The RÉPENER Linked Dataset

Editor(s): Oscar Corcho, Universidad Politécnica de Madrid, Spain; Jens Lehmann, University of Leipzig, Germany
Open review(s): Danh Le Phuoc, DERI, Galway, Ireland; Raúl García-Castro, Universidad Politécnica de Madrid, Spain; Michael Lutz, European Commission – Joint Research Centre, Italy

Álvaro Sicilia[a,*], German Nemirovski[b], Marco Massetti[a] and Leandro Madrazo[a]
[a] *ARC Enginyeria i Arquitectura La Salle, Universitat Ramon Llull, Barcelona, Spain*
E-mail: *{asicilia, mmassetti, madrazo}@salle.url.edu*
[b]*Business and Computer Science Albstadt-Sigmaringen-University of Applied Sciences Albstadt, Germany*
E-mail: *nemirovskij@hs-albsig.de*

Abstract. The dataset presented in this paper constitutes one of the outcomes of RÉPENER, a research project co-funded by the Spanish National RDI plan. It contains integrated information of the Spanish territory regarding energy certification, building monitoring, and geographical data. The integration has been carried out by means of semantic technologies. The adherence to the Linked Data principles guarantees the application of standard methods of accessing data as well as the links to the existing dataset on the Web of Data. The dataset is a knowledge base for end-users. It can be useful for stakeholders involved in the improvement of energy efficiency of buildings to improve their decision-making.

Keywords: energy efficiency, energy certification, data integration process, ontology, Linked Data

1. Introduction

Nowadays, improving the energy efficiency of new and existing buildings is a key issue in European Union policies. In order to design and build more efficient buildings, it is necessary to have a better knowledge of the relationship between design and operation, that is, between the initial design objectives and the actual performance of the building. Likewise, the improvement of existing buildings requires an extensive knowledge of their actual performance. Altogether, there is a need to have integrated access to energy information at the different stages of the building life-cycle –from design to construction and to operation. In fact, having access to the information on request and with the appropriate quality has become crucial for stakeholders involved in the improvement of building energy performance. Having access to this information may help in the design of new buildings, in the renovation of existing ones, and in the tuning of building energy management systems.

The dataset presented in this paper combines data from multiple sources in order to create a knowledge base which helps end-users in their decision making process. It contains energy-related data including the physical and environmental characteristics of a building, use profiles and consumption values. This dataset is one of the outcomes of the RÉPENER research project [1]. The data is exploited by a Semantic Energy Information System (SEíS) which provides services to different user profiles to analyze the data.

This paper is structured as follows. Section 2 describes the main features of the data sources. Section 3 presents the data modelling. Section 4 describes the current use of the dataset including the SEÍS services. Section 5 presents the related work. And, finally, in section 6 the conclusions are summarized.

2. The RÉPENER dataset

The goal of the dataset is to collect data from the different stages of the building life-cycle.

[*] Corresponding author.

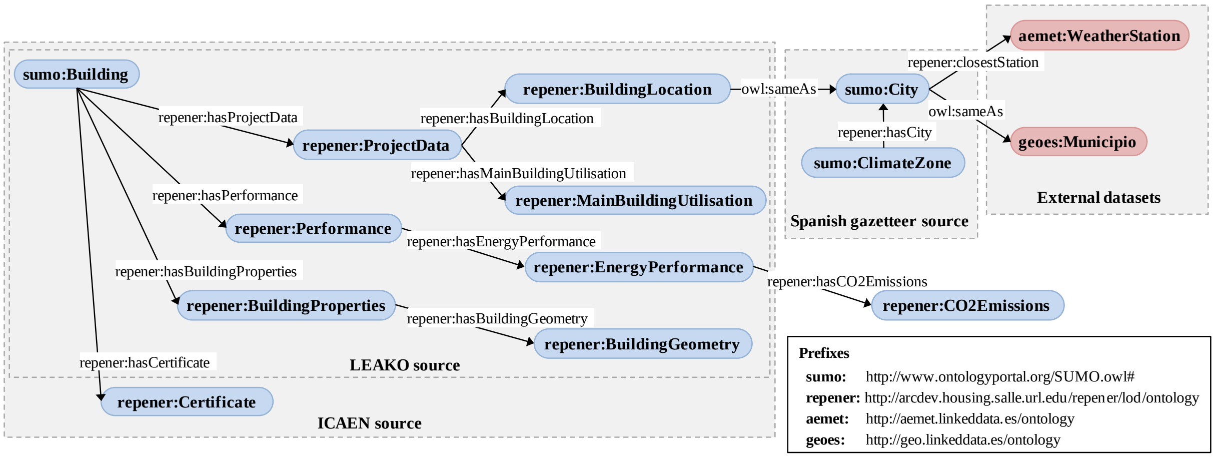
sumo:Building
repener:hasProjectData
repener:ProjectData
repener:hasBuildingLocation
repener:BuildingLocation
owl:sameAs
sumo:City
repener:closestStation
aemet:WeatherStation
owl:sameAs
geoes:Municipio
repener:hasCity
sumo:ClimateZone
repener:hasMainBuildingUtilisation
repener:MainBuildingUtilisation
repener:hasPerformance
repener:Performance
repener:hasEnergyPerformance
repener:EnergyPerformance
repener:hasCO2Emissions
repener:CO2Emissions
repener:hasBuildingProperties
repener:BuildingProperties
repener:hasBuildingGeometry
repener:BuildingGeometry
repener:hasCertificate
repener:Certificate
LEAKO source
ICAEN source
Spanish gazetteer source
External datasets
Prefixes
sumo: http://www.ontologyportal.org/SUMO.owl#
repener: http://arcdev.housing.salle.url.edu/repener/lod/ontology
aemet: http://aemet.linkeddata.es/ontology
geoes: http://geo.linkeddata.es/ontology

Fig. 1. RÉPENER’s ontology excerpt.

The dataset is the result of the integration of three data sources: energy certifications provided by the Catalan Energy Institute (ICAEN), consumption data facilitated by Leako and geographic information from the Geographical Information National Institute (CNIG).

Energy certifications of buildings, collected by public administrations, specifically by ICAEN[1] are the first and main source of data. The data comprises energy certifications and their simulated performance during several stages of the building life-cycle including design and refurbishment. ICAEN provides data in a single spreadsheet in which each row is an energy certificate of a specific building. Each energy certification contains the energy rating of the building, energy consumptions, types of the HVAC (Heating, ventilation, and air conditioning) systems, and geometric features such as the built surface. The ICAEN facilitated more than 1800 energy certifications of which 202 were selected because they contained simulation data. Since some relevant attributes, such as consumptions and emissions were not available, approximation values have been derived from existing studies on the energy consumption of buildings [2], the standard values of ISO [3] while taking into account the Spanish regulations.

The second source of data contains monitoring data of buildings. It is provided by Leako[2], a Basque company dedicated to HVAC installation, distribution and control, which maintains a Paradox database of energy consumption data (e.g., thermal consumption for air and water heating, and water consumption) and indoor conditions (e.g., air temperature) for several buildings.

In the first place, we initially considered using the GeoLinkedData.es dataset[3], but because it lacked detailed data about cities (such as population, surface, or elevation), the Spanish gazetteer –provided by the Geographical Information National Institute (CNIG)[4]– was selected instead. It is a Microsoft Access database which stores geographical data on the populated areas of the Spanish territory including their population, area, elevation and geometry specified in Universal Transverse Mercator (UTM) coordinates. This source does not include a climate zone classification which is relevant for the SEÍS services as described in section 4. For this reason, we have estimated the climate zone for each populated area based on the Spanish Building Code (CTE) which provides a distribution of climate zones per capital province.

3. Dataset modelling

RÉPENER’s ontology has been used to specify the data schemas of the individual sources mentioned above in a single model. A comprehensive description of the ontology design process is provided in [4]. The domain of the ontology is the building energy performance. It adopts many elements from energy standards such as the energy certification of buildings defined by the DATAMINE project [5] and the ISO CEN standards that follow the European Directive

[1] http://www.gencat.cat/icaen
[2] http://www.leako.com
[3] http://geo.linkeddata.es
[4] http://www.cnig.es

2002/91/EC (for example, ISO 13790:2008). These standards cover some areas of the core ontology. They are defined as follows: general project data (e.g., location and use), performance indicators (e.g., energy consumption and CO2 emissions), building properties (e.g., geometric characteristics), outdoor environment (e.g., outdoor temperature and solar radiation), operation (e.g., occupation, comfort levels, thermostat regulation), and certification (e.g., energy efficiency rating and certification-process methodology).

RÉPENER's ontology uses an upper-ontology. The Suggested Upper Merged Ontology (SUMO) [6] has been selected because it can be applied for reasoning and inference purposes. It includes domain-related units of measure such as meter, watt, or joule. The RÉPENER ontology is coded in the OWL *DL-Lite*$_A$ formalism which outperforms –in terms of computability in specific cases such as conjunctive queries over large data volumes– the conventional OWL language. The ontology embraces 71 classes and 100 properties in *DL-Lite*$_A$ style, implemented with 858 axioms. Figure 1 shows a small part of the ontology including classes, object properties, and links to external datasets (repener:closestStation and owl:sameAs).

3.1. Data transformation

The dataset has been created and updated through an ETL (Extract, Transform and Load) process, which converts the data sources into RDF according to RÉPENER's ontology. The components of the process can be seen in Figure 2. The challenge of the process resides in the heterogeneity of the sources –spreadsheets, Paradox database, and Microsoft Access– with a direct impact on the extract phase. The implementation of the three phases is described below:

Extract. Paradox is an obsolete database which does not provide interfaces to be used by current tools. For this reason, a script has been implemented to move the contents of the Paradox files to a MySQL database which is reachable by a D2R Server. In addition, the data extracted from Paradox files have been aggregated from hourly to monthly values since the SEÍS services do not require low levels of data aggregation. The ICAEN spreadsheet has been also migrated to a MySQL database.

Transform. This phase consists on creating a D2RQ [7] mapping file for each source. Mappings have been carried out by ontology engineers, translating each table and column of the databases to reflect the correct term and property from the ontology. Some classes, such as *repener:ConditionedFloorArea*, have to relate themselves to units of measure. For this reason, additional mappings have been done. Furthermore, resources contain annotation properties such as *rdfs:label*. Finally, the values of the use of building (*repener:mainBuildingUtilisation*) have been converted –through D2RQ language constructs– to the classification provided by the DATAMINE project [5], an international domain reference. In this way, third parties from other countries are able to understand the data.

Load. Since all three sources have been mapped to the same ontology, their integration directly merges the three RDF dumps. The resulting file has been uploaded to a Virtuoso server[5].

The dataset updating is carried out manually and the ETL process is executed with the new data because the data sources update frequency is very low.

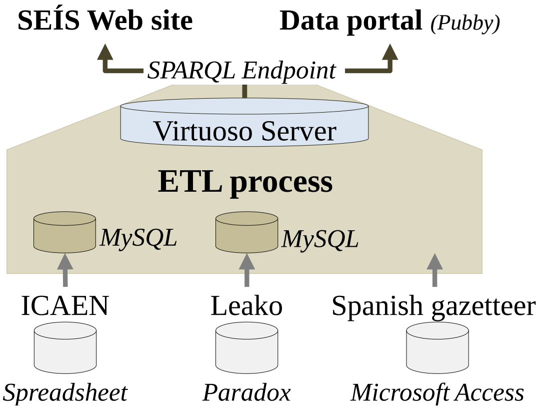

Fig. 2. Components of the RÉPENER dataset creation process.

3.2. URI design

All elements of the dataset have this base URI: *http://arcdev.housing.salle.url.edu/repener/lod/*. The concepts and properties of RÉPENER's ontology can be found under this URI: *http://arcdev.housing.salle.url.edu/repener/lod/ontology*/{class|property}. Each concept has some annotation properties such as *rdfs:label*, *rdfs:comment*, *repener:reference*, and *repener:author*. Comment and reference properties are important because of their usage on RÉPENER's website[6], helping users to un-

[5] http://virtuoso.openlinksw.com
[6] http://arcdev.housing.salle.url.edu/repener

derstand the data they are visualizing and the energy standard data is based on. An example of a concept is *repener:CO2emissions* (see *http://arcdev.housing.salle.url.edu/repener/lod/ontology/CO2emissions*).

Regarding the resources, the URI pattern, selected to identify the instances, uses Patterned URIs solution [8]. This pattern was selected as people are able to read it and it is easily generated from a database where identifiers (for example primary keys) are always present. Furthermore, adding a class name to the base URI mitigates the problem of generating different individuals with the same identifier but different class. Generally, the Natural Keys pattern [8] has been applied to model the URI identifiers (for example, http://arcdev.housing.salle.url.edu/repener/lod/resource/city/Lloret_de_Mar). In this case, a text property of the resource has been converted using the urify[7] pattern which applies a URL encoding and converts the spaces to underscores. In some cases, the identifier has been created following URL Slug pattern [8] to ease dataset exploration.

3.3. Data linking

The ETL process described in the previous *3.1 Data transformation* section has been the first step of the data integration process. The second step is to interconnect the data from the different sources in order to provide combined access to data that has originated from different sources and domains. We have adopted two strategies to connect the data sources:

- The same URI patterns in different data sources have been used to model the same type of resources. This can be done if the sources contain the same values for describing the data. For example, *sumo:ClimateZone* in which resources are generated by both the ICAEN source and the Spanish gazetteer. In both sources, the climate zones are identified with a character and a number, based on the Spanish Building Code. For instance, a climate zone resource such as C2 (see *http://arcdev.housing.salle.url.edu/repener/lod/page/climatezone/C2*) contains data from both sources.
- When the strategy previously described could not be applied, internal links between the data were generated. The SILK framework, described in [9], has been used to connect the building location resources (from ICAEN's and Leako's sources) with the populated places (from the Spanish gazetteer) using *owl:sameAs* relations.

The data sources have also been connected to external datasets, such as the Aemet meteorological dataset[8] and the GeoLinkedData.es, thus enriching the Web of Data with Spanish geospatial data. In total, 783 links have been established with the Aemet dataset and 7160 links with the GeoLinkedData.es dataset.

The Aemet dataset provides climate data from the Spanish Meteorological Office gathered from 204 weather stations across Spain. This connection is relevant since the outdoor environmental properties of the buildings can be enhanced with the data monitored by the Aemet's weather stations. The SILK framework has been configured to discover *repener:closestStation* links between *repener:City* and *aemet:WeatherStation* instances using a geographical distance measure with a maximum distance of 50 kilometres between the city and the station.

The GeoLinkedData.es dataset publishes diverse information sources of the National Geographic Institute of Spain (IGN-E) and the National Statistic Institute in Spain (INE), among others [10]. Some of the data in this dataset complements those of RÉPENER. This is an advantage for users since they then have access to different but complementary information of the same domain. The connection to the GeoLinkedData.es dataset is significant due to the presence of geographical relations between other entities. These are the cases of province capitals (*geoes:esCapitalDe*) [9] and parts of a region (*geoes:formaParteDe*) [10]. Furthermore, this dataset already contains links to the GADM dataset which provides different geometry descriptions of a spatial element for different scales. In this case, an aggregation of a character-based distance measure *(Levenshtein*) and a geographical distance have been designed to generate *owl:sameAs* links between *repener:City* and *geoes:Municipio* instances. The geographical distance is useful as it voids false positives when cities with the same name are located in different areas.

[7] http://d2rq.org/d2rq-language#dfn-uri-pattern

[8] http://aemet.linkeddata.es/

[9] http://geo.linkeddata.es/ontology/esCapitalDe

[10] http://geo.linkeddata.es/ontology/formaParteDe

3.4. Data publishing

Data is accessible through the SPARQL endpoint provided by the Virtuoso server, used by RÉPENER's data portal and by the SEÍS end-user services. The data portal has been implemented with the Pubby[11], a tool which provides ontology and data following the Linked Data principles.

The dataset includes the outputs of the ETL process as well as the links generated to internally connect the sources and the links to external datasets. Table 1 provides a summary of the main features of the dataset.

Table 1

Overview of the dataset features

VoID file	http://arcdev.housing.salle.url.edu/repener/void/repener.ttl
Homepage	http://www.seis-system.org
DataHub entry	http://datahub.io/dataset/repener-building-energy
Ontology file	http://arcdev.housing.salle.url.edu/repener/repener.owl
License	http://creativecommons.org/licenses/by/3.0/
Base URI for instances	http://arcdev.housing.salle.url.edu/repener/lod/resource/
SPARQL endpoint	http://arcdev.housing.salle.url.edu/repener/sparql
Graph name	http://arcdev.housing.salle.url.edu/repener/lod
Example class	http://arcdev.housing.salle.url.edu/repener/lod/page/ontology/TotalPrimaryEnergy
Example resource	http://arcdev.housing.salle.url.edu/repener/lod/page/building/001B00126908P0
Number of triples	150297
Number of distinct subjects	18962
Number of distinct objects	26097
owl:sameAs links	7239
repener:closestStation links	783

4. Dataset exploitation

The dataset is mainly exploited by the four end-user services which have so far been integrated into the SEÍS system[12]. SEÍS accesses RÉPENER's dataset endpoint directly to retrieve the data. Furthermore, the labels and tooltip descriptions are retrieved from RÉPENER's ontology with SPARQL queries. In the next sections the four SEÍS services are described.

4.1. Examples of energy efficient buildings

Users of this service wish to explore cases of energy-efficient buildings which meet a particular design criteria. Firstly, users specify the city or postal code of the location of the building. The main use of the building (e.g., Residential or Office) also has to be specified. Afterwards, users specify the energy uses and performance indicators that are important in their context. A list of the buildings which meet the inputs from the users is retrieved from the dataset by submitting SPARQL queries to the endpoint. The energy-efficient buildings are visualized in a table showing the different performance indicators. The results can also be explored graphically, in a heat map implemented on top of Google Maps showing the energy efficiency concentrations. Once a building is selected, a report of its main attributes is shown to the users. The report is structured according to the main taxonomy of the RÉPENER's ontology.

4.2. Performance benchmarks

This service benchmarks the main performance indicators of the dataset of buildings before and after its refurbishment. The indicators included are: heating consumption (*repener:HeatingConsumption*), CO2 emissions (*repener:CO2emissions*), among others. Users provide the location and a main use to filter buildings included in the benchmark. The benchmark of the performance indicators is shown to the user in two separated columns, one for energy efficient buildings and another for the non-efficient buildings. Two values are displayed for each indicator, before and after the renovation of buildings. In addition, and as a way of providing more information, its percentage of improvement is shown. In this way, users can find out the common values of energy-efficient buildings and compare them with the ones that correspond to non-efficient buildings.

4.3. Energy efficient design patterns

The goal of this service is to identify the correlations between the design variables and the energy performance of energy-efficient buildings. The service recognizes the common properties of the buildings such as prevalent orientation of the window area (*repener:PrevalentOrientationOfWindowArea*), or solar contribution for hot water (*repener:SolarContributionForHotWater*). This kind of analysis helps the users to identify which design op-

[11] http://www4.wiwiss.fu-berlin.de/pubby/

[12] http://www.seis-system.org

tions would reduce the energy consumption in the case of refurbishment.

4.4. Enter a building simulation

This service is carried out by an energy consultant who uploads the data of a simulation to the system with the final goal of ranking and comparing it with the existing data in the dataset. The data provided by the users is entered following the ontology structure, ensuring the compatibility with the existing data. Once the data is uploaded, the system ranks the input building within the list of buildings. Furthermore, the service compares the input building with the benchmarks of energy efficient buildings and all buildings.

4.5. Example query

The dataset can be accessed directly, submitting SPARQL queries to the endpoint. The following query is an example of retrieving building properties from the dataset:

```
prefix repener:
<http://arcdev.housing.salle.url.edu/repener/lod/ontology/>
prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
prefix sumo: <http://www.ontologyportal.org/SUMO.owl#>
SELECT ?bid ?floorArea ?lat ?long ?primaryenergy ?station
FROM <http://arcdev.housing.salle.url.edu/repener/lod>
WHERE {
   []    repener:hasBuilding ?b;
         repener:value ?climatezone.
   FILTER (regex(?climatezone, "C2", "i")).
   ?b    a <http://www.ontologyportal.org/SUMO.owl#Building>;
         repener:hasProjectData [repener:hasBuildingLocation ?bl];
         repener:hasBuildingProperties
[repener:hasBuildingGeometry [repener:hasConditionedFloorArea
[repener:conditionedFloorAreaValue ?floorArea]]];
         sumo:hasPerformance [repener:hasEnergyPerfomance
[repener:hasTotalPrimaryEnergy
[repener:totalPrimaryEnergyValue ?primaryenergy ]]];
         repener:buildingId ?bid.
   ?bl   owl:sameAs ?c.
   ?c    geo:lat ?lat;
         geo:long ?long.
   OPTIONAL {
         ?c repener:closestStation ?station. }
    } order by ?primaryenergy
```

This query retrieves a list of buildings with some of their attributes. The properties are: the building ID, conditioned floor area, geographical coordinates of the location, the primary energy use of the building and the closest weather station using the links of the Aemet dataset. This last property is optional since not all of the building locations have a link to a weather station. The list is ordered by the primary energy use and filtered by the “C2” climate zone.

5. Related work

Recent projects such as Reegle[13] use Linked Open Data technologies to access energy-related data that has been obtained from open sources [11]. In the same line, the Open Energy Information (OpenEI) [14] online platform provides with free and open access to energy-related data, models, tools, and information which has been made available via Linked Open Data standards. With regard to these projects, the distinguished features of the dataset of RÉPENER are the scale and source of the data. While Reegle and OpenEI platforms offer energy-related data at a national level –policies, regulations, energy production or renewable resource– RÉPENER’s dataset collects data for specific buildings including physical characteristics, environmental characteristics, use profiles, and performance indicators from different phases of the building life-cycle.

6. Conclusions

In this paper we have presented a dataset which integrates data from different sources according to RÉPENER’s ontology. One of the difficulties has been to integrate various sources which use three different storage systems, including an obsolete Paradox database. A data integration process based on semantic technologies helped to overcome this problem. RÉPENER’s dataset has been linked to the datasets of Aemet and GeoLinkedData.es which cover the entire Spanish territory. The links connecting the different datasets enable the development of new services by third parties, such as a correlation analysis between building energy data, weather observations, and demographic data.

The main shortcoming of the dataset is its size which is relatively small (18962 entities at this moment) as compared to the average size of the Linked Data cloud (591632) [15]. In spite of these figures, this dataset is bound to grow for two reasons. Firstly, users can upload an energy simulation calculation to the SEÍS system. Secondly, a new law has been implemented which requires all existing buildings to have an energy certification. As a result of the application of this law, ICAEN has collected 50.000 new certifications so far, including those from new build-

[13] www.reegle.info
[14] www.openei.org
[15] http://stats.lod2.eu/stats

ing types which had not been previously considered such as office, commercial, educational, sports and trade facilities, among others. Even though the current coverage of the certifications source is restricted to Catalonia, its ultimate purpose is to encompass the whole of Spain.

To increase its visibility the RÉPENER dataset has been registered in DataHub.org, a metadata repository for Open Data which runs under the license of Creative Commons Attribution.

RÉPENER's dataset can contribute to the improvement of energy-efficient buildings, giving end-users the opportunity to make more informed decisions based onto the qualified data obtained from multiple sources they now have access to.

Acknowledgments

RÉPENER has been supported by the research program BIA 2009-13365, funded by the Spanish National RDI Plan 2009-2012. Fàtima Galán has participated in the ontology modelling and Marta Salgado in the implementation of the SEÍS system.

References

[1] L. Madrazo, Á. Sicilia, M. Massetti, and F. Galan. Semantic modelling of energy-related information throughout the whole building lifecycle. In Gudnason and Scherer, editors, *eWork and eBusiness in Architecture, Engineering and Construction*, pages 381–387. Taylor & Francis Group, London, 2012.

[2] IDAE, Energy consumption Analysis in the Spanish residential sector. Final Report. URL http://www.minetur.gob.es/energia/desarrollo/EficienciaEnergetica/CertificacionEnergetica/DocumentosReconocidos/OtrosDocumentos/Calificaci%C3%B3n%20energ%C3%A9tica.%20Viviendas/Cond_acept_anexos.pdf.

[3] ISO 13790:2008. Energy performance of buildings - Calculation of energy use for space heating and cooling.

[4] G. Nemirovskij, Á. Sicilia, F. Galan, M. Massetti, and L. Madrazo, Ontological Representation of Knowledge Related to Building Energy-efficiency. In D. Cacciagrano, editor, *Sixth International Conference on Advances in Semantic Processing (SEMAPRO 2012)*, pages 20–27. Curran Associates, Inc., 2012.

[5] V. Corrado, S.P. Corgnati, and M. Garbino. Energy Consumption Data Collection with DATAMINE. In C. M. Joppolo, editor, *Proceedings of the Energy, Climate and Indoor Comfort in Mediterranean Countries conference (Climamed 2007)*, pages 803–816. AICARR, 2007.

[6] I. Niles and A. Pease. Towards a Standard Upper Ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages: 2–9. ACM Press, 2001.

[7] C. Bizer and R. Cyganiak. D2RQ – Lessons learned. *Position paper at the W3C Workshop on RDF Access to Relational Databases 2007*. URL http://www.w3.org/2007/03/RdfRDB/papers/d2rq-positionpaper/

[8] L. Dodds and I. Davis. Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data. URL http://patterns.dataincubator.org/book/linked-data-patterns.pdf.

[9] J. Volz, C. Bizer, M. Gaedke and G. Kobilarov. Silk – A Link Discovery Framework for the Web of Data. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Kingsley Idehen, editors. *Proceedings of the 2nd Workshop about Linked Data on the Web (LDOW 2009)*, pages 1–6. M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen, 2009.

[10] A. De León, V. Saquicela, L. M. Vilches, B. Villazón-Terrazas, and F. Priyatna. Geographical linked data: a Spanish use case. In A. Paschke, N. Henze, and T. Pellegrini, editors, *6th International Conference on Semantic Systems (I-SEMANTICS '10)*, pages 1-3. ACM Press, 2010.

[11] F. Bauer, D. Recheis, and M. Kaltenböck. data.reegle.info – A New Key Portal for Open Energy Data. In Jiří Hřebíček, Gerald Schimak, and Ralf Denzer, editors, *9th IFIP WG 5.11 International Symposium (ISESS 2011)*, volume 359 of *IFIP Advances in Information and Communication Technology*, pages 189-194. Springer, 2011.