and standard definitions.

The data categories are divided in two major groups: 1. those which concern data on energy systems, energy quantities and boundary conditions, and 2. those concerning contextual data. The first group contains the categories of energy data (e.g. CO₂ emission coefficient, CO₂ emissions, delivered energy, energy demand, energy supply etc.), climatic data (e.g. air temperature, solar irradiance, wind speed, relative humidity etc.), and building technical data (e.g. space heating systems, energy generator, mechanical ventilations, type of walls etc.). Contextual data includes energy costs (e.g. running costs and refurbishment costs), environmental data (e.g. air pollutants and air quality), legislative constrains such as energy performance requirements, geographical and land registry data (e.g. land lots, land value, land classification, etc.), socio-economic and demographic data (e.g. gender, level of education, tenure, income etc.).

The resulting vocabulary requires a common and shared terminology. With this purpose, international technical standards, research projects, and European directives were consulted to obtain the definitions of the terms, the relations between concepts and the symbols and units of the quantities.

The initial vocabulary is specified in the form of an energy standard table. Each category in this table contains numerous terms identified by the various activities. The initial vocabulary contains the description of the terms, and the relations between terms and, in this regard, it can be equated with a formal ontology specification.

Building an initial vocabulary is an important intermediate step towards the design of a semantic energy model. It simplifies formal ontology coding significantly by using a formal language, such as OWL. This task was carried out following the methodology for structuring and semantically modelling energy and contextual data developed in the SEMANCO project (Corrado and Ballarini, 2012, 2013).

The initial vocabulary is composed of 24 categories including building use, climate and building geometry. Around 1000 terms were collected including; descriptions, references, units, and type of data. 18 standards (e.g. ISO/IEC CD 13273-11, ISO/IEC CD 13273-22, EN 156033 and the EN ISO 15927-14) and 16 references (e.g. research project, public recommendations, European directives) were used to create the energy standard tables.

4.1.3 Mapping data sources to vocabularies

The goal of the third task is to map the data entities of the data sources –identified in the activities of the use cases– to the initial vocabulary. If a target data source is a relational database, then the fields of their tables are mapped to the terms of the initial vocabulary. The mappings are specified by data owners and domain experts using a table template. For example, Table 1 shows the mappings of the Manresa census data source.

¹ISO/IEC CD 13273-1:2012. Energy efficiency and renewable energy sources. Common international terminology. Part 1: Energy Efficiency.

²ISO/IEC CD 13273-2:2012. Energy efficiency and renewable energy sources. Common international terminology. Part 2: Renewable Energy Sources.

³EN 15603:2008. Energy performance of buildings - Overall energy use and definition of energy ratings..

⁴EN ISO 15927-1:2002. Hygrothermal performance of buildings. Calculation and presentation of climatic data. Part 1: Monthly and annual means of single meteorological elements.

Data source	Data name (in the Data source)	Data name (in the vocabulary)	Data category (in the vocabulary)
Manresa census	ID	Building	Building
Manresa census	NUMCOD	Address	Building
Manresa census	DOMCOD	Address	Building
Manresa census	ADRDESC	Address	Building
Manresa census	TITULACIO	Education_Level	Housing
Manresa census	SEXE	Household_Type	Housing

Table 1. An activity description

As illustrated in Table 1, the term 'Address' contains in the initial vocabulary it is mapped to the terms NUMCOD, DOMCOD and ADRDESC from the targeted data source. This information is used as an input for the fifth task -Mapping data sources- explained later. Unfortunately, not all of the terms contained in the data sources can be univocally mapped to the initial vocabulary, so it is necessary that an ontology expert deals with some of the less evident mappings. In these cases, ontology experts have three alternatives: to modify/extend the initial vocabulary (which is the most often selected choice); to implement non-trivial mapping preferences; or to specify complex queries.

Nine different data sources have been mapped to the initial vocabulary including census and cadastre records, building typologies, neighbourhoods, energy coefficients among others. In total, more than 60 mappings are established between the data entities of the data sources and the initial vocabulary.

4.1. 4 Ontology coding

The fourth task is focused on the codification of the semantic energy model, as a formal ontology based on the *DL-Lite_A* formalism which outperforms most other description logic formalisms when managing data distributed in heterogeneously structured sources (Poggi et al., 2008). The coding of the semantic energy model is carried out by SEMANCO's ontology editor (Figure 9) described by Wolters, Nemirovski and Nolle (2013). This editor provides a user-friendly interface which facilitates the participation of domain experts in the ontology building process. Besides, the editor supports the coding of *DL-Lite_A* axioms to represent domains and ranges of object properties which require the processing of reasoning. These two features are the main reasons for the development of a bespoke editor instead of using an existing one such as Protégé⁵ or TopBraid Composer⁶. The SEMANCO ontology editor offers the user two simultaneous views of an ontology: one for editing the taxonomy of concepts, and another one for editing the graph of non-subsumption relations.

⁵ <http://protege.stanford.edu>

⁶ http://www.topquadrant.com/products/TB_Composer.html

eeBDM at ICT4SP

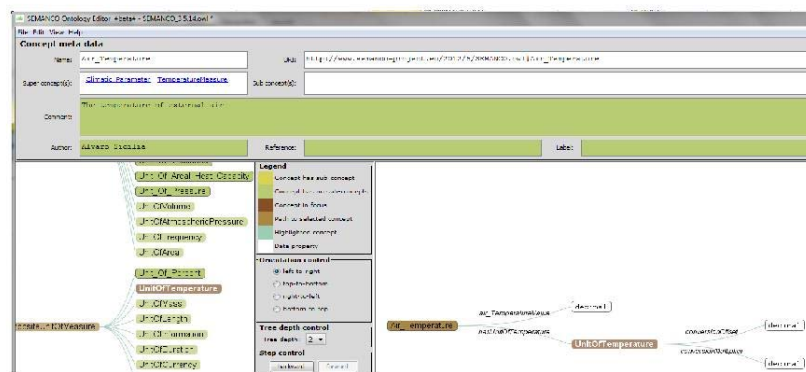


Figure 9. SEMANCO's ontology editor (© Albstadt-Sigmaringen University)

Annotations are key components of an ontology, which enable users to understand its structure and the criteria adopted in their conceptualisation. The ontology editor enables users to define four types of annotation properties for each concept; label, comment, reference and author. The values of the annotation properties are taken directly from the energy standard tables; such as the name, the description and the reference.

Following a modular approach to ontology design, the semantic energy model is built with modules of the Suggested Upper Merged Ontology (SUMO). In this way, each concept of the semantic energy model is subsumed at least by one concept of SUMO. SUMO was selected, rather than DOLCE, PROTON, General Formal Ontology (GFO), and Basic Formal Ontology (BFO) because of its simplicity of understanding, applicability for reasoning and inference purposes, the ability to apply units of measurement to data, and the number of concepts it contains related to the urban planning domain.

The outcome of this task is the creation of a global ontology based on the SUMO upper-ontology encompassing 592 concepts and 468 relations implemented with 3459 axioms in *DL-Lite_A* style.

4.1.5 Mapping data sources

The aim of this task is to apply the informal mappings produced in the previous task to transform the contents of the data sources into RDF resources. After coding the mappings, using a formal language of a dedicated middleware, the data stored in relational databases becomes available for SPARQL querying in terms of the target global ontology.

These mappings are implemented with declarative mapping languages, which offer rich expressive features helping to adjust rigid relational schemas to real cases. In SEMANCO for D2RQ (Bizer and Cyganiak, 2007) was selected. It is supported by the D2R server, a mature and stable lightweight middleware. Nevertheless, other software products, such as Quest (Rodriguez-Muro and Calvanese, 2012) using standard mapping language R2RML are also being tested.

The creation of such mappings is a complex process, which involves experts from different domains with different skills. The process requires them to understand both the structure of the ontology and the data sources. To support their work, two environments were developed using D2RQ and R2RML language. The OWL mapping extractor to extract an OWL ontology file and a D2RQ mapping file from the structure of a relational database, and the ontology mapping collaborative web environment that provides a graphical interface to assist non-ontology experts to implement the mappings (Figure 10).

In the integrated platform, both the experts' knowledge, captured through the use case methodology (use case and activities templates), as well as the links to the external data sources are available through the SEIF (Figure 11). This combination of knowledge and information constitutes the base for creating energy models for a particular urban area.

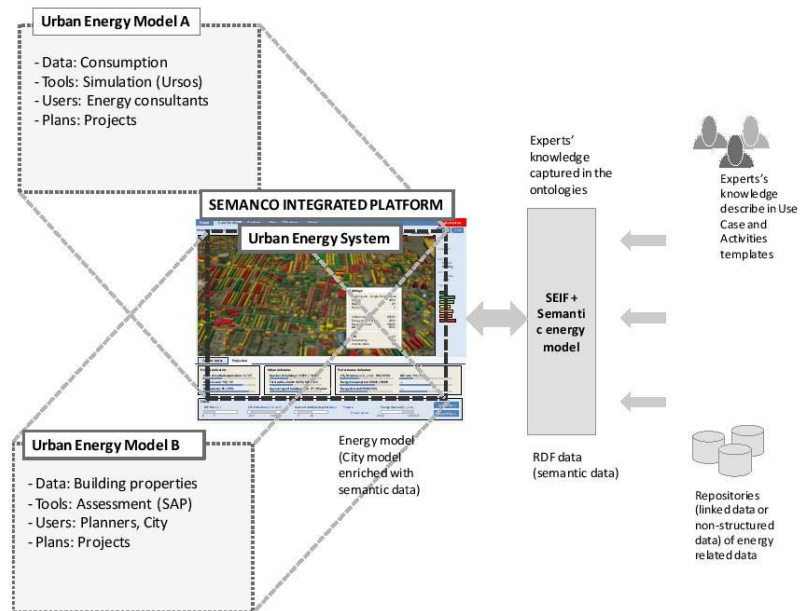


Figure 11. Different models providing partial views of the overall urban energy system

Urban energy models are constructed in an asynchronous manner by adding energy related information to a geometric model created with the 3dMaps software of Agency9 (a project partner). For this purpose, the platform provides different kinds of tools:

- Embedded; tools which are part of the platform and developed specifically for it.
- Interfaced; existing tools (e.g. simulation, assessment) which can interact with other tools and services in the platform.
- External; existing tools that can use data exported from the platform and generate data that can be imported to it.

Within a particular energy model domain experts can represent the existing conditions of the urban system (descriptive model), analyse the future evolution of the system (predictive model), explore different scenarios for future development (exploratory model) and propose improvement plans and evaluate projects to improve the performance of the urban energy system (planning model)⁷ using multicriteria decision analyses tools⁸.

⁷ These four types of models are identified in Echenique (1972).

⁸ Yamaguchi and Shimoda (2010) provide an example of the application of a set of tools to analyse alternatives to improve energy performance in a district within a given energy model.

eeBDM at ICT4SP



Figure 12. Integrated platform (© SEMANCO)



Figure 13. Semantic data explorer (© ARC Engineering and Architecture La Salle)

The platform has been designed to support services for four user groups:

- Domain experts. They collaborate in the construction of an energy model (e.g. describing use cases and activities, defining terms of the ontology), and/or they interact with the model (e.g. extracting reports, enriching the energy model with new data). They produce and evaluate alternative plans to

eeBDM at ICT4SP

improve the performance of the urban energy system, and they provide advanced data analyses services to other experts.

- **Ontology engineers.** They collaborate with domain experts in the maintenance and enhancement of the system's ontology. With this purpose, they use the tools developed for the project to create the energy model as a global ontology (Ontology Editor), to carry out the semantic integration process (Ontology mapping environments), and to verify the outputs of the process (Semantic data explorer).
- **Platform developers.** They assist experts in the integration of new tools and data in the platform.
- **Non-experts.** They interact with the platform –either by themselves or assisted by a domain expert– to visualize the energy data using different tools provided by the platform (3Dmodels, tables and diagrams), to extract the information they need and derive conclusions from it.

Once the project is completed, the integrated platform will provide a generic structure to support the development of services based on the exploitation of the semantic data and the tools interacting with them. Most important, it will be possible to incorporate into the platform additional energy systems from urban areas other than the three case study areas included in the SEMANCO project.

6 Conclusions

In the first two years of the SEMANCO project partners have devised and implemented a methodology to capture experts' knowledge –that is, the implicit knowledge, which experts possess that emerges as they are confronted with a particular problem concerning the performance of an urban energy system– with the purpose of creating a semantic framework to support decision making in energy efficient urban planning. This knowledge has been formalised as a global ontology created with the participation of domain experts and ontology engineers. As a result, a Semantic Energy Information Framework (SEIF) has been created, which provides access both to the experts' knowledge, captured by the terms and relations that form the ontology, and to information required by different energy models based on the ontology. A prototype of the integrated platform, which is currently being finalised, will facilitate access to the energy models for different types of users. Overtime, the use of the platform's services will support the addition of more energy related data, as well as enhancing the system's ontology with new terms and relations. SEMANCO's platform will provide a generic, flexible and open, structure that facilitates the continuous development of complex models of urban energy systems carried out with the participation of the different users and stakeholders.

The results of the SEMANCO project are therefore contributing to the development of integrated urban energy models which can help agents involved to improve the efficiency of urban energy systems by enabling a better understanding of the complexity of the issues involved. In this regard, the most relevant outputs of the project are not its end-products (e.g. the integrated platform and the various tools devised to build the ontologies) but rather, the comprehensive semantic framework which integrates energy accounting methods, energy related data, and energy assessment tools.

7 Acknowledgments

SEMANCO is being carried out with the support of the FP7 Program "ICT systems for Energy Efficiency" of the European Union with the grant number 287534. The use cases, activities, and mapping tables were created with the collaboration of project partners. The energy standard tables were collated by Vincenzo Corrado and Ilaria Ballarini from Politecnico di Torino. The ontology editor is being developed by Michael Wolters from the Albstadt-Sigmaringen University. The SEMANCO global ontology is being coded by German Nemirovski and Álvaro Sicilia. The semantic data explorer is being developed by Joan Pleguezuelos, from ARC Engineering and Architecture La Salle. The authors would like to thank Dr. Tracey Crosbie, from UoT, for making a final review of the manuscript.

8 References

1. Bizer, C. and Cyganiak, R. 2007. D2RQ – Lessons learned. Position paper at the W3C Workshop on RDF Access to Relational Databases, Cambridge, United Kingdom.
2. Corrado, V., and Ballarini, I. 2012. SEMANCO Deliverable 3.2: Guidelines for Structuring Energy Data. Accessed July 31, 2013 at website: http://semanco-project.eu/index_html_files/SEMANCO_D3.2_20130121.pdf
3. Corrado, V., and Ballarini, I. 2013. SEMANCO Deliverable 3.3: Guidelines for Structuring Contextual Data. Accessed July 31, 2013 at website http://semanco-project.eu/index_html_files/SEMANCO_D3.3_20130321.pdf
4. Echenique, M. H. 1972. Models: A discussion. In L. Martin and L. March (editors) *Urban space and structures*. London: Cambridge University Press.
5. Fernandes, B.C.B., Guizzardi, R.S.S. and Guizzardi, G. 2011. Using Goal Modelling to Capture Competency Questions in Ontology-based Systems. *Journal of Information and Data Management*, Vol. 2, No 3, pp. 527-540.
6. Fernández-López M. 1999. Overview of methodologies for building ontologies. In *Proceedings of the IJCAI-99 workshop on ontologies and problem-solving methods: lessons learned and future trend*, Stockholm, Sweden.
7. Gruber, T. R. 1992. A Translation Approach to Portable Ontology Specifications. Accessed July 31, 2013 at website <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>
8. Gruber, T. 1995. Towards principles for the design of ontologies used for knowledge sharing. In *International Journal of Human Computer Studies*, Vol. 43 (5-6), pp. 907-928.
9. Keirstead, J., Jennings, M., and Sivakumar, A. 2012. A review of urban energy system models: approaches, challenges and opportunities. *Renewable and Sustainable Energy Reviews*, 16(6), pp. 3847–3866. Accessed July 31, 2013 at website <https://spiral.imperial.ac.uk/bitstream/10044/1/10206/4/review.pdf>
10. Keirstead, J., and Shah, N. (editors). 2013. *Urban Energy Systems: An Integrated Approach*. Urban energy systems : an integrated approach. London: Routledge.
11. Madrazo, L. (editor) 2012. SEMANCO Deliverable 1.8: Project Methodology. Accessed July 31, 2013 at website http://semanco-project.eu/index_html_files/SEMANCO_D1.8_20120921.pdf
12. Nemirovski G., Nolle A., Sicilia Á., Ballarini I., and Corrado V. 2013. Data Integration Driven Ontology Design, Case Study Smart City. *The Semantic Smart City Workshop (SemCity-13)*, Madrid, Spain.

eeBDM at ICT4SP

13. Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M and Rosati, R. 2008. Linking data to ontologies. *Journal on Data Semantics*, pp. 133–173.
14. Rodríguez-Muro, M. and Calvanese, D. 2012. High Performance Query Answering over DL-Lite Ontologies. In *Proceedings of 13th International Conference on Principles of Knowledge Representation and Reasoning (KR 2012)*, pp. 308-318.
15. Shah, N. 2013. Modelling urban energy systems. In J. Keirstead and N. Shah (eds.) *Urban Energy Systems. An Integrated Approach*. London and New York: Routledge.
16. Steinberger, J. and Weisz, N. 2013. City walls and urban hinterlands: the importance of system boundaries. In A. Grubler and D. Fisk (editors) *Energizing Sustainable Cities. Assessing Urban Energy*. London: Routledge.
17. Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E. and Gangemi, A. 2012. *Ontology Engineering in a Networked World*. Berlin: Springer.
18. Uschold, M. and King, M. 1995. Towards methodology for building ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing*, held in conjunction with IJCAI-95, Cambridge, United Kingdom.
19. Wolters, M. Nemirovski G, and Nolle, A. 2013. ClickOnA: An Editor for DL-LiteA based Ontology Design. *26th International Workshop on Description Logics*. Ulm, Germany.
20. Yamaguchi, Y. and Shimoda, Y. 2010. District-scale simulation for multi-purpose evaluation of urban energy systems. *Journal of Building Performance Simulation*, Vol. 3, Issue 4.

Data Integration Driven Ontology Design, Case Study Smart City. *3rd International Conference on Web Intelligence, Mining and Semantics, 2013*

Data Integration Driven Ontology Design, Case Study Smart City

German Nemirovski,
Andreas Nolle
Albstadt-Sigmaringen University of
Applied Sciences
Albstadt, Germany
nemirovskij@hs-albsig.de
nolle@hs-albsig.de

Álvaro Sicilia
ARC Ingeniería i Arquitectura
La Salle
Universitat Ramon Llull
Barcelona, Spain
asicilia@salle.url.edu

Ilaria Ballarini,
Vincenzo Corado
Department of Energy (DENERG)
Politecnico di Torino
Torino, Italy
vincenzo.corrado@polito.it
illaria.ballarini@polito.it

ABSTRACT

Methods to design of formal ontologies have been in focus of research since the early nineties when their importance and conceivable practical application in engineering sciences had been understood. However, often significant customization of generic methodologies is required when they are applied in tangible scenarios. In this paper, we present a methodology for ontology design developed in the context of data integration. In this scenario, a targeting ontology is applied as a mediator for distinct schemas of individual data sources and, furthermore, as a reference schema for federated data queries. The methodology has been used and evaluated in a case study aiming at integration of buildings' energy and carbon emission related data. We claim that we have made the design process much more efficient and that there is a high potential to reuse the methodology.

Categories and Subject Descriptors

D.3.3 [Information Storage and Retrieval]: Systems and Software – *distributed systems*.

General Terms

Performance, Design, Standardization

Keywords

Ontology Design, Ontology Mapping, Description Logic, DL-Lite family, Data integration, Semantic Web.

1. INTRODUCTION

In the last decade, the paradigm of Semantic Web has gained lots of new ideas through approaches that focus on data integration and semantic interoperability. The cloud of Linked Opened Data has been growing rapidly and become one of the central components of Semantic Web. According to W3C, in 2011; it included over 31 billion RDF triples, stored in over 295 data sources¹. The utmost advantage of federation of distributed data

through interlinking using RDF triples is expected in areas where the heterogeneity of data builds a critical obstacle for its processing. This is when:

- large volumes of data have been stored in data sources supporting different data models,
- data describing characteristics of similar items has been generated using different standardization systems,
- measures characterizing equal physical quantities have been specified using different units of measurement, for example, following standards adopted in different countries.

The Smart City cluster clearly features all of these properties. Approaches like “sustainable low-carbon city” use statistic data for energy consumption and CO₂ emission of buildings that has been collected over many years in municipalities, energy and development companies, architecture offices and standardization organizations. The data stock is basically managed by relational database systems using wide diversity of data models. Taking this into concern, properties of ontologies specifying data semantics become crucial for the integration of this data into the Semantic Web environment.

In this paper we present a methodology for ontology design based on a series of document templates, tools and specifications. This methodology focuses on the requirements emerging in the context of data integration. Its application and effectiveness is shown in examples originated from the SEMANCO project² targeting the development of tools and data integration for the needs of the Smart City cluster.

A case study is highlighted in section 2 as an example of the variety of decisions that can be made in ontology design. Section 3 presents related work. Sections 4 to 8 illustrate details of the methodology. In section 9 we present the most important results and conclusions.

2. CASE STUDY WEATHER DATA

SEMANCO ontology has been developed as a mediator for integration of buildings' technical and statistical data, distributed in a set of heterogeneously structured data sources. Similar ideas of ontology driven data integration can be found in Calvanese [6] and Wang [26]. All data sources use relational schema. The

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS'13, June 12-14, 2013 Madrid, Spain

Copyright © 2013 ACM 978-1-4503-1850-1/13/06... \$10.00

¹ <http://lod-cloud.net/state/>

² <http://www.semanco-project.eu>

ontology should help to interlink this data according to its semantics, facilitate federated querying for the entire data stock and enable semantic interoperability of tools to operate on these data. Thereby relation between the ontology and the integrated sources can be expressed in terms of [18]: a so-called global ontology is defined as a union of elements of local ones, representing the schemas of sources being integrated.

Let us illustrate the desired solution with an example. Given a data source 1 containing city names, weather station names and distances between cities (Table 1). The source 2 contains names of weather stations, temperature values measured at these stations and the dates when they were measured (Table 2).

Table 1. Data source 1

City	Weather station	Distance
Terrassa	Viladecavalls	7
Terrassa	Granollers	25
Manresa	Pont de Vilomara	4
Manresa	Torre d'en Roca	16
Manresa	Ajuntament de Navarçles	15

Table 2. Data source 2

Weather station	Temperature	Date
Viladecavalls	32.3	08.08.12
Granollers	34.5	08.08.12
Pont de Vilomara	38.2	08.08.12
Torre d'en Roca	37.0	08.08.12
Ajuntament de Navarçles	33.2	08.08.12

Let us suppose that a user requests the temperature values for cities measured at particular dates. The expected result of this query will be the following:

```
32.3   Terrassa   08.08.12   (1)
38.2   Manresa   08.08.12
```

To generate these results we have to know which temperature measures are related to particular cities. This information is not contained in the data directly. Nevertheless, a human agent after a short consideration of the data will be able to conclude that weather stations that are close enough to particular cities (For example, less than 10 km apart) can deliver the temperature values of these cities. This simple semantic implication, logical for humans; needs to be specified for the purposes of automated data retrieval, explicitly.

One option is to code the semantics in a query. A SPARQL query returning these results can look like this:

```
SELECT ?temp ?city ?date (2)
WHERE {
  _:ws hasTemperatureMeasure ?tm.
  _:ws relatedTo ?city.
  _:ws distancedBy ?dist
  ?tm hasValue ?temp
  ?tm hasDate ?date
  FILTER (?dist < 10)
}
```

After this query is analyzed by a federated query processor, its parts are sent to particular sources. Afterwards, the results of

subqueries are aggregated as shown in [9]. Yet, the same results could be targeted by a much simpler query:

```
SELECT ?temp ?city ?date (3)
WHERE {
  ?city hasTemperatureMeasure _:tm
  _:tm hasValue ?temp.
  _:tm hasDate ?date
}
```

However, in this case, if the semantic described above is missing in the query, we have to specify it somewhere else, e.g. in a TBox. The role inclusion in line seven of the code below contains one part of the information missing in the query. Namely, it connects the concepts City and TemperatureMeasure (the connection is missing in the data sources).

```
!hasTemperatureMeasure ⊆ City (4)
!hasTemperatureMeasure ⊆ TemperatureMeasure
!closestTo ⊆ City
!closestTo ⊆ WeatherStation
!measuredTemperature ⊆ WeatherStation
!measuredTemperature ⊆ TemperatureMeasure
closestTo ◦ measuredTemperature ⊆ hasTemperatureMeasure
!hasDate ⊆ TemperatureMeasure
Range(hasDate) ≡ rdf:date
!hasValue ⊆ TemperatureMeasure
Range(hasValue) ≡ rdf:decimal
```

If the query and the TBox are specified as shown above, another part of the semantic is still missing: neither TBox nor the Query specify the rule for identification of the closest weather station to a city. Such a rule can be specified in a mapping of the corresponding data source, for example:

```
?ws closestTo ?city →
SELECT weatherStation from DS1 ds1_a WHERE
city='Manresa' and distance=(select
min(distance) from DS1 ds1_b where
ds1_b.distance < 10 and
ds1_b.city=ds1_a.city);
```

Such mappings are supported by tools for publishing of relational databases into a Semantic Web context. These tools rewrite SPARQL queries into SQL format and transform the query results to RDF triples. One of the most popular tools of this sort is D2R Server [3] another perspective mapping tool is Quest [22]. The mapping shown above could look in the D2R syntax as follows:

```
Data source 1: (5)
map:ds1_city a d2rq:ClassMap;
d2rq:dataStorage map:database;
d2rq:uriPattern "city/@@ds1.city@@";
d2rq:class :City.

map:ds1_weatherstation a d2rq:ClassMap;
d2rq:dataStorage map:database;
d2rq:uriPattern "station/@@ds1.weatherstation@@";
d2rq:class :WeatherStation.

map:ds1_cityhasweatherstation a d2rq:PropertyBridge;
d2rq:belongsToClassMap map:ds1_weatherstation;
d2rq:property :closestTo;
d2rq:uriPattern "city/@@ds1.city@@";
d2rq:condition "ds1.distance < 10".
```



```

Data source 2: (6)
map:ds2_weatherstation a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:uriPattern "station/@@ds2.weatherstation@@";
  d2rq:class :WeatherStation.

map:ds2_temperature a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:uriPattern "tempmeasure/@@ds2.temperature@@";
  d2rq:class :TemperatureMeasure.

map:ds2_temperaturevalue a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:ds2_temperature;
  d2rq:property :hasValue;
  d2rq:column "ds2.temperature".

map:ds2_temperaturedate a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:ds2_temperature;
  d2rq:property :hasDate;
  d2rq:column "ds2.date".

map:ds2_weatherstationtemperature a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:ds2_weatherstation;
  d2rq:property :measuredTemperature;
  d2rq:uriPattern "tempmeasure/@@ds2.temperature@@".

```

Question is: Which one of these two alternatives is better? Is there a third one? The choice between alternative designs is not only a question of designers' taste. It may have consequences for business processes, for example it could influence their performance or completeness and soundness of the query results.

In the following sections we will present instruments that determine ontology design decisions at different stages of a project, targeting data integration and semantic interoperability of tools.

3. RELATED WORK

The design of formally specified ontologies has been object of research since the early 1990s. Important work in this context was published by Gruber [14] and Uschold and King [25]. The former work is one of the most quoted in the field of semantic web. Its author defines the properties of ontological knowledge representation with relation to the requirements of engineering sciences. The approach of Uschold and King addresses the design process of ontologies, specifying four phases: identifying ontology purposes, building the ontology, evaluating and documenting. The ontology building phase is subdivided into three steps: 1) ontology capture, 2) ontology coding and 3) integration of existing ontologies. This approach has been further elaborated, for instance in Fernandes [12]. A survey of up-to-date methodologies for ontological design can be found in [10]. It became evident that the methodology per se is not enough. It should be supported by design patterns, document templates, tools or platforms, guiding developers along the methodology steps and making complex design tasks, easier. This requirement led to the development of ontology tools such as Protégé [17], WebODE [1] and OntoEdit [24]. Recent comparative studies of such tools are provided in Khondoker [16] and in Kapoor and Sharma [15]. Fonou-Dombouy and Magda Huisman [13] provide an interesting case study for ontology design.

Furthermore an important aspect of the design methodology is the selection of the formalism, a set of rules and constructors for the ontology specification. The right selection of the formalism usually determines the compromise between the expressive power of the ontology and the processing efficiency of the knowledge represented by the ontology. For example, the Description Logic that is mostly used as the formal basis for the ontology specification comprises a family of formal languages. Some of them like *SOJA* (D), *SRQ* (D), or DL-LiteR have been used as basis for different OWL dialects, i.e. OWL DL, OWL 2 and OWL QL respectively³.

Further approaches related to particular aspects of the proposed methodology are referred to in following sections.

4. METHODOLOGY

From the example provided in section 2 we have learned that the semantics of data can be expressed as a union of elements (concepts, roles and axioms) expressed by an ontology TBox and data source mappings. Furthermore, in [21] this issue is discussed more formally. It is shown that TBox and mappings generally supplement each other. However, they may have unnecessary overlaps. Moreover, as shown above, TBox and mappings specifications should be designed with respect to the required queries. Vice versa, as shown in [7], ontology design determines the efficiency of conjunctive queries, as in the case of queries aiming at retrieval of data properties of individuals (instances of ontology concepts).

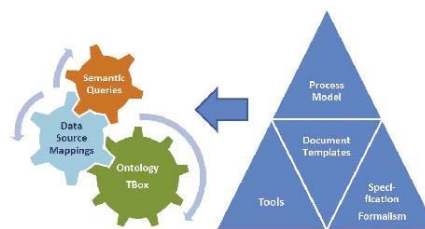


Figure 1. Dependencies between data integration items and parts of the methodology for ontology design

Cross dependencies between queries, TBox definition and mappings increase complexity of the ontology design process. Such dependencies can be easily overlooked by designers. This can lead to severe consequences while a query is processed, like incompleteness of query results or problems with its answer time. The proposed methodology addresses this issue by taking these dependencies into account. As shown in Figure 1 it combines four components: i) an integrated process model for ontology design and data integration ii) a set of document templates supporting designers in every phase of the design/integration process, iii) a set of tools for implementation of TBox and data source mappings exploiting iv) a specification formalism adapted for requirements on data integration.

We argue that the proposed methodology helps to make complex design decisions, for example to decide where to specify parts of query semantics, as described in section 2.

³ <http://www.w3.org/TR/owl2-overview>

5. PROCESS MODEL

The ontology design process is divided into three phases: i) vocabulary building, comprising use cases specification, building of an initial vocabulary and informal mapping of data sources' vocabularies, ii) implementation; that implicates TBox coding and integration of data sources with the help of formal mappings, and iv) evaluation implying the usage of informal specification of final vocabulary and of use cases generated at the beginning of the process. In doing so each of the following phases takes as input the specifications developed in the previous one (Figure 2).

Subdivision of the design process into phases was initially proposed by Uschold and King [19]. Authors defined three phases 1) ontology capture: for instance definition, naming and description of ontology concepts, roles and relations between them; 2) ontology coding: for example specifying the classes and roles using one of the formal languages, for instance OWL; 3) integration of existing ontologies into business processes and tools. This approach has been elaborated thoroughly, in further research work adding some new details like iterations [18] or new phases like scoping, evaluation and documentation [15].

The most important difference between the proposed model and the aforementioned approach is its specialization on data integration. This issue is explicitly addressed by steps 3 and 5. Coming back to the example from section 2, the proposed methodology used already in step 3 would help to identify the conflict between the information required by the user (the temperatures of cities) and the information available in the data sources (temperatures are not associated with cities but with the weather stations that have measured them). Furthermore, in step 5, the design that solves this conflict would be developed. Bringing the query, the TBox and the data source mapping in correspondence with each other is an example of a design that resolves this conflict.

As this will be shown; in the vocabulary building phase the design decisions are supported by document templates and in the implementation phase by a formalism designed to fulfill requirements of data integration, as well as by tools for ontology design and data source mapping.

6. VOCABULARY BUILDING

The vocabulary building phase is divided into three steps which increasingly capture knowledge from the context where the ontology is going to be used and the data sources to be integrated.

6.1 Vocabulary Capture

As mentioned above, we consider query design as an important part of the ontology design. Furthermore, queries are formulated by users or by tools controlled by users. Hence, for understanding the nature of potential queries it is important to take into consideration the users' perspective.

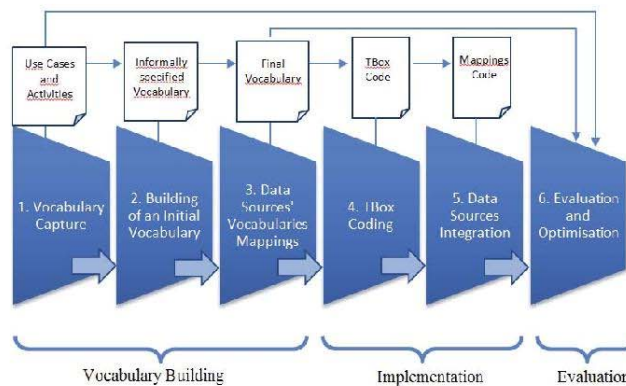


Figure 2. Process model

In the proposed methodology, this task is solved by the use case specifications generated at the beginning of ontology design process. Each use case specification contains a set of activities interconnected by flow lines, determining their sequences. An activity can occur in multiple use cases, so that a network of activities emerges, as shown below. Such specifications help to understand the users' requirements, the needs for data, its semantics, the vocabulary and the desired level of values aggregation. Starting the ontology design process with the specification of the users' perspective is not new: [12] describe an approach of goal modeling which is close to the one presented in this paper. Yet the goal modeling serves to prepare the so called "competency questions", also referred to in [22]. However as long as integration of data sources and information retrieval is focused on, the use cases and activity specifications provide an ideal basis for the formulation of semantic queries. As shown in table 3, an activity description contains a field for specification of all data related to this activity. On the contrary, "competency questions" only appear to be a good instrument for concepts capturing and less appropriate for query design.

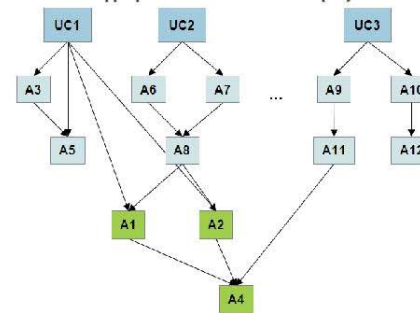


Figure 3. Relationships between activities and use cases

Table 3. An activity description (short version) generated using activity design template which is a part of the proposed methodology

Acronym	A9		
Goal	Determination of characteristics of urban environment		
Urban Scale	Meso –Macro (urban area)		
Process scale	Operational		
Actors	<ul style="list-style-type: none"> • The municipality (councillors of urban planning, housing, environment and countryside, ...) (stakeholder) • Urban Planners, from public authorities or from private companies • Public company of social housing • Owner/promoter of the building • Neighbours association (stakeholder) 		
Related national/local policy framework	<ul style="list-style-type: none"> • National energy code and national technical building construction code (CTE, and RITE) • Nation , regional and local urban planning regulations 		
Issues to be addressed	<ul style="list-style-type: none"> ▪ Volumetric information of the buildings conforming the urban area (to obtain profile of shadows) ▪ Geography of the Area ▪ Location and volume of other urban elements <ul style="list-style-type: none"> ○ Climatic information (Horizontal radiation, wind speed, relative humidity, external temperature) 		
Input Data			
	Name	Description	Domain
	Vector Maps from Manresa GIS	Polygon map showing 3D geometry (buildings footprint, perimeter and height) of the buildings of the urban area	Geography, Manresa GIS
	GIS maps with topographic information	Topographic information of the urban area and surroundings	Geography, Manresa GIS
	Horizontal radiation	Amount of W·h/m ²	Climatic
	Wind speed	Speed of the wind in m/s at the nearest weather station	Climatic
	Relative humidity	Relative humidity at the nearest weather station	Climatic
	Air temperature	Outside Temperature at the nearest weather station	Climatic

Possible semantic queries related to “Air temperature”, referred to the last data entry in the table above, are shown in section 2 of this document. However, no information about the available data is accessible, in this step. For this reason, it is still not clear how the term of “nearest weather station” can be interpreted. Therefore, the query design probably would look similar to (3) at this stage. The query (2) can be formulated only after the available data would have been analyzed.

6.2 Building of an Initial Vocabulary

The second step of the building vocabulary phase is focused on the constitution of an initial vocabulary (Figure 2). The names of the data items in the activity specifications are integrated into the vocabulary from standardization systems, taxonomies of terms or data models, well known in the Smart City context. The correct terminology, the definitions of data names and the relationships among concepts are based on technical standards (For instance, EN ISO 13786⁴, EN 15193⁵, EN 15251⁶ and NREL/TP-550-

38600⁷) and on scientific literature. These references also provide the symbols and the units of the defined quantities, if applicable. The emerging initial vocabulary includes terms and the relations between them. The initial vocabulary is specified in the form of an excel table using the corresponding template. One extraction of such vocabulary is shown in table 4. In this table the name of a relation connecting two terms is written left to these terms. The tree structure of the table determines the other term that is connected by the relation, e.g. *Air Temperature* is a *Climatic Parameter*.

On the one hand, the table shown in table 4 is an important intermediate step towards TBox design. It effectively prepares TBox coding using a formal specification language such as (4), shown in section 2. On the other hand, the completeness of the vocabulary within the use cases originated from the smart city context is guaranteed by the involvement of data specified in the activity description, as the one shown in table 4. For example, the term *Air Temperature* is part of the vocabulary specification twice, once as a climatic parameter and once as a value measured by a weather station. The resulting vocabulary is subdivided in categories, such as building use, climate, and building geometry. Each of these categories contains numerous data names identified in diverse activity descriptions.

⁴ Thermal performance of building components. Dynamic thermal characteristics. Calculation methods.

⁵ Energy performance of buildings.

⁶ Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics.

⁷ Standard Definitions of Building Geometry for Energy Evaluation.

Table 4. An activity description (short version) generated using activity design template

Name/Acronym	Description	Reference	Type of data	Unit
Climate	climatic data	-	-	-
has Climatic_Parameter	climatic parameter	-	-	-
is Air_Temperature	the temperature of external air	EN ISO 15927-1	real	°C
is Solar_Irradiance	radiation power per area generated by the reception of solar radiation on a plane	EN ISO 15927-1*	real	W/m ²
has Solar_Irradiance_Type	type of solar irradiance	-	string	-
is Direct_Solar_Irradiance	irradiance generated by the reception of solar radiation on a plane from a conical angle which surrounds concentrically the apparent solar disk	EN ISO 15927-1*	string	-
is Diffuse_Solar_Irradiance	irradiance generated by the reception of scattered solar radiation from the full sky hemisphere on a plane, with the exception of that solid angle which is used to measure the direct solar irradiance	EN ISO 15927-1*	string	-
is Global_Solar_Irradiance	irradiance generated by reception of solar radiation from the full hemisphere on a plane	EN ISO 15927-1*	String	-
...				
Stationary_Artifact		-	-	-
is Weather_Station		-	-	-
measuredTemperature	Air_Temperature the temperature of external air	EN ISO 15927-1	Climate	°C

6.3 Data Sources' Vocabularies Mappings

In the last step of the vocabulary building phase, Data Sources' Vocabularies Mappings (Figure 2), the names of the data items, used in sources to be integrated, are mapped on the initial vocabulary; as shown in table 4. In the case of relational databases, the fields of a table will be mapped to the terms of the vocabulary. This is done by mapping tables as the one shown in table 5.

Table 5. An activity description (short version)

Data source	Data name (in the Data source)	Data name (in the vocabulary)	Data category (in the vocabulary)
Cataluña Building Data BuildingParametersNONDomestic	average set point temperature	Air_Temperature	Building
Cataluña Building Data BuildingParametersNONDomestic	USE	Building_Use	Building
Cataluña Building Data BuildingParametersNONDomestic	DATE	Year_Of_Construction	Building
Cataluña Building Data BuildingParametersNONDomestic	Orientation main façade	Main_Orientation	Building
Cataluña Building Data BuildingParametersNONDomestic	Orientation main façade: East	MISSING	Building
Cataluña Building Data BuildingParametersNONDomestic	Orientation main façade: West	MISSING	Building

In the data source analyzed in this table, the vocabulary term *Air Temperature* was identified under the name of *average set point temperature*. The corresponding table element serves as an instruction for the following coding of the mapping files. However, not all of the data fields, in the considered document, could be mapped unambiguously (see missing correspondences in table 5). Now, designers are facing three alternative options: to change the initial vocabulary; to implement non-trivial mappings like (5) or to specify complex queries like (2).

7. IMPLEMENTATION

7.1 TBox Coding

The proposed methodology is exploiting the *DL-Lite_A* formalism for the ontology coding and design. The main reason for the use of *DL-Lite_A* was its special features designed w.r.t the requirements of data integration [20]. Furthermore *DL-Lite_A* serves as a basis for the OWL QL profile of OWL 2, designed for the purpose of data accessing/management⁸.

As stated in [18], the most important features of *DL-Lite_A* are the following: 1) domain and range of properties can be specified only for functional data properties; and 2) definition of an object property connecting two OWL classes with each other, has to be modelled by means of axioms and not by specifying the property's domain and range. For example, two following axioms in DL notation use subsumption (\sqsubseteq), existence quantification (\exists) and inversion ($^{-}$) to express that the class *BuildingGeometry* relates to the class *Building* via the *hasGeometry* property.

$$\exists \text{hasExternal_Temperature} \sqsubseteq \text{Building}$$

$$\exists \text{hasExternal_Temperature}^{-} \sqsubseteq \text{External_Temperature}$$

Although domains and ranges of properties are not explicitly specified in the code, if an ontology specification is valid, they can be inferred by reasoner software to be visualized by the user. In this context, using conventional ontology editors like Protégé is time consuming and prone to errors, if used for coding of numerous axioms.

⁸ <http://www.w3.org/TR/owl2-profiles/>

The ontology editor developed in the SEMANCO project provides an instrument to generate a set of axioms defining a relation between two concepts only by a mouse click in the context menu. Besides that, the ontology editor facilitates on the fly inferring of properties' domains and ranges, and enables simultaneous representation of subsumptions' taxonomy with the properties graph (Figure 4). These three features make this editor (to our knowledge) a unique tool for editing *DL-Lite_A* ontologies.

resources. To do so, the mappings established in the step Data sources' vocabularies mappings (step 3) are coded as relations between a relational database and the target ontology TBox created in step 4. These mappings are usually implemented with declarative mapping languages which offer rich expressive features to bring the rigid relational schemas to real cases. The prime example is the RDB to RDF Mapping Language (R2RML)⁹ which became a W3C recommendation in September, 2012 and it is currently being implemented in several projects. However,

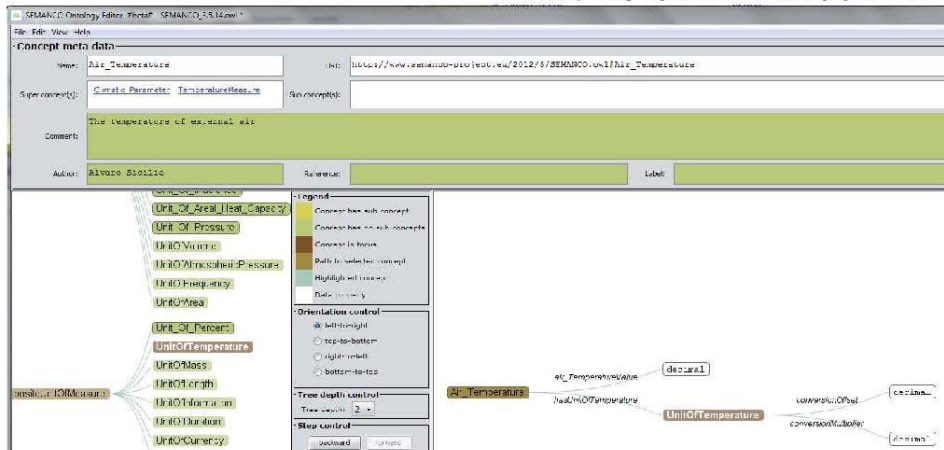


Figure 4. Ontology Editor presents the ontology graph using on the fly inferring

It is important to notice that the selection of a specific formalism like *DL-Lite_A* immediately determines (restricts and simplifies) the TBox design. Returning to the questions formulated in section 2, when *DL-Lite_A* is used, line seven in (4) cannot be specified as follows. The constructor for roles chaining is not a part of this DL language.

closestTo ◦ measuredTemperature ⊆ hasTemperatureMeasure

Consequently, corresponding semantics should be specified somewhere else outside of TBox, e.g. in the query or in the data source mapping. The desirable effect can be achieved by replacing the query (3) through the following:

```
SELECT ?temp ?city ?date
WHERE {
  ?city closestTo _ws.
  _ws measuredTemperature _tm.
  _tm has Value ?temp.
  _tm hasDate ?date
}
```

(7)

Hence the selection of *DL-Lite_A* formalism determines not only specification of TBox but also the form of semantic queries and/or mappings. The last statement is not illustrated here due to lack of space.

7.2 Mapping Data Sources

This step uses the outputs generated by the previous two steps (Figure 2) to transform the contents of the data sources into RDF

other languages can be used for the same purpose, e.g. R₂O [2] and D2RQ [4].

Two environments were developed within the SEMANCO project to help the data sources mapping processes based on D2RQ language: a) the OWL mapping extractor with the purpose of extracting an OWL ontology file and a D2RQ mapping file, reading the structure of a relational database; b) the ontology mapping collaborative web environment that provides a graphical interface to assist non ontology experts to implement the mappings (Figure 5).

The extractor tool uses a configuration file –written in Turtle¹⁰ syntax– to extract the structure of the database. The default tool's behavior is to map each table and column of the database as a class. This can be customized by removing statements or modifying the attributes of the configuration file. The outputs of the extractor tool are an OWL and a D2RQ mapping files like in cases (5) and (6).

8. EVALUATION

After a comparative analysis, we have adapted some ideas related to ontology evaluation described in Gómez-Pérez, [8], Obrst [19], Gangemi [5], and Nemirowskij [18]. In particular w.r.t data integration as the purpose of ontology design, the proposed methodology comprises evaluation of the following three

⁹ <http://www.w3.org/TR/r2rml/>

¹⁰ <http://www.w3.org/TR/turtle/>

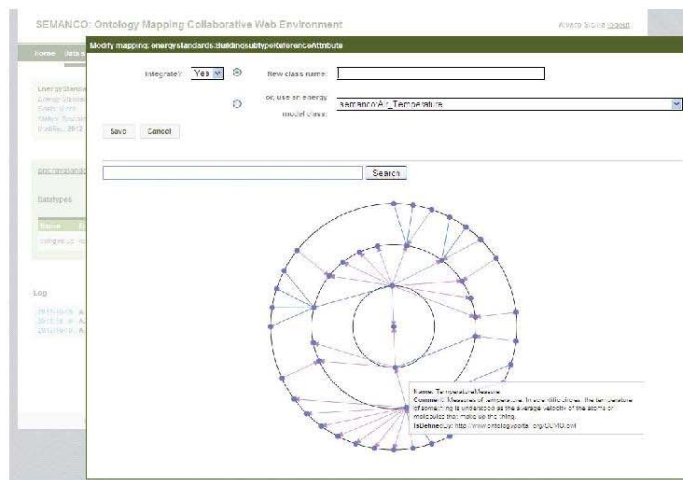


Figure 5. Ontology mapping web environment: ontology graphical representation

Village and the target TBox contains the subsumption $Village \sqsubseteq City$. If this is not the case, the query results cannot be considered complete. Therefore mappings, TBox or queries should be altered.

Computational efficiency: in the focus of this method is the evaluation of query processing. All queries to be evaluated are designed w.r.t. use cases and activity description, specified in step 1 of the design process. Alternative design approaches can be compared to each other, directly. The following table illustrates the method by comparing processing of the query (2) specified in section 2 and the query (7) shown in section 7.1 and using mappings (5)

properties of the resulting ontology, corresponding to three data integration items (Figure 1):

- TBox Intelligibility: the ability of actors that use the ontology to understand the ontology structure.
- Mappings compliance: correspondence of mappings with the TBox
- Computational efficiency: the ability of the ontology to support conjunctive querying on high efficiency level, i.e., with a comparatively short response time.

TBox Intelligibility: especially as a consequence of frequent vocabulary mappings in step 3, there is a risk that the initial structure of the vocabulary designed in step 2 changes significantly and its semantics get unintentionally altered. For the purposes of intelligibility testing, independent testers are asked to find concepts by navigating along the TBox graph. The navigation is done using the editor described in section 7.1. The evaluation is carried out by two independent groups of users, for example of computer science students, and experts in the field of building energy. Each tester is offered a list of terms to find in the ontology. The average score of each group is measured, compared to the shortest navigation path. Our experiments have shown average scores of 97.30%, and 91.20% for each group correspondingly.

Mappings compliance: as stated in [21], a new TBox emerges as a result of a data source mapping. The goal of this evaluation strategy is to make such a TBox explicit and to compare it with the target specified in step 4. This is done by generating an OWL code, out of mapping files. The task is carried out by the mapping environment described in 7.2. As mentioned in section 2, TBoxes generated from mappings should be subset of the target. On the other hand such TBoxes have to contain concepts and properties used in basic graph patterns of queries, e.g. lines 2, 3 and 4 of (3) or lines 2, 3, 4 and 5 of (7). If the query (3) is in use w.r.t. target TBox (4), at least one of the mapping TBoxes should contain the concepts *City*, *TemperatureMeasure*, *Date*, and properties *hasTemperatureMeasure*, *hasValue* and *hasDate*. Alternatively, these elements should be inferable w.r.t. entailment regimes [11], for instance, if a query contains a basic graph pattern "?city a City", it is sufficient if a mapping TBox contains a concept

and (6). The query (2) uses slightly simpler mappings. Five measures have been made for each query. While there is no difference w.r.t. completeness (the right column); the second query constantly shows better time performance.

Table 6. Query performance evaluation

Query ID	Time (in minutes, seconds, and milliseconds.)	Records retrieved
(2)	1:33:45.384	16566
(2)	1:31:08.581	16566
(2)	1:32:23.737	16566
(2)	1:30:35.088	16566
(2)	1:31:36.434	16566
(3)	1:17:30.026	16566
(3)	1:17:17.816	16566
(3)	1:17:33.300	16566
(3)	1:17:46.940	16566
(3)	1:17:27.311	16566

An obvious explanation for this is that the mathematical comparison "datasource1.distance < 10" is specified in the mapping is carried out by native methods of a data source that perform better than ones specified in a SPARQL query "FILTER (?dist > 10)" and consequently, running on RDF data.

9. CONCLUSIONS

In this paper we have described a methodology for ontology design addressing the needs of approaches using ontologies for data integration. We have shown that in this case the design process apart from the ontology TBox has to target semantic queries and mapping of data source. The methodology includes four components: a process model, a set of document templates, a specification formalism *DL-Lite_A* and a set of tools for the simplification of the coding.

To our knowledge the methodology is unique. There are a few approaches addressing ontology design, the most relevant ones are mentioned in this paper. However, none of the existing methodologies put the data integration into focus. Hence, these approaches basically target the development of a TBox and in some cases of an ABox, but do not address query and mapping design.

The efficiency of the approach as a whole, and of its components as well, has been proved by its application. The complete approach has been applied in the SEMANCO project. Within the first 18 months of project time, 592 TBox concepts and 468 relations in *DL-Lite_A* style have been implemented with 3459 axioms, 244 corresponding mappings have been done and 25 queries have been tested.

Furthermore, the ontology editor and the mapping tool presented in this paper have been designed to address generic problems of data integration. During the last year, previous versions of ontology editor and of the mapping tool have been applied in other projects concentrated on data integration issues. This is the case of REPENER. It is estimated that around 71 TBox concepts, 100 relations using 858 axioms in *DL-Lite_A* style have been developed, using these tools. Moreover, the high level of standardization and modularization of the code – the code has been developed using Jena¹¹ and CodeIgniter¹² frameworks – simplify the customization of tools and their reuse for alternative purposes.

Acknowledgement

The main contribution of this work has been developed under SEMANCO project, which is being carried out with the support of the Seventh Framework Programme “ICT for Energy Systems” 2011–2014, under the grant agreement no. 287534.

10. REFERENCES

- [1] Arpírez, J.C., Corcho, O., Fernández-López, M., Gómez-Pérez, A. 2001. WebODE: a scalable workbench for ontological engineering. In *Proceedings K-CAP '01 Proceedings of the 1st international conference on Knowledge capture*. 6-13, ACM New York, USA.
- [2] Barrasa, J., Corcho, O. and Gómez-Pérez, A. 2004. R2O, an Extensible and Semantically Based Database-to-Ontology Mapping Language. In *Proceedings of the Second International Workshop on Semantic Web and Databases (SWDB 2004)*. Springer, 1069–1070.
- [3] Bizer, C. and Cyganiak, R. 2006. D2R Server - Publishing Relational Databases on the Semantic Web. In *Proceedings of 5th International Semantic Web Conference (ISWC '06)*.
- [4] Bizer, C. and Cyganiak, R., 2007. D2RQ – Lessons learned. Position paper at the *W3C Workshop on RDF Access to Relational Databases*. Cambridge, USA.
- [5] Gangemi, Catenacci, A.C., Ciaramita, M. and Lehmann, J. 2005. Ontology evaluation and validation: an integrated formal model for the quality diagnostic task. *Technical report*, Laboratory of Applied Ontologies-CNR, Italy.
- [6] Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D. and Rosati, R. 1998. Description Logic Framework for Information Integration. In *Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR-98)*. Italy.
- [7] Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M. and Rosati, R. 2007. Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated Reasoning*, 39(3), 385–429.
- [8] Gómez-Pérez, A. 2004. Ontology evaluation. *Handbook on Ontologies*, Staab, S. and Studer, R. Eds., (1st ed.), Chapter 13, 251-274, Springer.
- [9] Görlitz, O. and Staab, S. 2011. Federated Data Management and Query Optimization for Linked Open Data. In *New Directions in Web Data Management*, A. Vakali & L.C. Jain Eds.: 1, SCI 331, 109–137.
- [10] Contreras, J. and Martínez-Comeche, J. 2008. Ontologías: ontologías y recuperación de información. DOI=http://www.sedic.es/gt_normalizacion_tutorial_ontologias.pdf
- [11] Glimm, B. 2011. Using SPARQL with RDFS and OWL Entailment regimes. In *Reasoning Web 2011*, volume 6848 of *Lecture Notes in Computer Science*, 137–201.
- [12] Fernandes, B.C.B., Guizzardi, R.S.S. and Guizzardi, G. 2011. Using Goal Modeling to Capture Competency Questions in Ontology-based Systems. *Journal of Information and Data Management*, Vol 2, No 3, 527-540.
- [13] Fonou-Dombeu, J.V. and Huisman, M. 2011. Combining Ontology Development Methodologies and Semantic Web Platforms for E-government Domain Ontology Development. *International Journal of Web & Semantic Technology (IJWesT)* Vol.2, No.2.
- [14] Gruber, T. 1995. Towards principles for the design of ontologies used for knowledge sharing. In *International Journal of Human Computer Studies*, Vol. 43 (5-6), 907-928.
- [15] Kapoor, B. and Sharma, S. 2010. A Comparative Study Ontology Building Tools for Semantic Web Applications. *International journal of Web & Semantic Technology (IJWesT)* Vol.1, Num.3.
- [16] Khondoker, M.R., Mueller, P. 2010. Comparing Ontology Development Tools Based on an Online Survey. In *Proceedings of the World Congress on Engineering 2010 Vol 1 WCE 2010*, London, U.K.
- [17] Knublauch, H., Ferguson, R.W., Noy, N. F. and Musen, M. A. 2004. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In *Proceedings of the Third International Semantic Web Conference*, Lecture Notes in Computer Science, Hiroshima, Japan, November 7-11., 229-243.
- [18] Nemirowski, G., Sicilia, A., Galán, F., Massetti, M. and Madrazo, L. 2012. Ontological Representation of Knowledge Related to Building Energy-efficiency. In *Proceedings of the Sixth International Conference on Advances in Semantic Processing*, Barcelona.
- [19] Obrst, L., Ceusters, W., Mani, I., Ray, S. and Smith, B. 2007. The evaluation of ontologies. In *Revolutionizing Knowledge Discovery in the Life Sciences*, C.J.O. Baker, and K.-H. Cheung, Eds., Chapter 7, 139-158, Springer.
- [20] Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M. and Rosati, R. 2008. Linking data to ontologies. *Journal on Data Semantics*, 133–173.
- [21] Rodríguez-Muro, M. and Calvanese, D. 2012. High Performance Query Answering over DL-Lite Ontologies. In *Proceedings of 13th International Conference on Principles of Knowledge Representation and Reasoning*, Rom.
- [22] Rodríguez-Muro, M. and Calvanese, D. 2012. Quest, a System for Ontology Based Data Access. In *OWLED 2012*.
- [23] Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E. and Gangemi, A. 2012. *Ontology Engineering in a Networked World*. Berlin: Springer.

¹¹ <http://jena.apache.org/>

¹² <http://ellislab.com/codeigniter>

- [24] Sure, Y., Erdmann, M. Angele, J., Staab, S Studer, R. and Wenke, D. 2002. OntoEdit: Collaborative Ontology Development for the SemanticWeb, In *Proceedings of First International Semantic Web Conference (ISWC 2002)*. Horrocks and Hendler Eds. Vol. 2342 of LNCS, 221–235, Springer-Verlag Berlin.
- [25] Uschold, M. and King, M. 1995. Towards methodology for building ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*, Canada.
- [26] Wang, J., Lu, J., Zhang, Y., Miao, Z. and Zhou, B. 2009. Integrating Heterogeneous Data Source Using Ontology. *JOURNAL OF SOFTWARE*, VOL. 4, NO. 8.