DOCTORAL THESIS

Title	**Supporting Tools for Automated Generation and Visual Editing of Relational-to-Ontology Mappings**
Presented by	**Álvaro Sicilia Gómez**
Center	**Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle**
Department	**ARC Architecture Representation Computation Research Group**
Directed by	**Dr. German Nemirovski** **Dr. Leandro Madrazo Agudín**

SUPPORTING TOOLS FOR AUTOMATED GENERATION AND VISUAL EDITING OF RELATIONAL-TO-ONTOLOGY MAPPINGS

per

Álvaro Sicilia Gómez

dirigida per

Dr. German Nemirovski
Dr. Leandro Madrazo Agudín

UNA TESI SOTMESA A ETSEEI LA SALLE – UNIVERSITAT RAMON LLULL
PER AL GRAU DE
PHILOSOPHIAE DOCTOR (PH.D.)
PROGRAMA OFICIAL DE DOCTORAT
TECNOLOGIES DE LA INFORMACIÓ I LA COMUNICACIÓ I LA SEVA GESTIÓ
DESEMBRE 2016

Acknowledgements

First of all, I would like to thank Leandro Madrazo for offering me the possibility to do research and German Nemirovski for awakening my interest in the Semantic Web. I am most grateful for their implication and scientific support in my professional career and in reaching this milestone. Thank you, Leandro and German, for your guidance on this research, for your patience with my initial inexperience on several aspects of the subject, and for your understanding in general.

Special recognition is due to my colleagues at the ARC Engineering and Architecture La Salle group at the Ramon Llull University. I especially want to thank those who have helped me to go through this research: to Gonçal Costa, for sharing the PhD journey; to Marta Salgado, Joan Pleguezuelos, and Eric Ortet, for offering technical suggestions with regard to software development.

As usual, parts of this thesis are based on previous publications. I would especially like to thank all of my colleagues in those various publications. Among them, special thanks to Andreas Nolle.

Mi más profundo agradecimiento a mi familia y amigos por su paciencia y ánimo en todos estos años. Gracias.

Per últim, a qui m'ha recolzat i ha patit la meva reclusió per llegir, programar, i escriure. Luisa, sense tu hagués sigut molt mes difícil. Gràcies.

Resum

La integració de dades amb formats heterogenis i de diversos dominis mitjançant tecnologies de la web semàntica permet solucionar la seva disparitat estructural i semàntica. L'accés a dades basat en ontologies (OBDA, en anglès) és una solució integral que es basa en l'ús d'ontologies com esquemes mediadors i el mapatge entre les dades i les ontologies per facilitar la consulta de les fonts de dades. No obstant això, una de les principals barreres que pot dificultar més l'adopció de OBDA és la manca d'eines per donar suport a la creació de mapatges entre dades i ontologies.

L'objectiu d'aquesta investigació ha estat desenvolupar noves eines que permetin als experts sense coneixements d'ontologies la creació de mapatges entre dades i ontologies. Amb aquesta finalitat, s'han dut a terme dues línies de treball: la generació automàtica de mapatges entre dades relacionals i ontologies i l'edició dels mapatges a través de la seva representació visual.

Les eines actualment disponibles per automatitzar la generació de mapatges estan lluny de proporcionar una solució completa, ja que es basen en els esquemes relacionals i amb prou feines tenen en compte els continguts de la font de dades relacional i les característiques de l'ontologia. No obstant això, les dades poden contenir relacions ocultes que poden ajudar a la generació de mapatges. Per superar aquesta limitació, hem desenvolupat AutoMap4OBDA, un sistema que genera automàticament mapatges R2RML a partir de l'anàlisi dels continguts de la font relacional i tenint en compte les característiques de l'ontologia. El sistema fa servir una tècnica d'aprenentatge d'ontologies per inferir jerarquies de classes, selecciona les mètriques de similitud de cadenes en base a les etiquetes de les ontologies i analitza les estructures de grafs per generar els mapatges a partir de l'estructura de l'ontologia.

La representació visual per mitjà d'interfícies intuïtives pot ajudar els usuaris sense coneixements tècnics a establir mapatges entre una font relacional i una ontologia. No obstant això, les eines existents per a l'edició visual de mapatges mostren algunes limitacions. En particular, la representació visual de mapatges no contempla les estructures de la font relacional i de l'ontologia de forma conjunta. Per superar aquest inconvenient, hem desenvolupat Map-On, un entorn visual web per a l'edició manual de mapatges.

AutoMap4OBDA ha demostrat que supera les prestacions de les solucions existents per a la generació de mapatges. Map-On s'ha aplicat en projectes d'investigació per verificar la seva eficàcia en la gestió de mapatges.

Paraules clau. OBDA, mapatges entre les dades i les ontologies, generació automàtica de mapatges, representació visual de mapatges, aprenentatge d'ontologies, R2RML.

Resumen

La integración de datos con formatos heterogéneos y de diversos dominios mediante tecnologías de la Web Semántica permite solventar su disparidad estructural y semántica. El acceso a datos basado en ontologías (OBDA, en inglés) es una solución integral que se basa en el uso de ontologías como esquemas mediadores y mapeos entre los datos y las ontologías para facilitar la consulta de las fuentes de datos. Sin embargo, una de las principales barreras que puede dificultar más la adopción de OBDA es la falta de herramientas para apoyar la creación de mapeos entre datos y ontologías.

El objetivo de esta investigación ha sido desarrollar nuevas herramientas que permitan a expertos sin conocimientos de ontologías la creación de mapeos entre datos y ontologías. Con este fin, se han llevado a cabo dos líneas de trabajo: la generación automática de mapeos entre datos relacionales y ontologías y la edición de los mapeos a través de su representación visual.

Las herramientas actualmente disponibles para automatizar la generación de mapeos están lejos de proporcionar una solución completa, ya que se basan en los esquemas relacionales y apenas tienen en cuenta los contenidos de la fuente de datos relacional y las características de la ontología. Sin embargo, los datos pueden contener relaciones ocultas que pueden ayudar a la generación de mapeos. Para superar esta limitación, hemos desarrollado AutoMap4OBDA, un sistema que genera automáticamente mapeos R2RML a partir del análisis de los contenidos de la fuente relacional y teniendo en cuenta las características de la ontología. El sistema emplea una técnica de aprendizaje de ontologías para inferir jerarquías de clases, selecciona las métricas de similitud de cadenas en base a las etiquetas de las ontologías y analiza las estructuras de grafos para generar los mapeos a partir de la estructura de la ontología.

La representación visual por medio de interfaces intuitivas puede ayudar a los usuarios sin conocimientos técnicos a establecer mapeos entre una fuente relacional y una ontología. Sin embargo, las herramientas existentes para la edición visual de mapeos muestran algunas limitaciones. En particular, la representación de mapeos no contempla las estructuras de la fuente relacional y de la ontología de forma conjunta. Para superar este inconveniente, hemos desarrollado Map-On, un entorno visual web para la edición manual de mapeos.

AutoMap4OBDA ha demostrado que supera las prestaciones de las soluciones existentes para la generación de mapeos. Map-On se ha aplicado en proyectos de investigación para verificar su eficacia en la gestión de mapeos.

Palabras clave. OBDA, mapeos entre datos y ontologías, generación automática de mapeos, representación visual de mapeos, aprendizaje de ontologías, R2RML.

Abstract

Integration of data from heterogeneous formats and domains based on Semantic Web technologies enables us to solve their structural and semantic heterogeneity. Ontology-based data access (OBDA) is a comprehensive solution which relies on the use of ontologies as mediator schemas and relational-to-ontology mappings to facilitate data source querying. However, one of the greatest obstacles in the adoption of OBDA is the lack of tools to support the creation of mappings between physically stored data and ontologies.

The objective of this research has been to develop new tools that allow non-ontology experts to create relational-to-ontology mappings. For this purpose, two lines of work have been carried out: the automated generation of relational-to-ontology mappings, and visual support for mapping editing.

The tools currently available to automate the generation of mappings are far from providing a complete solution, since they rely on relational schemas and barely take into account the contents of the relational data source and features of the ontology. However, the data may contain hidden relationships that can help in the process of mapping generation. To overcome this limitation, we have developed AutoMap4OBDA, a system that automatically generates R2RML mappings from the analysis of the contents of the relational source and takes into account the characteristics of ontology. The system employs an ontology learning technique to infer class hierarchies, selects the string similarity metric based on the labels of ontologies, and analyses the graph structures to generate the mappings from the structure of the ontology.

The visual representation through intuitive interfaces can help non-technical users to establish mappings between a relational source and an ontology. However, existing tools for visual editing of mappings show somewhat limitations. In particular, the visual representation of mapping does not embrace the structure of the relational source and the ontology at the same time. To overcome this problem, we have developed Map-On, a visual web environment for the manual editing of mappings.

AutoMap4OBDA has been shown to outperform existing solutions in the generation of mappings. Map-On has been applied in research projects to verify its effectiveness in managing mappings.

Keywords. OBDA, Relational-to-ontology mappings, automated generation of mappings, visual representation of mappings, ontology learning, R2RML.

Tesi doctoral per compendi de publicacions

La present tesi doctoral s'acull a la normativa per a l'elaboració de tesis doctorals per compendi de publicacions de la Universitat Ramon Llull[1]. La normativa consta dels següents punts:

1. Una tesi doctoral per compendi de publicacions estarà formada per un mínim de tres articles sobre una mateixa línia d'investigació.
2. Només s'acceptaran articles de publicacions que disposin d'un sistema d'avaluació per consemblants i/o que estiguin indexades preferentment en bases de dades científiques internacionals.
3. Només s'acceptaran articles publicats, o acceptats per a la seva publicació, realitzats amb data posterior a la primera matriculació del doctorand als estudis de doctorat o màster universitari.
4. Els coautors dels articles publicats donaran la seva conformitat per escrit a la utilització de l'article com a part de la tesi del doctorand.
5. Els coautors dels articles publicats no formaran part del tribunal de la tesi.
6. Els coautors dels articles publicats i utilitzats en una tesi que no tinguin el grau de doctor renunciaran per escrit a utilitzar l'article en una altra tesi. La Comissió Acadèmica del Programa de Doctorat podrà considerar excepcions justificades en l'aplicació d'aquesta norma, amb el vistiplau de la Comissió de Doctorat de la URL.
7. La tesi comptarà amb una introducció general que presenti els treballs publicats i la contribució específica del doctorand/a, una justificació de la unitat temàtica, una copia de cada treball publicat, un resum global dels resultats, la seva discussió i les conclusions finals.
8. Per tot això anterior, caldrà sempre, abans del dipòsit de la tesi, una presentació de sol·licitud formal a la Comissió Acadèmica del Programa de Doctorat i la seva l'acceptació favorable, la qual vetllarà per la qualitat de les publicacions que es volen presentar per a la tesi. A la dita sol·licitud s'afegirà també un informe del director de la tesi indicant quina és la contribució específica del doctorand al treball presentat i la de la resta d'autors, si s'escau.

Serà necessari presentar l'acta d'aprovació de la Comissió Acadèmica del Programa de Doctorat a la Comissió de Doctorat de la URL en el moment de la tramitació ordinària de la Tesi.

[1] Aprovada per la Junta Acadèmica a 18 de setembre de 2008 i actualització aprovada per la Comissió de Doctorat de la URL del 13/07/2016

PhD Thesis by publication

This thesis is regulated according to the policies for the preparation of doctoral theses compendium of works from the Universitat Ramon Llull[2]. The legislation consists of the following points:

1. A doctoral thesis by compendium of publications will consist of a minimum of three articles on the same line of research.
2. The articles will be accepted if they have a peer review assessment system and / or are preferably indexed international scientific databases.
3. Only published articles or accepted for publication will be accepted. The articles should be written after the first registration of the doctoral studies of doctorate or master official.
4. The co-authors of published articles should give their written consent to the use of the article as part of the doctoral thesis.
5. The co-authors of published articles cannot be part of the thesis jury.
6. Non PhD co-authors of published articles will provide a written renouncement to use these articles in another thesis. In case of published articles have been written by more than one research team, the Doctoral Committee may consider exceptions in the application of this regulation.
7. The thesis will include a general introduction to present the published works and the specific contribution of the doctoral student, a justification for the thematic research line, a copy of every published work, an overall summary of the results, discussion and conclusions.
8. For everything mentioned above, at the beginning of the process of the thesis, a formal request should be presented to the Academic Committee of the Doctoral Program and get their favorable acceptance. The Commission shall ensure the quality of the publications intend for the thesis. The formal request should include a report created by the Thesis director describing the specific contribution of the doctoral student, and if needed of the other co-authors, to the published articles.

It will be necessary to present the certificate of approval of the Academic Committee of the Doctoral Program at the Doctoral Committee of the URL when ordinary processing of the thesis.

[2] Approved by the Academic Board on 18 September 2008 and updated approved by the Doctoral Committee on 13 July 2016.

This thesis meets all the points previously mentioned. The publications that form the compendium are the following:

1. Madrazo, L., Massetti, M., Sicilia, Á., Wadel, G., & Ianni, M. (2015). SEíS: A semantic-based system for integrating building energy data. *Informes de La Construcción*, *67*(537). http://doi.org/10.3989/ic.13.048
2. Sicilia, Á., Nemirovski, G., Massetti, M., & Madrazo, L. (2015). The RÉPENER linked dataset. *Semantic Web*, *6*(2), 131–137. http://doi.org/10.3233/SW-130131
3. Madrazo, L., Sicilia, Á., & Nemirovski, G. (2013). Shared Vocabularies to Support the Creation of Energy Urban Systems Models. In *4th Workshop organised by the EEB data models community ICT for Sustainable Places* (pp. 130–150). Nice, France: Publications Office of the European Union. http://doi.org/10.2759/40897
4. Nemirovski, G., Nolle, A., Sicilia, Á., Ballarini, I., & Corado, V. (2013). Data integration driven ontology design, case study smart city. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* (p. 43-52). Madrid, Spain: ACM Press. http://doi.org/10.1145/2479787.2479830
5. Sicilia, Á., & Nemirovski, G. (2016). AutoMap4OBDA: Automated Generation of R2RML Mappings for OBDA. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings* (pp. 577–592). Bologna, Italy: Springer International Publishing. http://doi.org/10.1007/978-3-319-49004-5_37
6. Sicilia, Á., Nemirovski, G., & Nolle, A. (2016). Map-On: A web-based editor for visual ontology mapping. *Semantic Web Journal*, –in press. Retrieved from http://www.semantic-web-journal.net/content/map-web-based-editor-visual-ontology-mapping-0

Index

List of Figures

List of Listings

List of Tables

Preface

This thesis started in 2013 in the ARC Engineering and Architecture La Salle[3] group at the Ramon Llull University. ARC is a multidisciplinary research group dedicated to the design, development and application of information and communication technologies (ICT) in the architecture. ARC has carried out numerous research projects aiming at integrating data from different sources to improve the decision making of different stakeholders such as building designers, urban planners, owners and energy experts.

In the RÉPENER and SEMANCO research projects, ARC has developed energy information systems to integrate energy-related data from different domains stored in diverse formats. Overcoming structural and semantic heterogeneity of data sources has become a challenge that has been addressed by applying Semantic Web technologies. However, the use of these technologies requires a considerable amount of human resources. The current tools and methods of the state-of-the-art for semantic data integration do not reduce the burden of the human intervention in the development of information systems.

The research of this thesis started within these projects, in which the limitations and problems with regard to the application of Semantic Web technologies for data integration in the energy field were identified and partially solved. The motivation and goals of this thesis emerged from these projects initiating two research lines which are the main contribution of this thesis:

- Automated generation of relational-to-ontology mappings
- Visual support for relational-to-ontology mapping editing

Both research lines have produced tools and techniques that address the goals of the thesis. The technological outputs of this research have been the **AutoMap4OBDA** and **Map-On** tools which overcome some of the limitations of the related research works. The tools and the source code is available to the research community.

The relation between this thesis and the ARC research projects is illustrated in Figure 1. Four publications have been included in this thesis describing the construction process of the energy information systems of the RÉPENER and SEMANCO projects. Two publications have been included in this thesis, one for each research line, describing the contributions and the outcomes of this research. The tools developed in this thesis have been applied in the ENERSI project, a spin-off of the RÉPENER project.

[3] http://arc.salleurl.edu

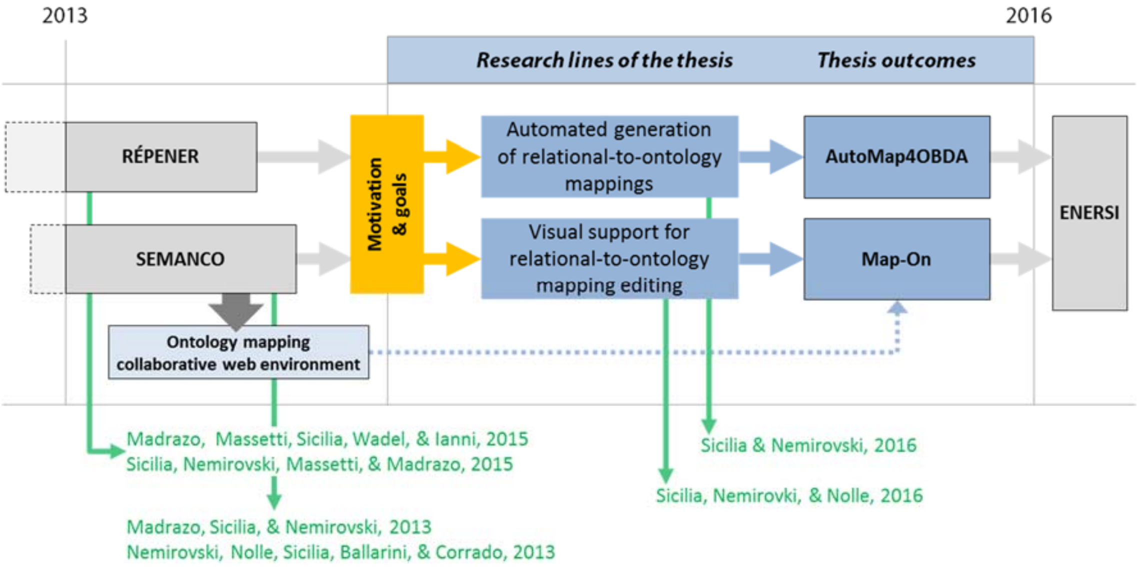

2013
2016
Research lines of the thesis
Thesis outcomes
RÉPENER
SEMANCO
Motivation & goals
Automated generation of relational-to-ontology mappings
Visual support for relational-to-ontology mapping editing
AutoMap4OBDA
Map-On
ENERSI
Ontology mapping collaborative web environment
Madrazo, Massetti, Sicilia, Wadel, & Ianni, 2015
Sicilia, Nemirovski, Massetti, & Madrazo, 2015
Madrazo, Sicilia, & Nemirovski, 2013
Nemirovski, Nolle, Sicilia, Ballarini, & Corrado, 2013
Sicilia & Nemirovski, 2016
Sicilia, Nemirovki, & Nolle, 2016

Figure 1. Research lines, outcomes and publications of this thesis in relation with ARC research projects.

This document is structured in the following chapters:

Chapter 1 Introduction presents the background and the motivation of this thesis. The energy information systems developed within the RÉPENER and SEMANCO projects are described. The goals and a summary of the contributions are presented.

Chapter 2 Concepts and Definitions provides the main concepts related to this research, such as Ontology-based data access (OBDA), relational database, ontology, and relational-to-ontology mappings. An illustrative example is described to show how a data source can be accessed using the OBDA techniques.

Chapter 3 Automated Generation of Relational-to-Ontology Mappings presents the work done in the automated generation of relational-to-ontology mappings. It introduces a formal description of an OBDA system. The main drawbacks of the current techniques for generating mappings between the data sources and the ontologies are identified. The techniques developed in the course of this research to overcome the limitations of state-of-the-art techniques are described. The implementation of the techniques are evaluated and compared with the current systems.

Chapter 4 Visual Support for Relational-to-Ontology Mapping Editing presents the work done with regard to the visual support for relational-to-ontology mapping editing. Tools for visual mapping and techniques for visual representing mappings between the data sources and the ontologies are introduced. The visual representation of mappings and the features of the mapping editor are described. A user study carried out to evaluate the usability of Map-On is introduced.

Chapter 5 Conclusion and Further Work discusses the results of the contributions of this research, draws the main conclusions, and presents further lines of research suggested by the results achieved so far.

The Appendix contains the abbreviations used in this document as well as the publications related to the research, and the contribution of the doctoral student.

1 Introduction

Developing information systems which integrate data from multiple sources involves challenges such as ensuring interoperability of systems by overcoming structural and semantic heterogeneity of data. Ontology-based data access, which relies on the use of ontologies as a mediator schemas and relational-to-ontology mappings to facilitate querying the data sources, is a comprehensive solution to address these challenges. The lack of tools to provide support for users in the creation and editing of relational-to-ontology mappings is one of the main barriers on the pathway towards developing semantics-based information systems. In this chapter the background and the motivation of this thesis – within the framework of Spanish and European research projects – are introduced. Finally, the goals and contributions of this thesis are presented.

1.1 Background

In recent years, the interdisciplinary character of numerous projects and applications has led to an increasing need for integrating data that is related to different knowledge domains, structured according to different schemas, and stored in different formats. In this context, the community of experts and stakeholders currently working with such heterogeneous data has grown considerably. The goal of integrating data is to have uniform access to a set of autonomous and heterogeneous data sources. Autonomous means that sources have usually been developed and maintained by different organizations. Heterogeneous sources are generally developed independently of each other; therefore, the data sources are stored in different systems and have different schemata even if they are representing the same domain (Doan, Halevy, & Ives, 2012).

Data integration processes must address system-level, logical and social challenges. The challenge at system-level is to ensure the interoperability of the distributed systems where the data is hosted. That is, data integration processes have to work with different formats, types of accessing and content syntax. The logical challenge has to do with how the data sources schemata have been designed. Schema of data sources are usually developed by different teams with their own design styles and view of the domain represented by the data source. Therefore, it is necessary to bridge the so-called semantic heterogeneity defined as the existence of disagreement about the meaning, interpretation, or intended use of the same type of data stored in diverse data sources' structures developed by different teams (Sheth & Larson, 1990). Social challenges emerge when data owners do not want to share their sources completely. Thus, legal issues might not allow access to particular data sources.

While social challenges can be solved by implementing non-technological approaches such as offering incentives to data owners to participate in the data integration process, the system-level and logical challenges require a technological solution. Semantic Web technologies, in particular the Ontology-based data access (OBDA) paradigm, can be useful in dealing with the interoperability of systems by solving semantic heterogeneity of data. In OBDA settings, the data sources are accessed using a high-level conceptual representation without the need to know how the data sources are actually organized (Calvanese et al., 2011; Poggi et al., 2008). Data queries, formulated in terms of their high-level conceptual representation, are rewritten with respect to the native data source schema and forwarded to the data source. The interoperability between systems is ensured by query rewriters which are specific tools to those systems. The semantic heterogeneity is alleviated by the use of a global representation that encompasses the different schemas of the data sources.

The conceptual representation of a domain can be realised by means of an ontology. According to Gruber (1993), an ontology is an explicit specification of a conceptualization. Where conceptualization is a simplified view of the world that is modelled for a particular purpose and the knowledge that one might want to model should be explicitly specified by means of concepts and relations formally coded in a particular language such as Ontology Web Language (OWL). Additionally, Borst (1997) defined ontology as a "formal specification of a shared conceptualization" pointing out that an ontology is a collective knowledge construction process in which various experts bring their vision and understanding of a particular domain.

In an OBDA solution, the main components are a data source, which contains data; an ontology, which represents a shared conceptualization of a domain; mappings between the data source and the ontology; and the query rewriter, which transforms queries initially referring the ontology concepts into an understandable form for the native system where the data is stored.

In this context, the development of mappings between the data source and the ontology is one of the key issues. Nowadays, manual development of such mappings is the widely adopted solution in academic and industry communities in spite of being extremely time-consuming and requiring high levels of human expertise (Savo et al., 2010). The development of relational-to-ontology mappings is a process that requires knowledge about a specific domain and technical skills in areas such as Entity Relationships modelling and ontology design. Finding users with this profile is difficult. Participation of domain experts and data owners, who usually are lacking the mentioned expertise, is significant. Therefore, the challenges involved in the manual creation of mappings represent still an important barrier in the adoption of the OBDA approach (Pinkel, Binnig, Kharlamov, & Haase, 2013).

1.2 Motivation: Semantic Energy Information Systems

Integration of heterogeneous data from different domains is one of the most challenging areas in the field of energy efficiency in buildings. Clearly, there is a need to have integrated access to energy-related data at the different stages of the building life-cycle in order to better understand the relationship between design and operation, in other words, between the initial design objectives and the actual performance of the building. In fact, having access to accurate information on request has become crucial for stakeholders involved in the improvement of the energy performance of buildings. Having access to this information may help in the design of new buildings, in the renovation of existing ones, and in the adjustment of building energy management systems.

Developing energy information systems – which integrate energy-related data – needs to address the same challenges described above. Indeed, in the last years, the development of new energy information systems has started to exploit Semantic Web technologies. Prime examples are the RÉPENER and SEMANCO research projects led by the group ARC Engineering and Architecture La Salle[4] at the Ramon Llull University.

1.2.1 The RÉPENER Project

The goal of RÉPENER[5], a project co-funded by the Spanish National R&D Plan 2010-2013, has been to design and implement an information system prototype which provides access to energy information using Semantic Web technologies. The result of the project has been SEíS[6], a semantic energy information system which integrates energy-related data of buildings and services to provide quality, accurate information to improve the decision making of different stakeholders such as building designers, facility managers, owners and energy experts (Madrazo, Massetti, Sicilia, Wadel, & Ianni, 2015). The services include the search for examples of energy efficient buildings, calculating building performance

[4] http://arc.salleurl.edu

[5] http://arc.salleurl.edu/repener/?lang=en

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

benchmarks, finding energy efficient design patterns, and comparing the energy performance of a building with the existing data. The SEíS services use an energy model formalised as an ontology to access the different data sources (e.g., energy certificates of buildings, energy certificates, building descriptions, simulation outcomes, energy monitoring, and climate data).

The SEíS energy model (i.e., an OWL ontology) has been developed based on existing energy information standards which encompass the building data as well as the contextual data – climate, economic and social – all of which impact buildings' energy efficiency. The ontology creation process involved energy-domain experts and ontology engineers. The collaborative process took into account the usage of standardised terminologies used in the energy domain and of certain agreed definitions that facilitate the understanding of the vocabulary among users. Standard definitions proposed by previous research projects like Datamine (Corrado, Corgnati, & Garbino, 2007) and by ISO and CEN standards such as ISO 13790:2008 (ISO, 2008) were used in the design of the SEíS energy model. The result has been an ontology that describes entities and relations, data types and units from all data sources (Nemirovski, Sicilia, Galán, Massetti, & Madrazo, 2012).

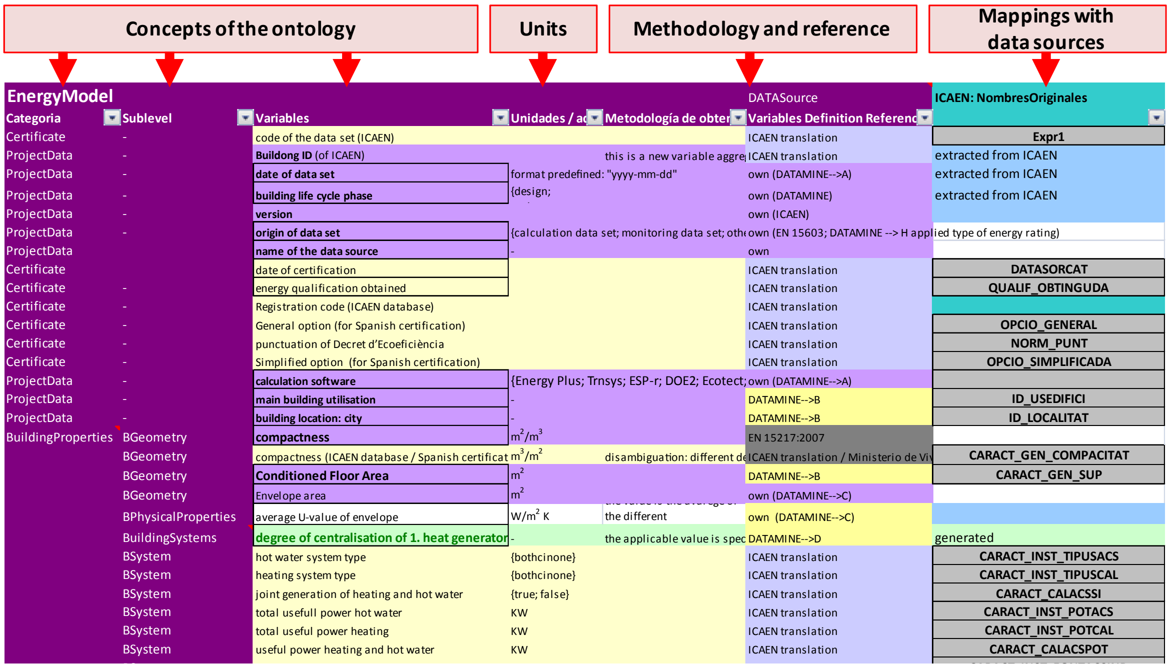
The SEíS system integrates three data sources: energy certifications provided by the Catalan Institute for Energy (ICAEN), consumption data facilitated by Leako – a company from the Basque Country dedicated to the installation, distribution and control of HVAC (Heating, Ventilation, and Air Conditioning) systems– and geographic information from the Geographical Information National Institute (CNIG). The data sources are heterogeneous in terms of domains (e.g., energy, geography, and climate) and they are provided in different formats: ICAEN data as Microsoft Excel documents, Leako's as a Paradox database, and the CNIG data as a Microsoft Access database. An ETL (Extract, Transform and Load) process has been devised to integrate the data sources using the energy model as a mediation schema. The sources have been manually mapped to the ontology using D2RQ language (Bizer & Cyganiak, 2007). Once the mappings are obtained they can be used to transform the sources into RDF (Resource Description Framework). The integration process followed the best practices in URI (Uniform Resource Identifier) design (Dodds & Davis, 2012). Data of SEíS system were linked to external open data sources such as GeoLinkedData[7] and Aemet[8] datasets. The result of the data integration process has been a linked dataset made public following the Linked Open Data principles. A comprehensive description of the data integration process can be found in (Sicilia, Nemirovski, Massetti, & Madrazo, 2015).

Considerable efforts of the data integration process have been dedicated to the development of the energy model and to the creation of mappings between the sources and the energy model. An Excel document has been used as tool to collect the knowledge from the domain experts in order unify the terms and identify relationships between them and

[7] https://datahub.io/dataset/geolinkeddata

[8] https://datahub.io/dataset/aemet

the data sources. The creation of such table has been necessary because of the lack of easy-to-use tools that enable non-ontology experts to participate in the ontology design process (Figure 2). The existing relational-to-ontology mapping tools were not mature enough when the SEíS system was being developed. Therefore, the mappings were created manually through the collaboration of domain experts and ontology engineers.



EnergyModel					DATASource	ICAEN: NombresOriginales
Categoria	Sublevel	Variables	Unidades / ac	Metodología de obter	Variables Definition Referenc	
Certificate	-	code of the data set (ICAEN)			ICAEN translation	Expr1
ProjectData	-	**Buildong ID** (of ICAEN)		this is a new variable aggre	ICAEN translation	extracted from ICAEN
ProjectData	-	**date of data set**	format predefined: "yyyy-mm-dd"		own (DATAMINE-->A)	extracted from ICAEN
ProjectData	-	**building life cycle phase**	{design;		own (DATAMINE)	extracted from ICAEN
ProjectData	-	**version**			own (ICAEN)	
ProjectData	-	**origin of data set**	{calculation data set; monitoring data set; othe		own (EN 15603; DATAMINE --> H applied type of energy rating)	
ProjectData		**name of the data source**	-		own	
Certificate		date of certification			ICAEN translation	DATASORCAT
Certificate	-	energy qualification obtained			ICAEN translation	QUALIF_OBTINGUDA
Certificate	-	Registration code (ICAEN database)			ICAEN translation	
Certificate	-	General option (for Spanish certification)			ICAEN translation	OPCIO_GENERAL
Certificate	-	punctuation of Decret d'Ecoeficiència			ICAEN translation	NORM_PUNT
Certificate	-	Simplified option (for Spanish certification)			ICAEN translation	OPCIO_SIMPLIFICADA
ProjectData	-	**calculation software**	{Energy Plus; Trnsys; ESP-r; DOE2; Ecotect;		own (DATAMINE-->A)	
ProjectData	-	**main building utilisation**	-		DATAMINE-->B	ID_USEDIFICI
ProjectData	-	**building location: city**	-		DATAMINE-->B	ID_LOCALITAT
BuildingProperties	BGeometry	**compactness**	m^2/m^3		EN 15217:2007	
	BGeometry	compactness (ICAEN database / Spanish certificat	m^3/m^2	disambiguation: different de	ICAEN translation / Ministerio de Viv	CARACT_GEN_COMPACITAT
	BGeometry	**Conditioned Floor Area**	m^2		DATAMINE-->B	CARACT_GEN_SUP
	BGeometry	Envelope area	m^2		own (DATAMINE-->C)	
	BPhysicalProperties	average U-value of envelope	W/m^2 K	the different	own (DATAMINE-->C)	
	BuildingSystems	**degree of centralisation of 1. heat generator**	-	the applicable value is spec	DATAMINE-->D	generated
	BSystem	hot water system type	{bothcinone}		ICAEN translation	CARACT_INST_TIPUSACS
	BSystem	heating system type	{bothcinone}		ICAEN translation	CARACT_INST_TIPUSCAL
	BSystem	joint generation of heating and hot water	{true; false}		ICAEN translation	CARACT_CALACSSI
	BSystem	total usefull power hot water	KW		ICAEN translation	CARACT_INST_POTACS
	BSystem	total useful power heating	KW		ICAEN translation	CARACT_INST_POTCAL
	BSystem	useful power heating and hot water	KW		ICAEN translation	CARACT_CALACSPOT

Figure 2. Excel sheet to identify concepts of the RÉPENER ontology and mappings with the data sources.

The results of the research carried out within the RÉPENER project have been published at:

- Madrazo, L., Massetti, M., Sicilia, Á., Wadel, G., & Ianni, M. (2015). SEíS: A semantic-based system for integrating building energy data. *Informes de La Construcción*, *67*(537). http://doi.org/10.3989/ic.13.048
- Sicilia, Á., Nemirovski, G., Massetti, M., & Madrazo, L. (2015). The RÉPENER linked dataset. *Semantic Web*, *6*(2), 131–137. http://doi.org/10.3233/SW-130131

1.2.2 The SEMANCO Project

The purpose of the SEMANCO[9] project – co-funded by the European Commission within the 7th Framework Programme 2011-2015 – was to provide an energy information system to help different stakeholders involved in urban planning (architects, engineers, building managers, local administrators, citizens and policy makers) to make informed decisions about how to reduce carbon emissions in cities. Unlike the RÉPENER information system, the SEMANCO system is not limited to a single building but it deals with groups of buildings within an urban area (Madrazo, Sicilia, & Nemirovski, 2013).

[9] http://semanco-project.eu

The energy information system uses Semantic Web technologies to integrate heterogeneous data distributed in different data sources with a variety of tools to support decision making which operate on the integrated data. The key component of the system is the Semantic Energy Information Framework (SEIF). This framework is the nexus between the different data sources – that employ diverse structural schemas, access methods and data semantics – and the energy analysis and simulation tools (Figure 3). The SEIF is an OBDA system which encompasses an ontology (Nemirovski, Nolle, Sicilia, Ballarini, & Corado, 2013), a set of relational-to-ontology mappings, and the federation engine ELITE (Nolle & Nemirovski, 2013) to unify the access to distributed relational data sources facilitated by the pilot cities (Manresa, Spain; Newcastle, UK; and Copenhagen, Denmark). All of these components have been developed within the SEMANCO project.

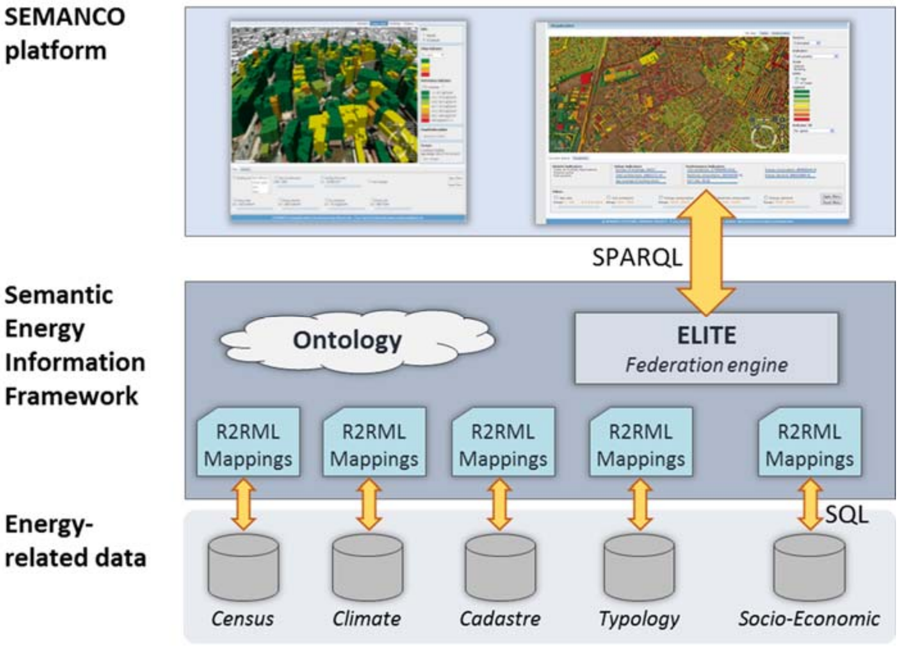


Figure 3. Energy information system developed in the SEMANCO project.

A methodology has been devised to elicit domain experts' knowledge with the purpose of creating the SEIF. This methodology embraces three main processes: use case definition, ontology building, and semantic data integration.

Use case definition. A set of use cases has been textually described by domains experts with the purpose of capturing the relations between actors, tools, and data which are set in interaction to fulfil a specific goal concerning to carbon reduction in a given urban area. The compound of interrelated actors, tools and data makes a use case. The activities encompassed by a use case are described using the Neon methodology for ontology engineering (Suárez-Figueroa, Gómez-Pérez, Motta, & Gangemi, 2012), as a set of requirements and competency questions.

Ontology building. In the ontology building process the knowledge encapsulated in the use cases is formalised as an ontology. As a first step in this process, domain experts and ontology engineers jointly created a vocabulary of terms and definitions which were compiled in the Energy Standard Tables (Corrado, Ballarini, Madrazo, & Nemirovskij, 2015). They are similar to the tables created within RÉPENER project but unlike them they include subsumption and aggregation relations among concepts. Finally, the Energy Standard Tables were formalised as an ontology using OWL.

Semantic data integration. The semantic data integration process takes the data sources identified in the use case definition process to translate them into RDF according to the ontology created in the ontology building process. The translation is guided by a set of mappings manually defined in a spread sheet by the domain experts who are familiar with the data sources and the ontology. Those mappings are coded using the declarative language R2RML (RDB to RDF Mapping Language)[10]. The SEIF receives queries from the external tools – using the terms of the ontology – and translates them into SQL using the R2RML mappings using Elite, a federation engine. The SEIF assures the interoperability between the data and the tools that use the data. The relation between data and tools is also handled by the SEIF (Sicilia, Madrazo, & Pleguezuelos, 2015).

With respect to the RÉPENER project, SEMANCO took a step forward in allowing domain experts, data owners and ontology engineers to integrate data sources using OBDA paradigm. To support the participation of different kinds of users two tools have been developed: Click-On, an ontology editor which hides the complexity of coding ontologies (Wolters, Nemirovski, & Nolle, 2013) and the Ontology Mapping Collaborative Web Environment which provides a visual interface to assist non-ontology experts in coding relational-to-ontology mappings (Madrazo et al., 2013). This tool reduces users' efforts in creating those mappings. However, the tool had some drawbacks which became evident as it was implemented during the project. It simplifies the manual creation of R2RML mappings but it does not automatically locate mappings between the elements of a relational source and the elements of the ontology. Thus, it does not offer a visual representation of the mappings.

The outcomes of the research carried out within the SEMANCO project have been presented at:

- Madrazo, L., Sicilia, Á., & Nemirovski, G. (2013). Shared Vocabularies to Support the Creation of Energy Urban Systems Models. In *4th Workshop organised by the EEB data models community ICT for Sustainable Places* (pp. 130–150). Nice, France: Publications Office of the European Union. http://doi.org/10.2759/40897
- Nemirovski, G., Nolle, A., Sicilia, Á., Ballarini, I., & Corado, V. (2013). Data integration driven ontology design, case study smart city. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* (p. 43-52). Madrid, Spain: ACM Press. http://doi.org/10.1145/2479787.2479830

[10] https://www.w3.org/TR/r2rml

1.3 Goals and Contributions

The objectives of the research presented in the following sections emerged from the development of energy information systems in RÉPENER and SEMANCO projects. In both projects, heterogeneous data from different energy-related domains have been integrated using the OBDA approach. The manual creation of relational-to-ontology mapping required a considerable amount of human resources, which led to the following problems:

- Considerable amounts of mappings had to be manually created from scratch despite the fact that the creation of some of them could have been automated based on the contents (i.e., data instances) of the data source.
- In this manual creation process some errors were introduced which required a lot of resources to detect and fix.
- Editing and maintenance of the mappings were carried out as modifications of text files by means of a text editor. This hindered the participation of people without ontology engineering skills. Moreover, it made the modification of the mappings and the error detection difficult.

These problems could have been solved if automated tools for finding and visualizing relations between the elements of a relational source and an ontology had been available. Therefore, we have followed the "divide and conquer" strategy to decompose the main problem which is the generation of relational-to-ontology mappings in an OBDA context into two subproblems or lines of work:

1) Automated generation of relational-to-ontology mappings. Proposing solid techniques for the automated generation of mappings between a relational source and an already existing ontology. Current relational-to-ontology mapping generators are far from solving real-world scenarios. Indeed, current mapping generators basically rely on the relational schema and do not fully take into account the contents of the relational source and the features of the ontology, i.e. the one to be mapped onto the relational source. The work conducted in this research contributes to overcome these limitations by:

- Developing innovative techniques for automatically extracting relational-to-ontology mappings between a relational data source and an existing ontology based on the intensive use of relational source contents and features of the ontology. These techniques infer class hierarchies from the contents of the relational source using an ontology learning technique, select the proper string similarity metric based on the ontology labels, and generate the mappings based graph structures.
- Implementing the new techniques in **AutoMap4OBDA**, a system which automatically generates R2RML mappings based on the intensive use of relational source contents and features of the ontology.

2) Visual support for relational-to-ontology mapping editing. Proposing visual representations and techniques for providing support users in the editing of mappings. Creating R2RML mappings requires advanced skills and expertise in ontology design and formal logic. For domain experts and data owners, who usually lack the aforementioned expertise, the main barrier is often the lack of a visual representation of the mappings. In practice, a visual representation could help them overcome the lack of expertise and complete the mapping task. The contributions to this line of work have been:

- A proposal of a visual representation of relational-to-ontology mappings which considers the structure of the relational source, the ontology, and the relation between them. The visual representation is based on a graph layout which is probably the most typical and the most commonly used form of ontology and mapping visualization.
- An implementation of the new visual representation in **Map-On**, a graphical web environment for ontology mapping which supports different kinds of users to manually establish relations between the elements of a relational source and an ontology in the context of an OBDA scenario. Usability studies have been carried out to validate the benefits provided by Map-On in the mapping editing process.

The scientific contributions of the doctoral student to the published works are described in the Appendix C.

2 Concepts and Definitions

The goal of this chapter is to introduce the main terminology used in this thesis such as the Ontology-based data access concept and its components: relational database, ontology, and relational-to-ontology mappings. Moreover, the languages to represent and query ontologies are presented. Thus, an illustrative example – taken from the research projects described in *Section 1.1 Motivation* – is given to show how the concepts are used and interrelated.

2.1 Ontology-Based Data Access

The term *Ontology-based data access* was introduced to describe the application of Semantic Web technologies in data integration systems (Calvanese et al., 2007). The main purpose of OBDA is to provide access to the data layer of an information system by means of queries over a domain specific conceptual layer rather than through direct access to an information system. The conceptual layer hides how the data is stored in the system. The typical candidate for the management of the data layer is a relational database management system and the candidate for implementing the conceptual layer is an ontology. The use of an ontology as a conceptual layer enables reasoning through the ontology to unveil relations hidden in the data layer. Thus, the domain knowledge coded in the ontology offers the possibility for automated query optimization. Similar terms used in literature are *accessing data mediated by an ontology*, *ontology-driven information systems*, and *ontology-based data management*.

The specification of a system that uses OBDA is a triple *<O, S, M>*, where *O* is an ontology, providing a conceptual specification of the domain of interest, *S* is an schema of a set of relational data sources, and *M* is a set of mapping assertions that describe the relation between the ontology and the data sources by means of queries through the ontology that are rewritten in terms of the data sources.

In OBDA, the relational data sources and ontologies are designed and developed independently. While data sources are created by an information technology team with the purpose of increasing the performance of the storage system, ontologies are created by a community of experts or by a standardization body with the purpose of reaching an agreement of a specific view of a domain of interest. In this document the terms target ontology and domain ontology are used indistinctly to refer to those ontologies which are used in an OBDA system as a conceptual layer.

2.2 Relational Database

The main purpose of a relational database is to manipulate data and the relations among those data (Abiteboul, Hull, & Vianu, 1995). That is, data is grouped in relations *R* (i.e., tables) of tuples (i.e., table row), in which each tuple is composed of attributes (i.e., columns). The attributes are defined by a name and a set of permitted values D_i for a particular domain. This way, a relation is a set of n-tuples where each tuple has the same type of attributes. The formal description of a relation considers it a subset of the Cartesian product of all attributes, in other words, it is the product of all possible n-tuples. The data stored in relations are uniquely identified and linked to related data through attributes

$$R \sqsubseteq D_1\, x\, D_2\, x\, \ldots\, x\, D_{n-1}\, x\, D_n$$

A relational database is a tuple $< \mathcal{R}, \Sigma >$, where $\mathcal{R}$ is a set of relations, and Σ is a set of integrity constraints. For example:

- A primary key is an attribute (or combination of attributes) of a relation which values uniquely identifies each n-tuple of that relation.
- A foreign key is an attribute (or attribute combination) of relation *R* that is not the primary key of *R* but its elements are values of the primary key of some relation *S*.
- Uniqueness constraint is an attribute which is a sequence of distinct values in *R*.

2.3 Knowledge Representation: Ontologies

An ontology is “a formal specification of a shared conceptualization” (Borst, 1997). The formal specification is created through concepts, attributes, values, relationships, roles and rules that describe a domain. The term ‘shared conceptualization’ indicates reaching a consensus among experts whereby the conceptualization represents the related knowledge domain. Ontologies are formed by concepts which are sets, collections, types of objects or kinds of things while the attributes are aspects, properties, features, and characteristics that an object can have. Concepts are related to each other by means of roles. An ontology specifies the items of a knowledge domain by means of axioms which are assertions and rules.

There are several classifications of ontologies. One of them distinguishes between the general ontologies (conceptualizing time, space and events) and domain ontologies (focusing on the resolution of a specific problem), for example conceptualizing academic records and documentation. The use of ontologies for representing knowledge enables the application of inference mechanisms aimed at discovering knowledge not previously established.

The ontology engineering community has developed different standard languages for knowledge representation. Web Ontology Language (OWL) (W3C, 2009) and Resource Description Language (RDF) (W3C, 2014) are two well established languages created by the World Wide Web Consortium with the purpose of describing the ontologies in a formal way. The SPARQL Protocol and RDF Query Language (SPARQL) (W3C, 2013) is the most used query language by the community, and the facto standard. This formalization brings the capability of being processed by computers allowing the inference, reasoning and bridging between ontologies.

A *putative ontology* is an ontology that have been automatically generated from a database schema using reverse engineering methods (Juan F. Sequeda, Tirmizi, & Miranker, 2008). In this document the term putative ontology is used to refer those ontologies that have been automatically generated from a relational source using transformation rules. Some authors name this kind of ontologies *bootstrapped ontologies*.

2.3.1 Web Ontology Language

OWL is the acronym of The Web Ontology Language, a set of languages that have been created for computers to process the content of information. OWL can be used to represent a particular view of the world using vocabularies of terms and relationships between those terms. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema by providing additional vocabulary along with formal semantics.

The languages in the OWL family use the open world assumption. Under this assumption, if a statement cannot be proven to be true with the available knowledge, then the conclusion that the statement is false cannot be drawn.

2.3.2 Resource Description Framework

Resource Description Framework (RDF) is a specification provided by the World Wide Web Consortium (W3C). The main goal of RDF is to provide a metadata data model to represent statements in the form of subject-predicate-object expressions. The subject and the object indicates a resource, and the predicate denotes a relationship between the subject and the object. The union of a subject, predicate, and an object is called triple. The subject of a triple is an Internationalized Resource Identifier (IRI) which is a string of characters used to identify a resource. The object of a triple can be an IRI and a literal which is composed of a lexical form (i.e., a Unicode string), a datatype IRI, and optionally a non-empty language tag.

In these terms, a particular view of the world is represented with a list of triples describing its characteristics. A collection of RDF triples inherently represents a labelled, directed multi graph. Despite this, RDF data often stored in relational databases are expected to perform better than dedicated stores such as triple stores or quad stores which also store the context of the triples. RDF data can be serialised in several formats such as: Turtle[11], N-Triples[12], N-Quads[13], JSON-LD[14], Notation3[15], and RDF/XML[16].

2.3.3 SPARQL Query Language

SPARQL is a specification proposed by W3C aimed at querying databases which contain data in RDF format. SPARQL is a language based on RDF and triple patterns which have to meet the output data. SPARQL has four query forms: SELECT (to return data that match a given triple pattern), CONSTRUCT (to generate a RDF graph), ASK (to return a Boolean if there is data or not), and DESCRIBE (to return a RDF graph describing a resource). The first part of the SELECT form specifies the variables which are going to be retrieved. Then, in the clause WHERE is defined the triple query which can contain static values or variables. Optionally an ORDER BY or FILTER clauses can be added too. The Listing 1 shows an example of a SPARQL query.

```
SELECT ?name ?email
WHERE {
    ?person rdf:type foaf:Person.
    ?person foaf:name ?name.
    ?person foaf:mbox ?email.
}
ORDER BY ?name
```

Listing 1. An example of a SPARQL query.

In the example of the Listing 1, the query requests all triples which meet the given pattern. The output of this query is all the triples which contain a subject which is a person and have a name and a mail box.

2.4 Relational-to-Ontology Mappings

Relational-to-ontology mappings describe how to transform instances of a database into instances of an ontology. A mapping M is a set of expressions of the form $Q \rightsquigarrow E$, where Q is an SQL query over a schema S and E is an element of the ontology O. Mappings are formalized in a declarative language such as R2RML (RDB to RDF Mapping Language)

[11] http://www.w3.org/TR/turtle/

[12] http://www.w3.org/TR/n-triples/

[13] http://sw.deri.org/2008/07/n-quads/

[14] http://www.w3.org/TR/json-ld/

[15] https://www.w3.org/TeamSubmission/n3/

[16] http://www.w3.org/TR/rdf-syntax-grammar/

which is a declarative language recommended by the W3C for expressing customized mappings from relational databases to RDF datasets according to an ontology (W3C, 2012).

An R2RML mapping is composed of a logic table which can be defined as a base table, a view (i.e. the result set of a stored query), or a SQL query. Thus, for instance, a mapping can relate instances retrieved from the logic table of the database with a class of an ontology, attributes of the relational table with data type properties, and relations between database instances to object properties of an ontology. In terms of R2RML, a TriplesMap is declared as a subject map described with an Internationalized Resource Identifier (IRI) generated from the logic table. Moreover, the data-to-object mappings are declared as predicate and object maps. The subject and objects maps describe how the IRIs should be generated using the columns specified in the logic table and the elements of an ontology. The Figure 4 shows the elements of a R2RML statement.

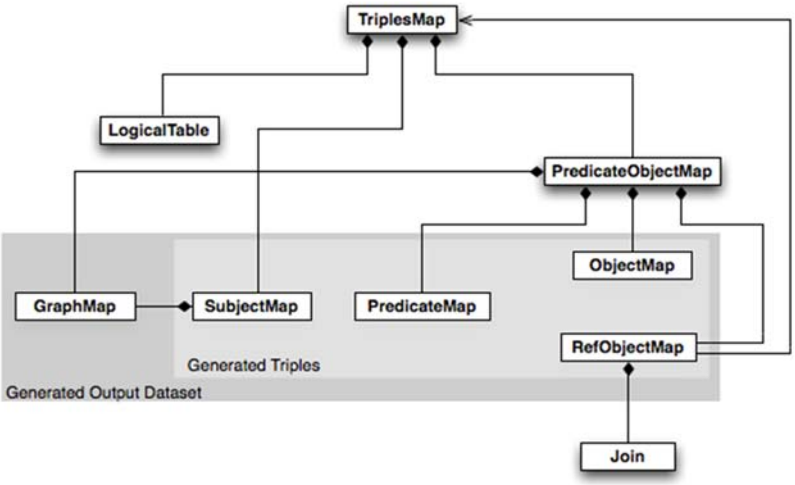


Figure 4. An overview of R2RML statement[17].

The triples map illustrated in Listing 2 uses a logic table based on a SQL query for a table called *Buildings*. The IRI of the subject map uses the *ID* column of the table and a concept called *Building* from the ontology. The object map is defined with the relation *hasAddress* and the column *Address* from the ontology.

```
<mapping1> a rr:TriplesMap;
  rr:logicalTable [
    rr:sqlQuery "SELECT ID, Address FROM Buildings"];
  rr:subjectMap [
    rr:template "http://example.com/building/{ID}";
    rr:class ex:Building];
  rr:predicateObjectMap [
    rr:predicate ex:hasAdress;
    rr:objectMap [rr:column "Address"]].
```

Listing 2. An example of an R2RML mapping.

Generating R2RML mappings between a database and an ontology conveys two main tasks: first, finding the correspondences between the elements of the database and the

[17] Figure taken from https://www.w3.org/TR/r2rml/

ontology (e.g., table *Buildings* relates to class *Building*) and second, obtaining the SQL views needed to generate the IRIs of the subject and predicate object maps. The creation of R2RML mappings requires technical skills in both SQL query design and in ontology engineering. The experts who create the mappings should understand the structure of the relational schemas and the ontology in order to find correspondences between the columns of the relational tables and the ontology entities. Moreover, users have to design SQL queries for the logic tables and the IRI patterns for the subject and object maps.

2.5 Illustrative Example

This section describes an example of how a data source of the RÉPENER project –the energy certifications provided by the ICAEN– can be accessed using OBDA techniques.

The ICAEN data source is structured as a single table which contains 35 columns and 1804 rows. Each row is an energy certification of a building. The energy certification can be performed during different phases of the life cycle of the building such as the design or operational phase. Each energy certification contains the energy rating of the building, different energy consumptions, the types of HVAC (Heating, Ventilation, and Air Conditioning) systems, and geometric features such as the built surface or the compactness (a ratio between the surface of a building and its volume). Table *1* specifies selected columns of the ICAEN table.

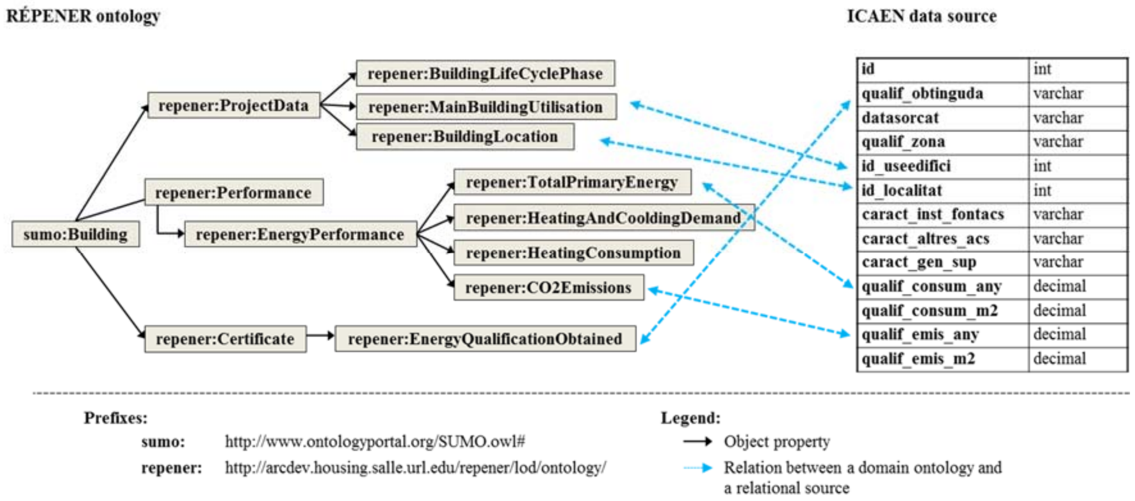
Table 1. *ICAEN data source structure: column names and descriptions*

Column name	Description
id	Id of the energy certification
qualif_obtinguda	Energy rating
datasorcat	Date when the certification was performed
qualif_zona	Climate zone where the building is located
id_useedifici	Building use
id_localitat	Name of the city or town where the building is located
caract_inst_fontacs	Energy carrier for domestic hot water
caract_altres_acs	Solar energy contribution to domestic hot water
caract_gen_sup	Built surface
qualif_consum_any	Yearly energy consumption
qualif_consum_m2	Yearly energy consumption per m_2
qualif_emis_any	CO2 yearly emissions.
qualif_emis_m2	CO_2 yearly emissions per m_2

The ontology selected is the energy model (Nemirovski et al., 2012) which has been designed during the RÉPENER project whose domain is energy performance of buildings. This ontology has been developed taking into account existing energy standards such as the energy certification of buildings as defined by the DATAMINE project (Corrado et al., 2007), the ISO and CEN standards following the European Directive 2002/91/EC (e.g., ISO 13790:2008) and the Standard Network Variable Types from LonWorks. These standards cover some areas of the ontology core which is defined as follows:

- **General project data**: Project descriptions which include its generic characteristics such as location, use, project execution data, and site description. Some examples of ontology classes are: *repener:MainBuildingUtilisation*, *repener:BuildingOwner*, and *repener:BuildingLocation*.
- **Performance**: Indicators regarding energy use such as energy demands or consumption of different energy carriers, CO_2 emissions and indoor conditions such as temperature and humidity. Examples of classes are: *repener:TotalPrimaryEnergy* and *repener:CO2emissions*.
- **Building properties**: Parameters which describe geometric characteristics, construction systems and building services. Examples of classes are: *repener:BuildingGeometry*, *repener:EnvelopeArea*, and *repener:ActiveSystems*.
- **Outdoor environment**: Climate characteristics and conditions of the physical environment such as outdoor temperature, wind speed and direction, and solar radiation. Examples of classes are: *repener:Climate* and *repener:ClimateZone*.
- **Operation**: Parameters regarding the usage and management of the building and its systems for maintaining comfort levels such as solar protection or thermostat regulation. Examples of classes are: *repener:PresenceTimePerDay* and *repener:TypicalHeatingSetPointTemperature*.
- **Certification**: It includes indicators to qualify a building based on performance such as the energy efficiency rating according to a conventional Spanish scale of (A, B, C, etc.). It also includes the certification-process methodology. Examples of classes are: *repener:EnergyQualificationObtained* and *repener:DateOfCertification*.

A set of R2RML mappings have been created to transform the contents of the ICAEN database into RDF according to the RÉPENER ontology (Figure 5).


RÉPENER ontology
ICAEN data source
sumo:Building
repener:ProjectData
repener:BuildingLifeCyclePhase
repener:MainBuildingUtilisation
repener:BuildingLocation
repener:Performance
repener:EnergyPerformance
repener:TotalPrimaryEnergy
repener:HeatingAndCooldingDemand
repener:HeatingConsumption
repener:CO2Emissions
repener:Certificate
repener:EnergyQualificationObtained
Prefixes:
sumo: http://www.ontologyportal.org/SUMO.owl#
repener: http://arcdev.housing.salle.url.edu/repener/lod/ontology/
Legend:
Object property
Relation between a domain ontology and a relational source

id	int
qualif_obtinguda	varchar
datasorcat	varchar
qualif_zona	varchar
id_useedifici	int
id_localitat	int
caract_inst_fontacs	varchar
caract_altres_acs	varchar
caract_gen_sup	varchar
qualif_consum_any	decimal
qualif_consum_m2	decimal
qualif_emis_any	decimal
qualif_emis_m2	decimal

Figure 5. Mappings between ICAEN source and the RÉPENER ontology.

For example, the mapping in Listing 3 relates the column *qualif_obtinguda* with the class *repener:EnergyQualificationObtained*. The example uses a simple SQL query, however the user can specify a complex query which could involve where, join, and group clauses among others.

```
<mapping1> a rr:TriplesMap;
  rr:logicalTable [
    rr:sqlQuery "SELECT qualif_obtinguda FROM icaen"];
  rr:subjectMap [
    rr:template "http://example.com/eqo/{id}";
    rr:class repener:EnergyQualificationObtained
  ].
```

Listing 3. An example of an R2RML mapping of the ICAEN database.

When The R2RML mapping in Listing 3 is used in an OBDA system the following RDF triples are generated (Listing 4).

```
<id1>    rdf:type    repener:EnergyQualificationObtained.
<id2>    rdf:type    repener:EnergyQualificationObtained.
<id3>    rdf:type    repener:EnergyQualificationObtained.
```

Listing 4. RDF triples generated by the R2RML mapping of Listing 3.

3 Automated Generation of Relational-to-Ontology Mappings

The contributions made in this research to the *automated generation of mappings between a relational source and an ontology* are described in this chapter. Current research documentation related to this area is summarized to support the contributions of this research. Three innovative methods are presented which are based on an intensive use of relational source contents and features of ontologies to increase the performance of the existing state-of-the-art mapping generators. The performance of the methods is evaluated and compared with the state of the art tools. Finally, the contributions and results are discussed.

3.1 Motivation and Goals

One of the main barriers in the adoption of OBDA systems for data integration is that the creation of mappings between a relational source and a domain ontology consumes a lot of resources. Domain experts and data owners are often required to create mappings manually. In doing so, they have to understand the structure of the relational source and the domain ontology. Automated mapping generation is a way to alleviate the users' burden. This way, users can focus on conceptual verification, fixing, and completing of mappings rather than on the mapping syntax.

Efforts towards the development of automated mappings between a relational source and a domain ontology have been carried out in several studies. Prime examples of the state-of-the-art systems are IncMap (Pinkel et al., 2013) and BootOX (Jiménez-Ruiz et al., 2015). In IncMap, an intermediate graph structure is created in order to map the relational schema and the domain ontology. Based on a flooding algorithm, IncMap can locate the mappings between the elements of the relational schema and the domain ontology. BootOX is based on a three-step process. First, a so-called putative ontology is generated from the relational source using direct mapping transformation rules such as relational tables becoming concepts, columns becoming data properties, and foreign/primary keys becoming object properties (Sequeda, Garcia-Castro, Corcho, Miranker, & Tirmizi, 2009). An example of a putative ontology and its database schema can be found in Figure 6. Second, BootOX uses LogMap (Jiménez-Ruiz & Cuenca Grau, 2011; Jiménez-Ruiz, Grau, Zhou, & Horrocks, 2012) as an ontology matching system to align the putative ontology and the domain ontology. Third, the final mappings are generated from the correspondences found by LogMap. Both tools – LogMap and BootOX – are compliant with R2RML recommendation.

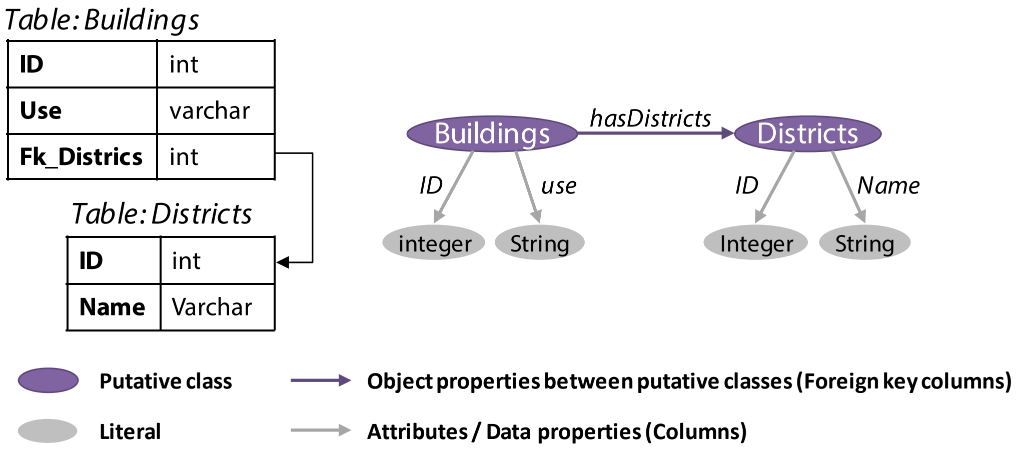


Table: Buildings

ID	int
Use	varchar
Fk_Districs	int

Table: Districts

ID	int
Name	Varchar

Buildings
hasDistricts
Districts
ID
use
integer
String
ID
Name
Integer
String
Putative class
Literal
Object properties between putative classes (Foreign key columns)
Attributes / Data properties (Columns)

Figure 6. Example of a putative ontology derived from a database schema.

Relational-to-ontology mapping generators have been evaluated using the RODI benchmark (Pinkel et al., 2016; Pinkel et al., 2015). The results of the evaluation have shown that current generators can address simple mappings. However, all systems failed on advanced tests such as those where relational data sources use design patterns that differ greatly from those used in ontologies. As a matter of fact, this is the most common OBDA scenario. One of the reasons for the poor performance is that current mapping generators basically rely on the relational schema and only barely take into account the contents (i.e., attribute values of the data instances) of the relational data source (i.e., database) and the features of the ontology. The contents of the database can have hidden relations between data that can help in the mapping generation process. For example, values of the attributes can be mined using terminology-based patterns to enrich the class hierarchy of the putative ontology and null values can reveal underlying class hierarchies. Indeed, BootOX has heuristics to analyse the contents of the database in order to infer a richer hierarchy of classes when the putative ontology is generated. However, those heuristics require input from the user; hence it is not a fully automated process.

With regards to the automated generation of mappings, the goal of this thesis is to improve relational-to-ontology mapping generators in order to obtain mappings of significantly higher quality compared to those generated by existing systems. To do so, we have developed innovative techniques that extensively utilize relational source content and features of the domain ontology to generate mappings and integrate these techniques into a relational-to-ontology mapping generator system.

3.2 Techniques for Improving the Mapping Generation

As previously stated, the relational-to-ontology mapping generation process is usually composed of three steps such as the process of the BootOX system. The steps are: 1) the generation of a putative ontology derived from a relational source, 2) the alignment of the putative and the domain ontologies by means of ontology matching techniques, and 3) the generation of the mappings according to the alignment. The main contribution of the line of work introduced in this chapter are three techniques which aim at improving each of the three steps of the relational-to-ontology mapping generation process (Figure 7). The three techniques (Sicilia & Nemirovski, 2016) are introduced in the following sections.

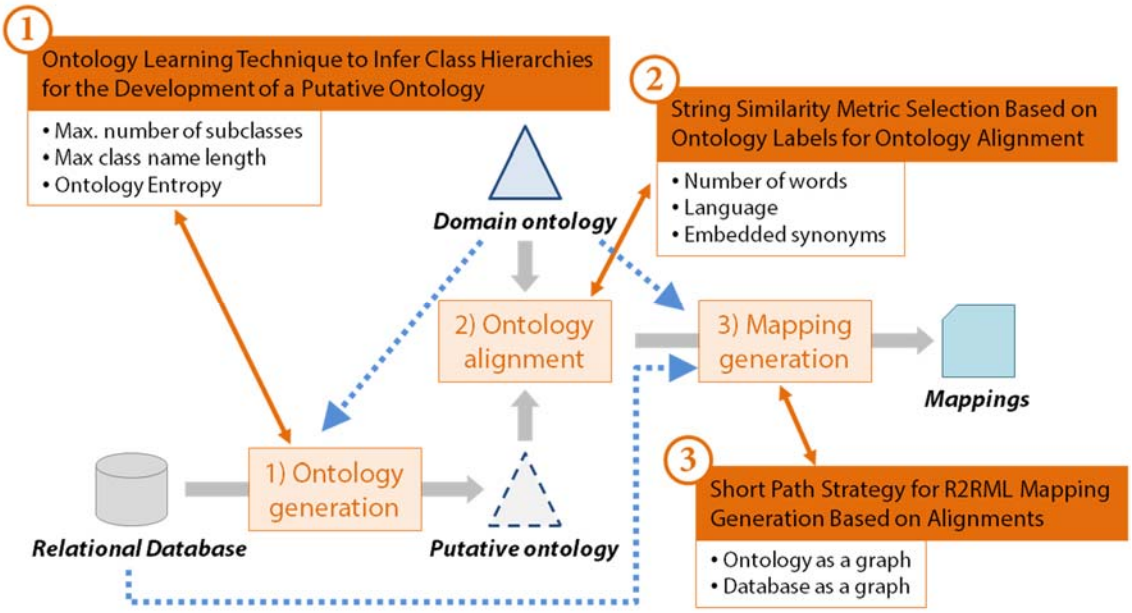

1
Ontology Learning Technique to Infer Class Hierarchies for the Development of a Putative Ontology
• Max. number of subclasses
• Max class name length
• Ontology Entropy
2
String Similarity Metric Selection Based on Ontology Labels for Ontology Alignment
• Number of words
• Language
• Embedded synonyms
Domain ontology
2) Ontology alignment
3) Mapping generation
Mappings
1) Ontology generation
Relational Database
Putative ontology
3
Short Path Strategy for R2RML Mapping Generation Based on Alignments
• Ontology as a graph
• Database as a graph

Figure 7. Ontology generation, ontology alignment, and mapping generation techniques.

3.2.1 Ontology Learning Technique to Infer Class Hierarchies for the Development of a Putative Ontology

In a basic process of putative ontology generation, a concept (i.e., OWL class) is obtained from each table of a relational source. Attributes of the table become data properties (i.e., OWL Datatype property) whose domain is the class generated for the table. Object properties (i.e., OWL object properties) are obtained from the relations between tables (i.e., primary/foreign keys). A comprehensive list of transformation rules can be found in (Sequeda et al., 2009).

The purpose of the *Ontology learning technique* developed in this research is to enrich the putative ontology which has been previously derived from a relational source. It is assumed, that the enriched putative ontology will have more opportunities than a regular putative ontology to be aligned with a domain ontology. Therefore, ontology matching techniques will find more correspondences between the enriched putative ontology and the domain ontology than correspondences between the putative ontology in its original form and the domain ontology. On the one hand, class hierarchies are extracted from the attribute values of the data instances of the relational source by means of the *Ontology learning technique* and incorporated into the putative ontology. On the other hand, the extraction process of a class hierarchy needs to take into account the characteristics of the domain ontology. The extraction process we have developed is guided by features of the domain ontology such as maximum number of subclasses, maximum class name length, and ontology entropy among others. This way, the class hierarchies obtained from the relational source will correspond with the features of the domain ontology better than when the features of the domain ontology are not taken into account.

The extraction process utilises the values of the attributes of tables of a relational source to derive concepts. These concepts have as a super class the concept obtained from the table name. For example, in Figure 8, the concept *Buildings* is obtained from table *Buildings.* The concepts *Residential*, *Office*, and *Stadium* are obtained from the values of the attribute *Use*. The main issue in this context is identifying attributes that can be mined to extract concepts and determining which values of those attributes can serve as concepts names.

A set of rules have been defined to discard attributes and values of those attributes. Examples of these rules are:

- the length of a value is much greater than the maximum length of a class name of the domain ontology,
- the number of different values of an attribute is greater than the maximum number of subclasses of any class of the domain ontology,
- the entropy of an attribute is greater than the maximum entropy of the ontology.

The concept of ontology entropy used in the rules is based on the work of Cerbah (2008) who applied similar methods to obtain enriched hierarchies from a relational source where the attributes were selected if they could reveal a specific role in the table. The concept of entropy in information theory is a measure of the uncertainty of a set of values. Therefore, the entropy of an attribute can be considered as the number of different values and the entropy of a class of an ontology can be the number of different subclasses of that class. Attributes with the highest entropy are usually the primary keys. The description of how the entropy concept is applied to calculate the entropy of attributes of a relational source and the classes of a domain ontology can be found in (Sicilia & Nemirovski, 2016).

For example, in Figure 8, the value *Very long street name* of the attribute *Address* is discarded because its length is 21 and the maximum class name length of the domain ontology is 11. Moreover, the values of the attribute *ID* are discarded because its entropy is greater (i.e., 9.32) than the entropy of the domain ontology (i.e., 5.31). In addition to the rules listed above, there are basic rules that are used to discard values such as *the value has to be a text* (i.e., numbers and Boolean values are not permitted) and the value cannot contain a URL. In the example of Figure 8, the values *1234* and *http://url.com* are discarded according to these rules.

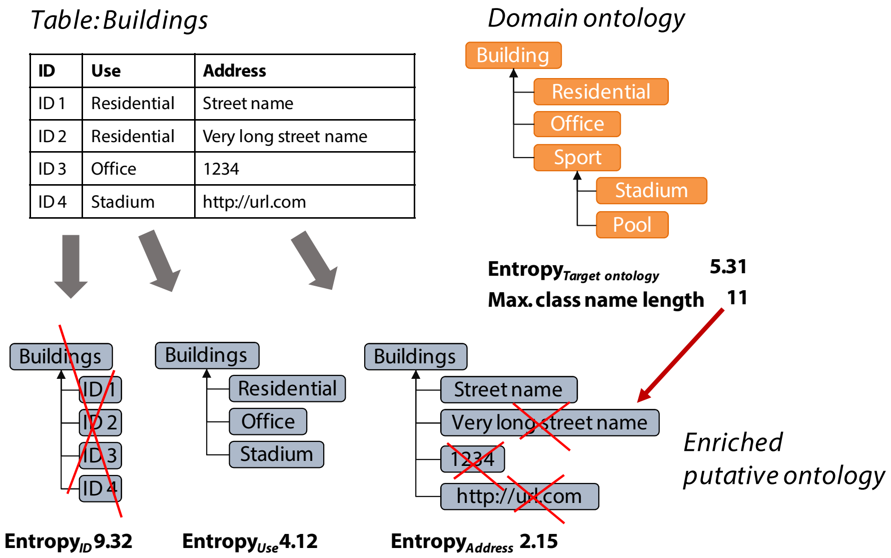


Table: Buildings

ID	Use	Address
ID 1	Residential	Street name
ID 2	Residential	Very long street name
ID 3	Office	1234
ID 4	Stadium	http://url.com

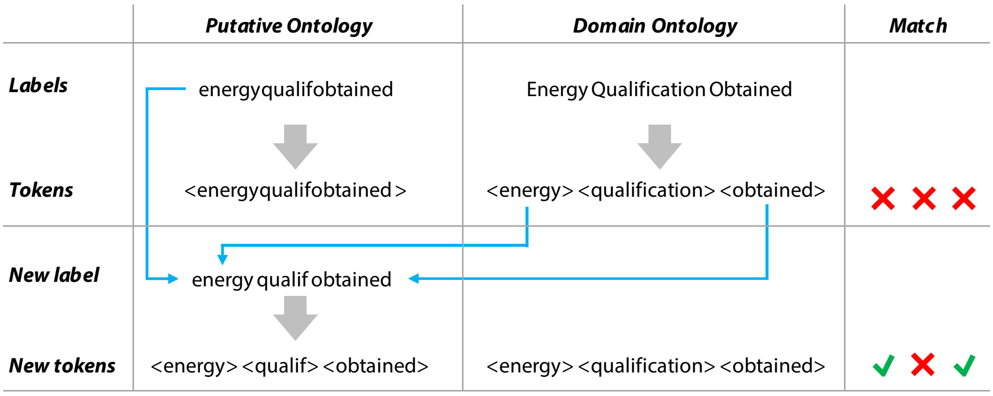
Figure 8. Examples of accepted and discarded values in ontology learning technique.

3.2.2 String Similarity Metric Selection Based on Ontology Labels for Ontology Alignment

The purpose of the *String similarity metric selection technique* is to improve the performance of ontology matching methods in the particular case of aligning a putative ontology and a domain ontology. Putative and domain ontologies may have differences that might affect the performance of the common ontology matching methods. Putative ontologies – which reflects the structure of a relational schemas – describe the syntactical structure on a very low level of granularity, but domain ontologies usually model high-level semantic information. Moreover, different design patterns are used in ontologies and relational schemas because the goal of a domain ontology is to reach a consensus on the representation of a domain while the goal of a relational source is to optimise the storage and the query of data. Full-featured ontology alignment systems rely on syntactic, semantic, and structural similarity metrics to find correspondences between entities in putative and domain ontologies. In the case of aligning putative (derived from a relational source) and domain ontologies the use of structural metrics may hinder the detection of correspondences. For example, the structural metrics take into account the similarity between domains and ranges of the putative ontology and the domain ontology. This way, object properties in spite of having the same name in the putative and in the domain ontology might not be matched due to differences in the structure of the putative and domain ontologies.

Therefore, the technique we propose uses string-based metrics instead of all the features of an ontology alignment system. The performance of string similarity metrics depends on the features of putative and domain ontologies. Chetham and Hitzler (2013, 2014) implemented StringAuto and PropString, string-based alignment systems for classes and properties which select the string similarity metric based on the features of the putative and domain ontologies. They take into account the number of words per entity label after tokenization, the language of the ontologies, and the existence of embedded synonyms.

The performance of PropString decreases when labels – i.e., concept name – of the putative ontology cannot be tokenized but the labels of the domain ontology can be tokenized and vice versa. For example, a matcher will not find a correspondence between the labels *energyqualifobtained* and *EnergyQualificationObtained* because the first one cannot be tokenized and the second label can be tokinized as the terms *Energy*, *Qualification*, and *Obtained*. To address this issue, we have extended PropString to carry out a tokenization process on labels of the putative ontology. Labels of the putative ontology which cannot be tokenized are modifed according to the tokens found in the labels of the domain ontology. For example, in Figure 9, the tokens of the labels *energyqualifobtained* and *EnergyQualificationObtained* do not match. Therefore, the label *energyqualifobtained* is modified with the tokens *Energy*, *Qualification*, and *Obtained*. First, label is modified with the token *Energy* to become *energy qualifobtained*. Then, with the token *Qualification* the label is not modified. Finally, with the token *Obtained* the final label becomes *energy qualif obtained*. The tokens of the new label are *energy*, *qualif*, and *obtained*. Two of them match with the tokens of the label of the domain ontology. This way, we increase the chances of finding correspondences.



	Putative Ontology	Domain Ontology	Match
Labels	energyqualifobtained	Energy Qualification Obtained	
Tokens	<energyqualifobtained >	<energy> <qualification> <obtained>	✗ ✗ ✗
New label	energy qualif obtained		
New tokens	<energy> <qualif> <obtained>	<energy> <qualification> <obtained>	✓ ✗ ✓

Figure 9. String similarity metric extension to enable tokenization.

3.2.3 Short Path Strategy for R2RML Mapping Generation Based on Alignments

The current mapping generators produce the mappings between the relational data source and the ontology in the process of generating a putative ontology. Later, the alignment between the putative ontology and the domain ontology is used to update the mappings. Finally, the modified mappings can be used to query the source by means of queries referring to the domain ontology. The main drawback of this approach is that the mappings strongly reflect the relational source structure which might not be the same as the structure of the ontology since different design patterns are usually used in ontologies and relational schemas. This may lead to the generation of incorrect mappings.

To overcome this problem, we have developed the *Short path strategy technique* whose purpose is to generate the final R2RML mappings according to the structure of the domain ontology. The technique initiates with the correspondences found by a matcher which relates elements of the putative ontology to elements of the domain ontology using a property (i.e., *owl:sameAs*). Subsequently, since not all the object properties of the domain ontology are matched with the object properties of the putative ontology, new correspondences for object properties are established following a connectivity rule which takes into account the domain ontology structure (Figure 10). Therefore, this technique does not generate mappings in two steps – first in the putative ontology and second as after the alignment – as current mapping generators does.

In order to assess the connectivity between elements (e.g., tables and classes), the putative ontology and the domain ontology are represented as two graph structures. A graph is generated from the putative ontology. Since it reflects the database structure, database tables and columns are nodes, and foreign keys relations are edges between correspondent columns. Another graph is derived from the domain ontology where the classes and properties are nodes while the edges are the domain and ranges of the properties. Two concepts – already aligned to putative classes – can be connected through an object property from the domain ontology if two conditions are met:

1. There is a path connecting them according to the target ontology.
2. There is a path between the pair of putative classes (i.e., tables) according to the database schema.

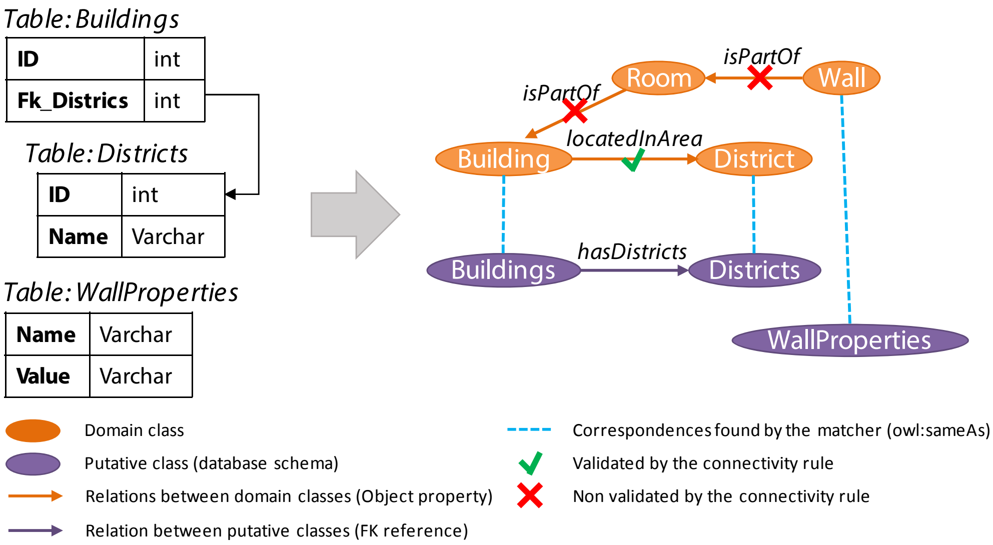


Table: Buildings

ID	int
Fk_Districs	int

Table: Districts

ID	int
Name	Varchar

Table: WallProperties

Name	Varchar
Value	Varchar

isPartOf
Room
isPartOf
Wall
isPartOf
locatedInArea
Building
District
hasDistricts
Buildings
Districts
WallProperties
Domain class
Putative class (database schema)
Relations between domain classes (Object property)
Relation between putative classes (FK reference)
Correspondences found by the matcher (owl:sameAs)
✓ Validated by the connectivity rule
✗ Non validated by the connectivity rule

Figure 10. Connectivity rule validation.

This way, the connectivity between concepts is assured at the level of the domain ontology – at least one object property will exists between those concepts – and at the level of the relational source – an SQL query can be obtained which involves the corresponding tables of both concepts and join clauses. The alignment between two classes (putative and domain) is reflected in the R2RML mapping by means of SQL queries. The architecture of the queries depends on the object properties that connect those classes. For example, in Figure 10, the concept *Building* is connected to the concept *District* with an object property (i.e., *locatedInArea*) and a SQL query can be generated to be used in the mapping. The query will be *SELECT * FROM Buildings JOIN Districts ON Buildings.Fk_Distrits = Districts.ID*. However, the connectivity rule fails when applied to the object property isPartOf since there are no foreign key relations between the tables *Buildings* and *WallProperties*.

The *Short path strategy technique* takes into account the possible differences between the structure of the putative and the domain ontologies. It can handle two cases which are illustrated in Figure 11. The first case occurs when the path with the minimum length between a pair of concepts of the domain ontology has more nodes (concepts) than the path with the minimum length between the aligned concepts of the putative ontology. To address this case, additional mappings (i.e., triples maps) are generated to make the final mappings consistent. The IRIs and SQL queries needed for those additional mappings are the same as the previous concepts. In the example of the Figure 11, the concepts *Room* and *Wall* will have the same IRI and correspond to the same SQL query as the *Building* concept. The second case occurs when the path with minimum length between a pair of concepts of the putative ontology has more nodes (i.e., concepts) than the path with the minimum length between the aligned concepts of the domain ontology. This case is solved by generating a SQL query with multiple joins clauses to connect all nodes of the path. The mappings obtained through *Short path strategy technique* for the example of Figure 11 can be found in Appendix B.

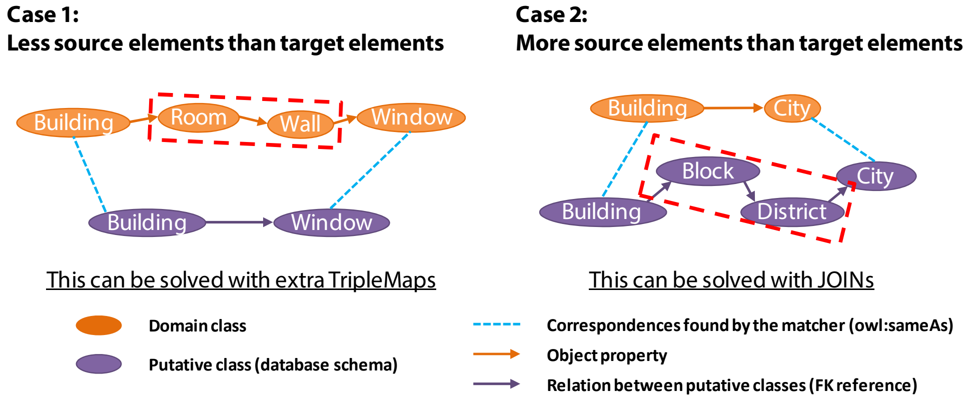

Case 1:
Less source elements than target elements
Building
Room
Wall
Window
Building
Window
This can be solved with extra TripleMaps
Case 2:
More source elements than target elements
Building
City
Block
City
Building
District
This can be solved with JOINs
Domain class
Putative class (database schema)
Correspondences found by the matcher (owl:sameAs)
Object property
Relation between putative classes (FK reference)

Figure 11. Differences between the structure of the putative and the domain ontologies.

The *Short path strategy technique* can address redundant mappings such as when a class of the domain ontology is aligned with more than one putative class. For each domain class, it is checked if the database tables referred by the correspondent putative classes are connected by a foreign key. For example if the class of a domain ontology *ex:Building* has been matched with the putative classes *put:Buildings* and *put:Buildgs*, the database tables corresponding to these two classes (e.g., the table *Buildings* and the table *Buildgs*) should be connected by a relation, i.e. one of the table should contain a foreign key referring to the primary key of the other table. This way, the foreign key *fkBuilding* from table *Buildgs* should refer to column *ID* of table *Buildings*. If such a relation is missing, the corresponding alignments are removed. With this rule we ensure that the IRIs templates of the subject maps are homogeneous and consistent for each target ontology class.

3.3 Evaluation and Results

The three techniques described above have been integrated in AutoMap4OBDA (**AM4O**), a system to automatically generate R2RML mappings from a relational database and a domain ontology in (Sicilia & Nemirovski, 2016). The performance of AutoMap4OBDA, when all three techniques mentioned above have been applied simultaneously, has been compared with the performance of existing state-of-the-art relational-to-ontology mapping generators using the RODI benchmark[18]. The RODI benchmark offers basic test scenarios from conference, geographical, and oil and gas domains; and mixed scenarios in the conference domain where the database schema has to be matched to an ontology from another scenario. Each scenario is composed of databases, ontologies, and a set of queries to test how the mappings generated by mapping generating system are performing. RODI simulates real-world scenarios by creating different databases with modifications to reproduce design patterns and anti-patterns in databases (e.g., *Adjusted* naming, *Restructured hierarchies*, *Combined case*, M*issing keys*, *partial denormalization*). Moreover, RODI includes a complex scenario compared with the conference scenarios in the domain of geographical data, scenarios which combine databases and ontologies from scenarios of the conference domain, and an actual real-world database and ontology in the oil and gas domain. For a further explanation of the scenarios addressed by RODI refer to (Pinkel et al., 2016; Pinkel et al., 2015). The mappings are evaluated in terms of the percentage of successful queries answered for each scenario. The mapping generators selected for the comparison were BootOX (Jiménez-Ruiz et al., 2015), IncMap (Pinkel et al., 2013), ontop (Rodríguez-Muro & Rezk, 2015), MIRROR (de Medeiros, Priyatna, & Corcho, 2015), COMA++ (Aumueller, Do, Massmann, & Rahm, 2005), and D2RQ (Bizer & Cyganiak, 2007). The results for BootOX (**B.OX**), IncMap (**IncM.**), ontop, MIRROR (**MIRR.**), COMA++ (**COMA**), and D2RQ have been obtained by RODI team (Table *2*).

The results demonstrate that AutoMap4OBDA achieves the top position for eleven out of seventeen scenarios and is in second position in three scenarios. The scores are based on average of per-test F-measure. The results of AutoMap4OBDA in the mixed scenarios such as *Target ontology: CMT*, *Target ontology: Conference*, and *Target ontology: SIGKDD* is not as good as the other scenarios because the level of semantic heterogeneity is much higher than in the basic scenarios. It is worth mentioning the *GeoData* scenario where the methods *String similarity metric selection based on ontology labels for ontology alignment* and *Short path strategy for R2RML mapping generation based on alignments* helped to find

[18] http://www.cs.ox.ac.uk/isg/tools/RODI/

considerably more properties than the other systems achieving a performance more than three times higher than the following system.

Table 2. *Overall scores of the state of the art tools and AutoMap4OBDA in RODI scenarios (scores based on average of per-test F-measure, best numbers per scenario in bold)*

Scenarios		**B.OX**	**IncM.**	**ontop**	**MIRR.**	**COMA**	**D2RQ**	**AM4O**
Adjusted naming	CMT	**0.76**	0.45	0.28	0.28	0.48	0.31	0.56
	Conference	0.51	0.53	0.26	0.27	0.36	0.26	**0.56**
	SIGKDD	0.86	0.76	0.38	0.30	0.66	0.38	**0.86**
Restructured	CMT	0.41	**0.44**	0.14	0.17	0.38	0.14	0.41
	Conference	0.41	0.41	0.13	0.23	0.31	0.21	**0.54**
	SIGKDD	0.52	0.38	0.21	0.11	0.41	0.28	**0.72**
Combined case	SIGKDD	0.48	0.38	0.21	0.11	0.28	0.28	**0.62**
Missing FK	Conference	0.33	0.41	-	0.17	0.21	0.18	**0.49**
Denormalized	CMT	0.44	0.40	0.20	0.22	-	0.20	**0.52**
GeoData	Classic Rel	0.13	0.08	-	-	-	0.06	**0.44**
Oil&Gas domain	User Queries	0.00	0.00	0.00	0.00	-	0.00	0.00
	Atomic	0.14	0.12	0.10	0.00	0.00	0.08	**0.23**
Target ontology: CMT	Conference	0.20	**0.35**	0.10	0.00	0.00	0.10	0.15
	SIGKDD	0.33	0.33	0.19	0.00	0.14	0.19	**0.38**
Target ontology: Conference	CMT	0.20	0.34	0.05	0.00	0.05	0.05	**0.39**
	SIGKDD	0.13	**0.30**	0.09	0.00	0.04	0.09	0.17
Target ontology: SIGKDD	CMT	0.51	**0.57**	0.19	0.00	0.24	0.26	0.41
	Conference	0.24	**0.44**	0.13	0.00	0.09	0.14	0.19
Average of the tests		0.36	0.37	0.15	0.10	0.20	0.18	**0.43**

Full-featured alignments systems – such as those used in the evaluation of BootOx and ontop among others – have difficulties matching object properties when the structure of the source (putative) ontology is not similar to the structure of the target (domain) ontology. AutoMap4OBDA does not directly match the object properties of the putative and domain ontologies, but the object properties are set by the *Short path strategy technique* described in Section 3.2.3. For example, the correspondences illustrated in Figure 12 can be found by AutoMap4OBDA but not by full-featured alignments systems. In *sigkdd_mixed* scenario, the domain ontology has the object property *isCommitteOf* whose domain is *Commitee* and range is Conference. Moreover, in the putative ontology the correspondent object property is *commitee* whose range class is *conferences* and whose domain classes are *best_paper_awards_committs*, *organizing_committees*, and *program_committees*. Those domain classes are subclasses of *committe* class in the domain ontology however an ontology matcher cannot match both object properties. In this case AutoMap4OBDA, after having aligned the classes correctly, fulfils the alignment of the object property in the *Short path strategy technique.*

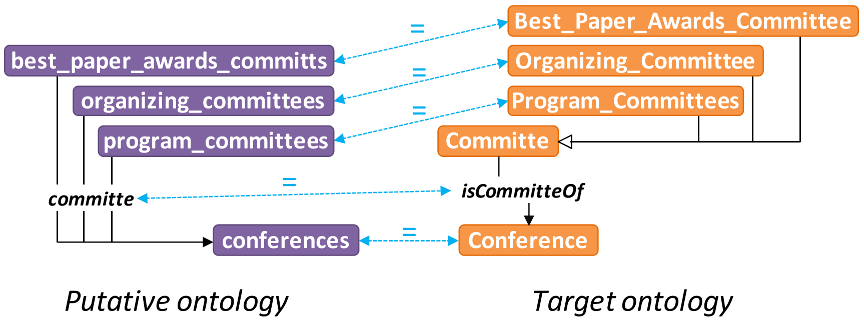


Figure 12. Example of correspondences found by AutoMap4OBDA in sigkdd_mixed scenario.

The favourable result in the *Oil&Gas domain Atomic* scenario – compared to other systems – has been achieved thanks to the *Ontology learning technique*. In this scenario, AutoMap4OBDA was able to find several mappings where the values of the columns are used to set classes of the subjectMap such as the mapping in Listing 6.

```
<mapping1_201> a rr:TriplesMap;
  rr:logicalTable [
    rr:sqlQuery "SELECT pipnpdidpipe FROM pipline
                 WHERE pipmedium = 'Oil'" ];
  rr:subjectMap [
    rr:template "http://.../oilpipeline/{pipnpdidpipe}";
    rr:class http://sws.ifi.uio.no/vocab/npd-v2#OilPipeline].
```

Listing 5. An R2RML mapping using values of the columns to filter the classes.

AutoMap4OBDA outperforms other innovative mapping generators because they cannot generate this kind of mapping in an automated way. However, no results have been achieved in *Oil&Gas domain User Queries* scenario. This is a real-world scenario where the queries go beyond returning a simple result list of all objects of one class.

Despite the favourable results, AutoMap4OBDA is far from being able to generate a full list of mappings derived from a relational database and a domain ontology. Indeed, the average F-measure obtained in the RODI scenarios is 0.43 which is not a remarkable result but it is a step forward in relational-to-ontology mapping generators since it has improved the results by 0.06 with regards to the next contender – IncMap – which has an average of 0.37.

The average execution time of AutoMap4OBDA – in an Intelcore i5 architecture with 10GB of RAM – has been less than 25 seconds per scenario for 15 scenarios, 57.96 seconds for the *Adjusted naming Conference* scenario, and 434.40 seconds for the *Oil&Gas* whose database has 70 tables with 250k records and the target ontology has 344 classes, 148 object properties, and 237 data properties. The complete results for each scenario are shown in Table 3. There is a strong dependence between the number of records that a database have and the execution time. Indeed, the correlation between them is 0.994. The reason behind this is that the *Ontology learning technique* queries the database to calculate the entropy of each attribute of the tables.

Table 3. *Comparative execution time of AutoMap4OBDA in RODI Scenarios*

Scenarios		**Number of tables**	**Number of records**	**Number of ontology classes**	**Execution time (s)**
Adjusted naming	CMT	48	9,153	31	6.05
	Conference	66	12,508	60	57.96
	SIGKDD	58	6,677	50	9.14
Restructured	CMT	32	5,386	31	4.42
	Conference	30	6,270	60	4.16
	SIGKDD	22	4,352	50	2.42
Combined case	SIGKDD	22	4,352	50	2.33
Missing FK	Conference	30	6,270	60	3.04
Denormalized	CMT	30	5,762	31	4.45
GeoData	Classic Rel	38	26,904	51	24.30
Oil&Gas domain	User Queries	70	257,784	344	465.74
	Atomic	70	257,784	344	434.40
Target ontology: CMT	Conference	32	5,386	60	2.94
	SIGKDD	32	5,386	50	2.59
Target ontology: Conference	CMT	30	6,270	31	2.17
	SIGKDD	30	6,270	50	2.73
Target ontology: SIGKDD	CMT	22	4,352	31	1.60
	Conference	22	4,352	60	2.68

3.4 Discussion

The goal to automating the relational-to-ontology mapping process is to reduce the burden of users who are deploying an OBDA system. Several systems have been developed with that purpose. However, as has been demonstrated through the RODI benchmark, they do not perform well, particularly in real-world scenarios. They basically rely on a relational schema and only barely take into account the contents of the relational data source and the features of the domain ontology. To increase the performance of relational-to-ontology mapping generators, three innovative techniques have been presented that make an intensive use of relational source contents and features of the domain ontology to generate mappings and these techniques have been integrated into in AutoMap4OBDA, a relational-to-ontology mapping generator system.

The evaluation of the performance of the techniques – implemented in the AutoMap4OBDA tool – represent a step forward in relational-to-ontology mapping systems. However, they are far from providing a universal solution for all types of mappings. In spite of recent advancements, there is still knowledge to be gained regarding automated mapping generation for real-world scenarios. In those scenarios, acronyms for naming tables and columns are commonly used by the data source developers. Multiple languages can be found in the same data source and the mappings depend on relations hidden in the data only known by data source developers. In those scenarios, queries are tailored by users with deep knowledge about the domain of discourse. The techniques proposed in this research can be enhanced to overcome these particularities of the real-world scenarios. Some of these particularities can be solved basing on the research carried out by the ontology alignment community (e.g., multi-language features).

During the design and development of the techniques presented in this research, the RODI benchmark has been continuously applied to evaluate their performance. Using this kind of benchmark based on a generic, effective, and reliable evaluation of the quality of computed mappings, has been useful to unveil design errors and implementation bugs. On one hand, the scenarios provided by RODI – in particular the synthetic scenarios from the Conference domain – have been helpful to make the techniques robust to errors. On the other hand, the real-world scenarios – such as the scenarios in the oil and gas domain – are essential to further enhance and improve the techniques. This is because these scenarios the queries are tailored by users with great knowledge with regard the data source and the domain ontology. The real-world scenarios can help to open new research directions with regard *automated generation of relational-to-ontology mappings*.

Following the results of the evaluation, we believe that generation of relation-al-to-ontology mappings is a task that cannot be completely automated. It will always be necessary that an expert in the domain from which the data originates validates and complements the mappings automatically generated by the system. A parallel contribution of this research – *visual support for relational-to-ontology mapping editing* – introduced in the next chapter is focused on providing representations and environments to help users without ontology engineering and database skills in the creation and maintenance process of relational-to-ontology mappings.

The outcomes of this research have been presented in:

- Sicilia, Á., & Nemirovski, G. (2016). AutoMap4OBDA: Automated Generation of R2RML Mappings for OBDA. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings* (pp. 577–592). Bologna, Italy: Springer International Publishing. http://doi.org/10.1007/978-3-319-49004-5_37

4 Visual Support for Relational-to-Ontology Mapping Editing

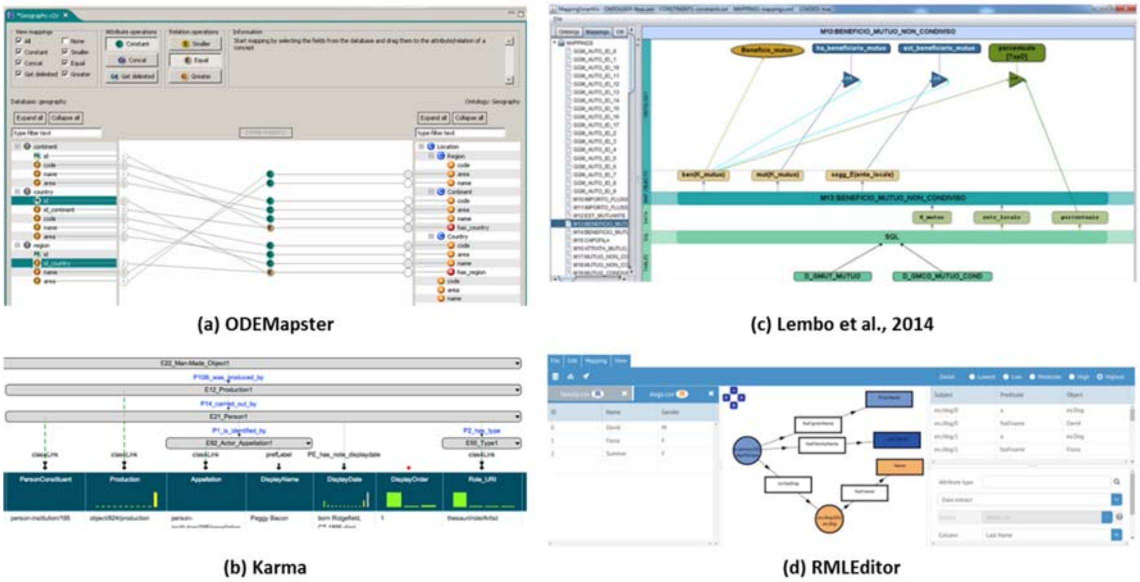
The work conducted in this research to provide *visual support for visual relational-to-ontology mapping creation* is introduced in this chapter. The motivation and goals are presented as a result of the study of the current research works carried out on this matter. A visual representation of relational-to-ontology mappings based on graph layouts is presented to aid users in creating and maintaining mappings. The tool Map-On, one of the outcomes of this research, is presented. It is a graphical environment which supports users to manually establish relations between elements of a database and of a domain ontology in the context of an OBDA scenario. An evaluation of the tool has been conducted to validate the usability of Map-On and to demonstrate that it can be used by non-ontology experts. Finally, the contributions and results are discussed.

4.1 Motivation and Goals

As stated in the previous chapter, one of the main barriers in the adoption of OBDA systems for data integration is that creation of relational-to-ontology mappings is a process requiring a high consumption of human resources. The contributions for automating the mapping generation presented in previous chapter can find around 43% of mappings, in some cases they can even reach 90%. Mappings which are not generated by these techniques should be manually created by users who have to understand the structure of the relational source and the domain ontology. Creating mappings requires knowledge of the mapping language (e.g., R2RML), advanced technical skills (e.g., SQL), and expertise in ontology design. Data owners have knowledge about their relational data sources and domain experts can understand parts of the domain ontology, however, these users lack technical expertise to create mappings. For non-ontology experts, the main barrier is often the lack of a visual representation of the mappings. In practice, ontology visualization techniques could help non-experts in ontology engineering to overcome the lack of expertise and to inspect, navigate, and verify the ontologies and mappings (Lanzenberger, Sampson, & Rester, 2009).

Several tools have been developed to assist experts in defining the mappings between the data sources and the ontologies. For instance, ontop (Calvanese, Cogrel, Komla-ebri, Kontchakov, & Lanti, 2015) and the mapping editor developed by Segupta (Sengupta, Haase, Schmidt, & Hitzler, 2013) aim to help advanced users instead of domain experts and data owners. However, these tools do not provide graphic visualization of any kind. Another group of tools includes editors with graphic visualization of mappings based on tree layouts such as ODEMapster (Priyatna, Villazón-Terrazas, Barrasa, & Schulte, 2011) (Figure 13a), Karma (Knoblock et al., 2012) (Figure 13b), and R2RML By Assertion (Neto, Vidal, Casanova, & Monteiro, 2013). The limitation of these tools lies in the use of tree layouts which are unable to represent the complete structure of the database schema, ontology and mapping by itself since the structure of an ontology can be an arbitrary complex graph. An example of a mapping editor that uses an advanced graph layout is the mapping visualization model presented by Lembo et al., (2014) (Figure 13c). In this case, the mappings are presented in a graph layout including three views focused on the mapping, the ontology, and the source. But a complete overview of the all mappings at once is not provided. RMLEditor is another example of an editor that presents the mappings using a graph layout (Heyvaert et al., 2016) (Figure 13d). The limitation of the mapping representation of RMLEditor is that the structure of the relational source is not included in the mapping representation. In Figure 13, the most representative visual representations of mappings are shown.

The motivation behind our research concerning *visual support for relational-to-ontology mapping editing* has been to overcome the limitations of the existing visualization tools mentioned above. Therefore, we have devised a visual representation of relational-to-ontology mappings that helps data owners and domain experts to understand the structure of the relational data sources, ontology and the mappings between them. Furthermore, we have developed Map-On, a graphical environment for ontology mapping which includes the visual representation devised in this line of work.


(a) ODEMapster

(c) Lembo et al., 2014

(b) Karma

(d) RMLEditor

Figure 13. State of the art of visual representations of mappings.

4.2 Visual Representation of Relational-to-Ontology Mappings

When considering the visualization of relational-to-ontology mapping, the representation of ontologies as a graph layout is probably the most natural and the most common technique that can be used, also for mappings. Indeed, graph layouts are more suitable for overviews and their flexibility can help users to maintain focus during mapping tasks (Fu, Noy, & Storey, 2013). A recent prominent example is VOWL, a visual language for visualizing ontologies as a force-directed graph layout (Lohmann, Negru, Haag, & Ertl, 2014). In a graph layout representation of an ontology, the concepts (i.e., classes) are displayed as nodes and the relations (i.e., properties) as edges. One of the handicaps of graph layouts is that they can become difficult to manage once the nodes being visualized exceed a certain number.

The mappings between a database schema and an ontology are a set of relations between their elements, in particular between columns of relational tables and elements of an ontology such as concepts and data properties. That is when, for instance, a column of a relational table is used to define the IRI of a R2RML subject map and a concept of the ontology is utilized to define the type of a R2RML subject map. The relations can have different cardinalities. For example, a column of a relational table can be mapped to different concepts of the domain ontology and one ontology concept can be mapped to more than one relational tables. Therefore, it becomes intuitive to represent the mappings graphically as edges between columns and concepts (Figure 14). Moreover, the relational sources can also be represented as graphs where the tables and columns are represented as nodes, relations among columns and tables as edges, and relations between tables – by means of foreign key constraints – as edges as well. For example, in Figure 14, the table *icaen_adm_cee* and its columns *ID*, *Superficie*, and *Provincia* are nodes while the edges among these nodes indicates that these columns belong to that table.

In the visual representation of relational-to-ontology mappings that we have devised, tables and their columns are visualized as purple rectangles connected with a solid purple line. The relationships between tables are shown as a purple dashed line between foreign key and primary key constraints (e.g., *ID* column from table *icaen_adm_cee* and column *ID_CEE* from table *icaen_adm_cee_tancaments_opacs*). The ontology concepts are represented as orange ellipses, the roles (i.e., object properties) as directed solid orange lines, and the attributes (i.e., data properties) as green ellipses. The mappings between the elements of the database and the ontology are displayed with dashed blue lines. The visual representation of a mapping is a top-down visualization in which the elements of the ontology and database schema (i.e., tables, columns, concepts, roles and attributes) involved in the mapping are visualized in one single representation similar to a global view. Through this visual language, a user can grasp both, the database and ontology structures.

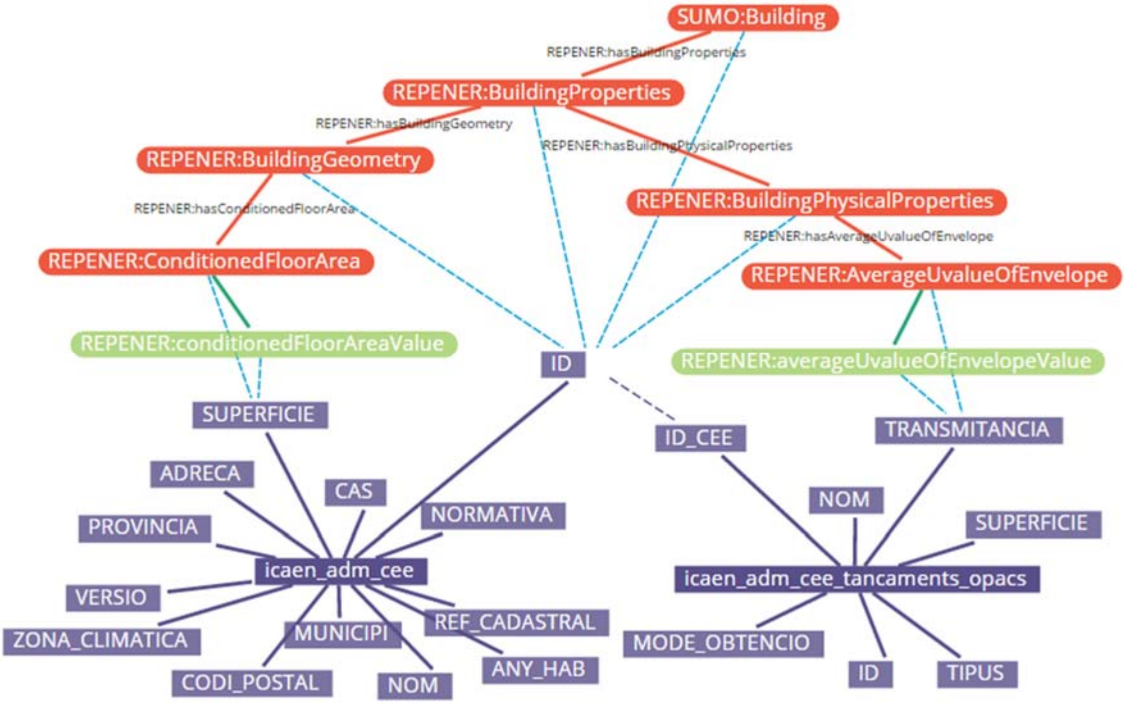

SUMO:Building
REPENER:hasBuildingProperties
REPENER:BuildingProperties
REPENER:hasBuildingGeometry
REPENER:hasBuildingPhysicalProperties
REPENER:BuildingGeometry
REPENER:hasConditionedFloorArea
REPENER:BuildingPhysicalProperties
REPENER:hasAverageUvalueOfEnvelope
REPENER:ConditionedFloorArea
REPENER:AverageUvalueOfEnvelope
REPENER:conditionedFloorAreaValue
ID
REPENER:averageUvalueOfEnvelopeValue
SUPERFICIE
ID_CEE
TRANSMITANCIA
ADRECA
CAS
NOM
NORMATIVA
SUPERFICIE
PROVINCIA
icaen_adm_cee
icaen_adm_cee_tancaments_opacs
VERSIO
REF_CADASTRAL
ZONA_CLIMATICA
MUNICIPI
MODE_OBTENCIO
ANY_HAB
CODI_POSTAL
NOM
ID
TIPUS

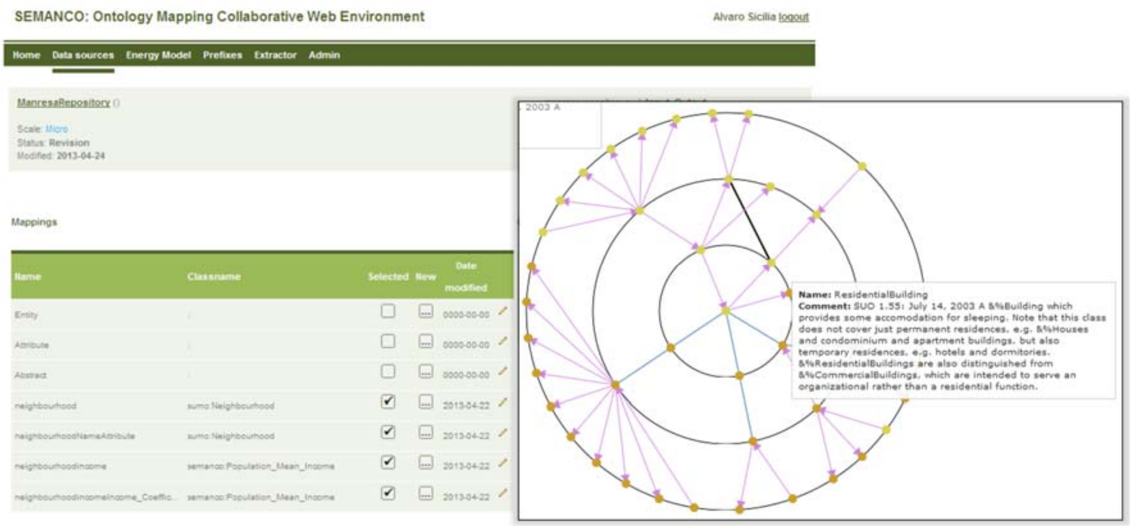
Figure 14. Visual representation of a mapping.

Furthermore, users need to personalise the visual representation of mappings. This can be achieved by placing the nodes in different positions in order to create a proper layout according to their understanding of the database and ontology structures. In scenarios with a considerable number of mappings the user might find it necessary to group the layouts in mapping spaces. These spaces are partial views of an entire picture of mappings between an ontology and a database. Such spaces contain a limited set of ontology and database entities and serve to partition a complex mapping task into a set of less complex and smaller tasks.

4.3 Map-On: A Web-Based Editor for Visual Ontology Mapping

Map-On has been developed to overcome the limitations of the existing tools for editing relational-to-ontology mappings. Map-On is a graphical environment for ontology mapping to help different kinds of users – domain experts, data owners, and ontology engineers – in the creation and maintenance of mappings between a database and a domain ontology using the R2RML recommendation (Sicilia, Nemirovski, & Nolle, 2016).

The development of Map-On tool started from on the Ontology Mapping Collaborative Web Environment tool developed in the SEMANCO project. This environment included a visual representation of the ontology using a radial graph layout. However visual representations of the mappings were not developed during the project. Because of this, users without technical skills found it difficult to use this tool (Figure 15). To solve this problem, the visual representation for relational-to-ontology mappings described in Section 4.2 has been included in Map-On.



ManresaRepository ()

Scale: Micro
Status: Revision
Modified: 2013-04-24

Mappings

Name	Classname	Selected	New	Date modified
Entity		□		0000-00-00
Attribute		□		0000-00-00
Abstract		□		0000-00-00
neighbourhood	sumo:Neighbourhood	☑		2013-04-22
neighbourhoodNameAttribute	sumo:Neighbourhood	☑		2013-04-22
neighbourhoodIncome	semanco:Population_Mean_Income	☑		2013-04-22
neighbourhoodIncomeIncome_Coeffic...	semanco:Population_Mean_Income	☑		2013-04-22

Name: ResidentialBuilding
Comment: SUO 1.55: July 14, 2003 A &%Building which provides some accomodation for sleeping. Note that this class does not cover just permanent residences, e.g. &%Houses and condominium and apartment buildings, but also temporary residences, e.g. hotels and dormitories. &%ResidentialBuildings are also distinguished from &%CommercialBuildings, which are intended to serve an organizational rather than a residential function.

Figure 15. Ontology visualization of the old version of Map-On.

The Map-On editor provides a graphical environment for ontology mapping creation using an interactive graph layout. The Map-On graphic user interface is based on a point-and-click paradigm where most of the user's actions are carried out with the cursor. The main benefits of this kind of interfaces are the high comfort and the diminished initiation barriers for those users who are lacking skills in mapping languages such as R2RML. Furthermore, the interface provides easy access to the elements to be mapped and fosters productivity, since complex mapping tasks can be carried out with fewer actions from the user (Figure 16).

Users can change the layout of the mapping representation by dragging the graph nodes, making the visualization clearer. Thanks to this feature they can create their own layouts. For example, ontology concepts can be positioned on top of the screen while tables and columns can be placed on the lower side. A point-and-click interface simplifies the mapping creation process. The interface provides suggestion lists of possible concepts, relations, and attributes to be used in the mappings. The editor automatically generates a R2RML document based on user inputs, producing specific IRI patterns and SQL queries.

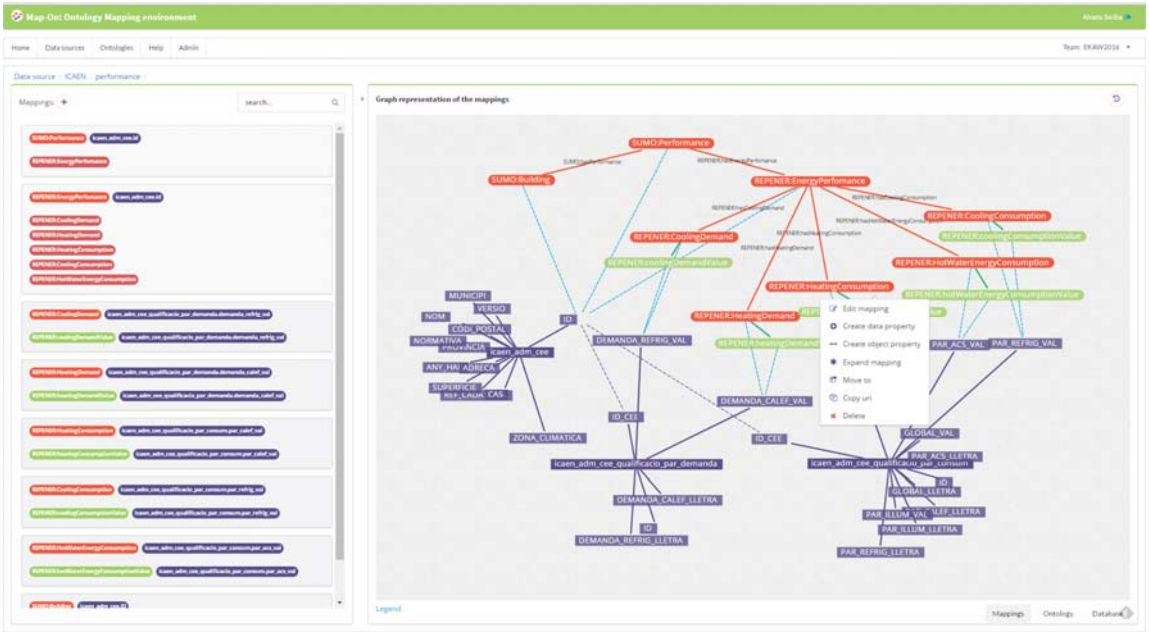


Figure 16. Map-On Interface.

Map-On implements the ontology-driven approach for editing the mappings. Namely, the user starts by selecting concepts of the ontology and subsequently generates R2RML statements by selecting elements of the relational source (i.e., columns) to obtain the proper IRI patterns and SQL queries. An alternative to the ontology driven approach is the database-driven approach which starts with selection of database elements followed by the generation of R2RML statements through selection of the proper domain ontology elements. As stated in (Pinkel et al., 2014), none of these approaches (i.e., ontology-driven and database-driven) is better. However, users with a background in database may be more familiar with the ontology-driven approach.

The Map-On editor automatically generates IRI patterns and logic tables (i.e., SQL queries) that are required by the R2RML statements. This is based on the concepts and columns included in the mappings created by the user. The IRI is generated using a patterned URIs solution (Dodds & Davis, 2012). This pattern was chosen because people are able to read it. The editor inspects the mappings created by the user for generating a valid SQL query which takes into account all the possible tables and columns considered in the mapping. For example, in Figure 16, when a user maps the concept *repener:EnergyPerformance* to the column *ID* of the table *icaen_adm_cee*, the following IRI and SQL query are generated for defining the subject map (Listing 6):

```
IRI: <base_iri>/energyperformance/{\"icaen_adm_cee.ID\"}
SQL: SELECT icaen_adm_cee.ID,
icaen_adm_cee_qualificacio_par_demanda.DEMANDA_REFRIG_VAL FROM
icaen_adm_cee JOIN icaen_adm_cee_qualificacio_par_demanda ON
icaen_adm_cee_qualificacio_par_demanda.ID_CEE = icaen_adm_cee.ID
```

Listing 6. Examples of an IRI and a SQL query for a triple map.

The Map-On features can be summarized as follows:

- Multiuser web environment for manual creation of relational-to-ontology mappings.
- Mapping spaces for distribution of the mapping creation process.
- Top-down visual representation of relational source schema, ontology structure, and mappings based on a graph layout which can be customised by users.
- Visual representation of an ontology using VOWL and a relational source based on Entity-Relationship diagrams.
- Input relational sources can be a SQL database or a tabular source such as comma separated values (CSV) file.
- Support of R2RML recommendation.
- R2RML documents generated by AutoMap4OBDA can be imported in Map-On.
- Automated generation of IRI patterns and SQL queries based on mappings defined by users.
- Dialog window in input boxes with suggestions of elements to be used in the mappings based on the text introduced by users.
- Point-and-click interface for reducing the effort required for mapping activities.
- Ontology-driven mapping approach, where the mapping process starts from the ontology instead of working with the database.
- Contextual menus to help users in mapping creation.
- Log of the activities carried out by users.
- Pop-ups with tips as an integrated help.

4.4 Evaluation and Results

A user study was conducted to validate the user performance of Map-On and to demonstrate that it can be used by non-ontology experts. The profile of the participants was similar. They were graduates and post-graduates, experts who knows the SQL language. However, participants did not have experience with the Semantic Web technologies (e.g., OWL, RDF, SPARQL, and R2RML). The user test was composed of a database and an ontology from the domain of conferences (e.g., authors, papers, committees, and reviews among others) which could be easily understood by the participants without teaching them basic concepts and their interrelations. The participants had to carry out three tasks, each of them involved mapping a class of the domain ontology and an element of the database. The tasks were to 1) relate authors, 2) relate authors with their submitted papers, and 3) relate conferences with their committees. The tasks were designed to increasingly carry out simple to complex mappings. The usability metrics to evaluate the results of the test were the effectiveness metric with the measure of accuracy – percentage of tasks correctly completed – and the efficiency metric with the completion time which is the time taken to complete the tasks.

Figure 17 summarizes the results for each usability metric obtained in the three tasks. The difference in the results was related to the complexity of the task. Task 1 involved the creation of a simple mapping while in Task 2 and Task 3 the participant had to create a complex mapping including relating different tables and columns. Additionally, the mappings created in Task 2 were based on the mappings produced in Task 1. The completion time for Task 1, was 5 minutes, 19.1 minutes for Task 2, and 3.7 minutes for Task 3. The difference between the completion time of Task 1 and 3 is due to the lack of knowledge of how to work with ontologies – in particular when referring to the creation of object properties to connect concepts of an ontology – and how to use the tool since participants were not trained beforehand. The completion time of Task 3 is similar to the time of Task 1 because participants had the chance to learn using the tool.

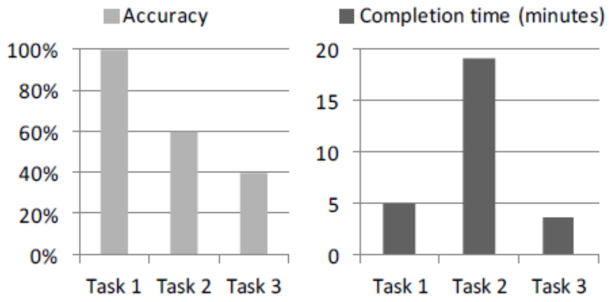


Figure 17. Accuracy and completion time results for the three tasks.

User satisfaction was measured with a post-test questionnaire based on the System Usability Scale (Brooke, 1996) whereby the participants rated some subjective statements with a five-item Likert scale (from 1-completely disagree to 5-completely agree). Results of the user satisfaction were obtained by calculating the mean of responses per answer (Figure 18). Ratings clearly show that Map-On is not perceived as a complex tool. Moreover, it demonstrated that most people could easily learn to use it. However, the participants neither agree nor disagree about feeling very confident using the tool. Finally, participants observed that a lot of prior knowledge is not required to use the tool.

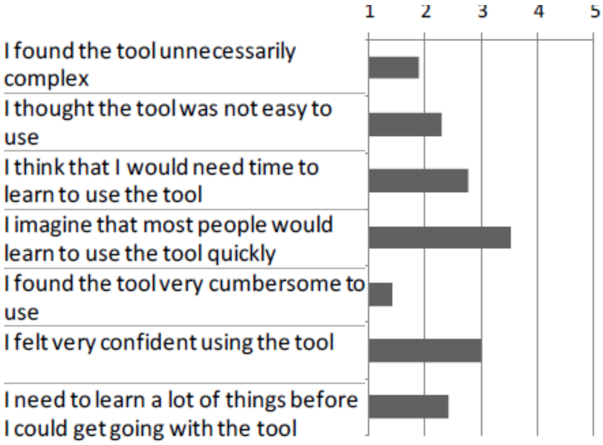
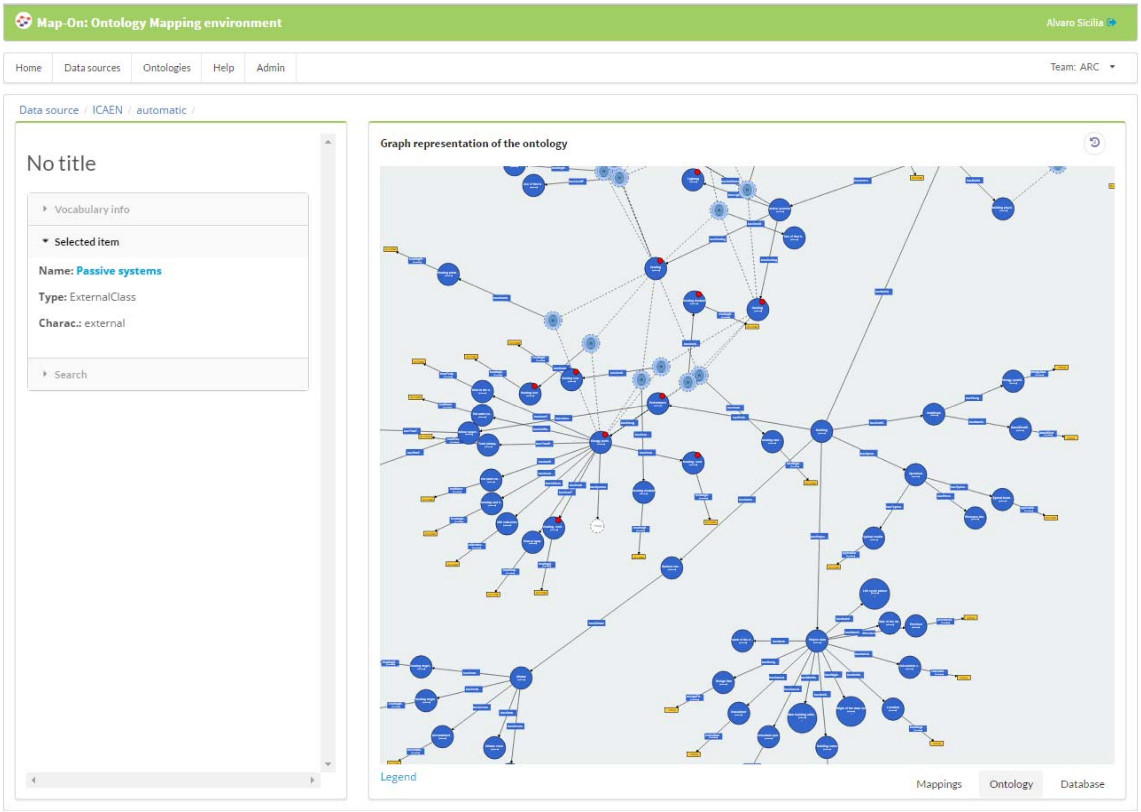


Figure 18. User satisfaction results.

Some observations were noted by the administrator during the test and by analysing the voice transcription and screen recording. Most of the participants stressed that the process of mapping creation should be more guided than it is now. The representation of the database schema in a graph basis was a bit confusing for some users since they are very used to working with relational views. All participants changed the layout of the mappings by dragging the graph nodes in a usual way. Some participants missed a visual representation of the whole ontology where they could see how the concepts are actually related. In order to address some of the issues raised in the user study a new tool iteration was developed to include a visual representation of the ontology using VOWL based on a force-directed graph layout (Lohmann et al., 2014) and the relational source using an basic Entity-Relationship diagram (Figure 19).



Name: Passive systems

Type: ExternalClass

Charac.: external

Figure 19. Visual representation of an ontology using VOWL.

The overall conclusion is that the Map-On editor can be used by non-ontology experts to manually generate mappings between a database and an ontology without previously acquiring a formal knowledge about several Semantic Web technologies such as OWL, RDF, and R2RML. Although the results of the complex tasks successfully completed were not as good as with the simple ones, the tool is easy to learn. That is, the completion times decreased in the last task. Moreover, these results could be improved with a previous training session where the tool is presented to the user in a detailed way. This was not done to avoid contaminating the test.

4.5 Discussion

Domain experts with missing skills in ontology development need visual support for relational-to-ontology mapping editing tools to help them to edit mappings. Visual representations of mappings and user-friendly interfaces help data owners and domain experts to establish mappings between a relational source and a domain ontology. Map-On supports those kinds of users thanks to a visual representation of mappings that takes into account the relational source schema, ontology structure and the mappings between their elements.

The manual creation of mappings is a difficult task which requires understanding the data source and the domain ontology. Moreover, having technical skills in native languages such as SQL, OWL, and R2RML is primordial. Editors like Map-On simplify the task of mapping, requiring only the selection of the proper elements of the data source and ontology to relate them. This way, users do not have to spend time on technical issues such as IRI and SQL generation. Thus, the visual representation of mappings based on a graph layout helps in understanding how the data source and ontology are structured. Indeed, users can personalise the layout of the mappings using a drag and drop feature. This feature is important because it does not restrict users to utilizing a fixed layout but enables them to modify the mapping layout according to their own interpretation of the data source and ontology structure. Different users who worked with Map-On (including participants of the user study) have modified the mapping layout according to their needs. Some of them followed the layout initially proposed by Map-On with slight modifications while others modified it significantly (e.g., elements of the data source in the right side and elements of the ontology in the left part).

The user study has confirmed that Map-On and its visual representation of relational-to-ontology mappings can help non-ontology experts to manually edit mappings. Some participants of the user study and colleagues who have used Map-On have asked for methods to automatically suggest mappings. To address this concern, mappings automatically generated by the techniques presented in the initial line of contributions of this thesis – *Automated generation of relational-to-ontology mappings* – in Section 3 can be loaded in the tool. This way, users can edit and complete those mappings. Map-On has been used in research projects (i.e., REPENER, SEMANCO and ENERSI) as technological solutions to implement energy information systems, in particular to integrate heterogeneous data based on the OBDA paradigm.

The outcomes of this research have been presented at:

- Sicilia, Á., Nemirovski, G., & Nolle, A. (2016). Map-On: A web-based editor for visual ontology mapping. *Semantic Web Journal*, –in press. Retrieved from http://www.semantic-web-journal.net/content/map-web-based-editor-visual-ontology-mapping-0

5 Conclusion and Further Work

This chapter summarizes the contributions of this research and suggests some objectives for future work. The conclusions of the two research lines are presented from a perspective less focused on results. Future lines of work that emerged from the conclusion of this research are described including the further work necessary to improve the performance of AutoMap4OBDA and to enhance the usability of Map-On.

5.1 Conclusion

Developing information systems which integrate data from multiple sources raises some challenges such as ensuring interoperability of systems by overcoming structural and semantic variety of data. OBDA is a comprehensive solution to address these challenges which relies on the use of ontologies as mediation schema for different data sources. However, one of the main barriers in the implementation of an OBDA system is the lack of tools to support the creation of mappings between data and ontologies. The development of semantic information systems in the domain of urban energy consumption in the RÉPENER and SEMANCO projects have required a substantial amount of human resources dedicated in particular to mapping creation. During the realisation of those projects the limitations of the OBDA technology became evident: the lack of an automated process for finding relations between elements of a relational source and an ontology, and the lack of visual representations of relational-to-ontology mappings that can help data owners and domain experts in the task of mapping creation. The research described in this thesis focused on the development of approaches to address these obstacles. The technological outputs of this research have been the **AutoMap4OBDA**[19] and **Map-On**[20] tools.

5.1.1 Automated Generation of Relational-to-Ontology Mappings

There is a need in real-world scenarios to integrate diverse data sources using Semantic Web technologies – such as decision-making related to increasing the energy efficiency of buildings – where relational sources and domain ontologies have been developed by independent teams focused on different purposes. Attempts at automating mappings between a relational source and a domain ontology have been conducted in previous research projects with comparatively poor performance results. One of the reasons for the bad performance is that current mapping generators basically rely on the relational schema and only barely take into account the contents of the relational data source and the features of the domain ontology.

To overcome this issue three approaches have been proposed in this research for relational-to-ontology mapping that are free of this disadvantage. The approaches are: an *Ontology learning technique* is applied to infer class hierarchies, the *String similarity metric selection technique* chooses the metric based on the domain ontology labels, and *Short path technique* applies graph structures to generate the mappings. Furthermore, these approaches have been implemented in the AutoMap4OBDA system, a full-featured R2RML mapping generator for OBDA scenarios. RODI benchmarking suite has been used to evaluate AutoMap4OBDA which outperforms the most advanced existing state of the art mapping generators.

The contribution of this research in the area of *automated generation of relational-to-ontology mappings* has provided a new kind of techniques that use contents of relational sources and features of ontologies. AutoMap4OBDA clearly outperforms the existing mapping generators. However, the performance of these techniques in real-world scenarios – such as *GeoData* and *Oil&Gas* scenarios of RODI benchmark – is lower compared to the scenarios using synthetic data.

As result of this research, we can conclude that the generation of mappings between a relational source and a domain ontology – which have been created independently from the source – cannot be fulfilled solely by a fully automated tool. The resulting mappings should be validated and amended by experts who are aware on the one hand of the structure of the data source and on the other hand of the domain knowledge described by the ontology. Despite this, there is still room to improve and extend the current techniques to enhance mappings in real-world scenarios.

[19] http://arc.salleurl.edu/automap4obda/

[20] http://semanco-tools.eu/map-on

5.1.2 Visual Support for Relational-to-Ontology Mapping Editing

As stated above the creation of relational-to-ontology mappings is a time consuming task, even though automated mappings generators can obtain around 45% of the mappings. In practice, the remaining mappings are created manually by users who are aware of the data source schema as well as having expertise in the domain described by the ontology. Moreover, such users should have knowledge of mapping languages (e.g., R2RML), advanced technical skills (e.g., SQL), and expertise in ontology design. It seldom occurs that these compound expertises are owned by a single user. Rather it is a team of experts who is responsible for the development of relational-to-ontology mappings. Such a team may involve data owners, domain experts and ontology designers. To assist those users in specifying relational-to-ontology mappings, several tools have been developed. These tools have certain limitations concerning visual representation of the mappings such as the structure of the relational source not being included in the mapping representation and the lack of a complete overview of the all mappings at once.

To overcome these limitations, an innovative visual representation of relational-to-ontology mappings has been developed that helps data owners and domain experts to understand the structure of the relational data sources, ontology and the mappings between them. The visual representation is based on a graph layout where elements of relational sources and domain ontology are represented as nodes and their relations as edges. Moreover, this representation has been implemented in the Map-On editor – a second release of the tool developed in the SEMANCO project – to support data owners, domain experts, and ontology engineers in the task of mapping editing. Map-On is a multiuser graphical web environment for the manual creation of mappings. R2RML mappings generated by AutoMap4OBDA system can be imported into Map-On, modified and extended by users using the visual representation of mappings mentioned above.

Map-On has been validated by its application in the data integration process of the ENERSI and SEMANCO projects. The graphical visualization of the mappings helped users to understand, evaluate, and correct the mappings. The Map-On editor can be used by non-ontology experts to manually generate mappings between a database and an ontology without acquiring a formal knowledge about the Semantic Web technologies such as OWL, RDF, and R2RML. Map-On is generic enough to be applied in other OBDA scenarios.

5.2 Further Work

The contribution of this research lies in the development of new techniques and tools to support non-ontology experts in the process of relational-to-ontology mapping creation with the ultimate goal being to integrate data sources for their information systems. The two research lines presented in this document represent a step forward in the automated generation of mappings and in the visual editing of mappings. However, there are several issues still pending.

- In the research line for automated generation of mappings, the real-world scenarios are hard to solve. Mainly, these scenarios require tailored queries which can only be created by users with extensive knowledge about the domain of discourse. The *Ontology learning technique* proposed in this thesis can be enhanced to generate sophisticated mappings which include those types of queries. For example, the enhanced techniques will be able to identify patterns and acronyms in names of tables and columns used by data source developers. Moreover, the current techniques are dependent on the existence of explicit relations between tables using foreign keys. In some cases those relations are not coded in the database but established in the queries by data owners. Therefore, mapping generators can identify those relations by analysing the values of the database following Subclass identification techniques (de Medeiros et al., 2015).
- Furthermore, an extension to the RODI benchmark – including real-world scenarios similar to *Oil&Gas domain* scenario described in Section 3.3 – will help developers of automated mapping generators to improve their techniques. The main feature to be included in AutoMap4OBDA is a translation module to address multi-language scenarios. This specifically implies to extending the *String similarity metric selection technique* with a language detector and translation mechanism. Moreover, semantic similarity techniques based on external resources will need to be explored to increase the performance of AutoMap4OBDA. Furthermore, future versions of the system will support different relational database management systems since the current version can work only with PostgreSQL. The *Ontology learning technique* will be modified to support the different particularities of the SQL syntax of each relational database management systems. Furthermore, the queries included in the R2RML document have to be accordingly adapted.
- The mapping environment – including the visual representation of relational-to-ontology mappings – developed in the course of this research does not entirely take into consideration the maintenance of mappings. That is, the versioning of mappings and annotations are not features of Map-On. How to visualise changes among different versions of mappings, and making those differences understandable to non-technician users is a pending issue that can be addressed as a continuation of this research. A visual representation of mappings that considers changes between different versions of mappings is an open issue. The work of Hascöet and Dragicevic (2012) in visual comparison of graphs can be adapted to represent different versions of mappings. Moreover, the functionality requirements identified by Lambrix et al. for ontology evolution systems can be considered to define visual representations for versioning mappings (Lambrix, Dragisic, Ivanova, & Anslow, 2016). This is particularly important in those scenarios where different users are working together with large relational sources and ontologies. The main feature to include in a future development of Map-On is the conditional mappings. For example, a mapping would be applicable only if certain conditions are met like the value of an attribute being greater than a particular number. This new feature implies a visual representation as well as the generation of the proper SQL queries for the R2RML statements. Another

enhancement of Map-On would be to integrate an existing OBDA system – such as morph-RDB (Priyatna, Corcho, & Sequeda, 2014) – for executing and evaluating the mappings generated by the editor. This way, end-users will see in real-time the RDF data according to the mappings that they have created. Thus, by exporting R2RML documents with Map-On properties (e.g., mapping spaces, position of the elements, user who created the mappings, and creation dates among others) it would be possible to share the mappings created by Map-On among different teams and colleagues.

- The research lines of this thesis provide techniques and tools to support the mapping process of a relational sources and ontologies. However, there is a lot of data which is available in non-relational sources such as XML and JSON. Providing support for non-relational data sources is an ambitious research line for both AutoMap4OBDA and Map-On. This requires the addition into AutoMap4OBDA of other mechanisms to extract class hierarchies from non-relational sources using *Ontology learning technique* as well as an extension of the *Short path strategy technique* to handle non-relational sources. With regard to Map-On, visual representations for those data sources have to be devised to include tree-layout based sources. Supporting heterogeneous data sources will lead to using an alternative mapping language such as RDF mapping language (RML), a generic mapping language defined to express customized mapping rules from heterogeneous data structures and serializations to the RDF data model (Dimou et al., 2014).

Bibliography

Abiteboul, S., Hull, R., & Vianu, V. (1995). *Foundations of databases*. Addison-Wesley Publishing Company.

Aumueller, D., Do, H.-H., Massmann, S., & Rahm, E. (2005). Schema and Ontology Matching with COMA++. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data SIGMOD 05* (pp. 906–908). New York, New York, USA: ACM Press. http://doi.org/10.1145/1066157.1066283

Bizer, C., & Cyganiak, R. (2007). D2RQ — Lessons Learned. In *W3C Workshop on RDF Access to Relational Databases* (pp. 1–10).

Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse. Technology* (Vol. PhD). Retrieved from http://doc.utwente.nl/17864/

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7. article.

Calvanese, D., Cogrel, B., Komla-ebri, S., Kontchakov, R., & Lanti, D. (2015). Ontop: Answering SPARQL Queries over Relational Databases. *Semantic Web Journal*, *Preprint*(Preprint), 1–17. http://doi.org/10.3233/SW-160217

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., … Savo, D. F. (2011). The MASTRO system for ontology-based data access. *Semantic Web*, *2*(1), 43–53. http://doi.org/10.3233/SW-2011-0029

Calvanese, D., Giacomo, G. De, Lembo, D., Lenzerini, M., Poggi, A., & Rosati, R. (2007). Ontology-Based Database Access. In *Proceedings of the 15th Italian Symposium on Advanced Database Systems* (pp. 324–331). Torre Canne di Fasano, BR, Italy.

Cerbah, F. (2008). Mining the content of relational databases to learn ontologies with deeper taxonomies. In *Proceedings - 2008 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2008* (pp. 553–557). Sydney, Australia: IEEE. http://doi.org/10.1109/WIIAT.2008.382

Cheatham, M., & Hitzler, P. (2013). String Similarity Metrics for Ontology Alignment. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference* (pp. 294–309). Sydney, NSW Australia: Springer. http://doi.org/0.1007/978-3-642-41338-4_19

Cheatham, M., & Hitzler, P. (2014). The properties of property alignment. In *Proceedings of the 9th International Workshop on Ontology Matching* (Vol. 1317, pp. 13–24). Riva del Garda, Trentino: CEUR-WS. org.

Corrado, V., Ballarini, I., Madrazo, L., & Nemirovskij, G. (2015). Data structuring for the ontological modelling of urban energy systems: The experience of the SEMANCO project. *Sustainable Cities and Society*, *14*(1), 223–235.

http://doi.org/10.1016/j.scs.2014.09.006

Corrado, V., Corgnati, S. P., & Garbino, M. (2007). Energy Consumption Data Collection with DATAMINE. In *Energy, Climate and Indoor Comfort in Mediterranean Countries* (pp. 803–816). Genova, Italy. http://doi.org/9788895620022

de Medeiros, L. F., Priyatna, F., & Corcho, O. (2015). MIRROR: Automatic R2RML Mapping Generation from Relational Databases. In *Engineering the Web in the Big Data Era: 15th International Conference, ICWE 2015* (pp. 326–343). Rotterdam, The Netherlands: Springer International Publishing. http://doi.org/10.1007/978-3-319-19890-3_21

Dimou, A., Sande, M. Vander, Colpaert, P., Verborgh, R., Mannens, E., & Van De Walle, R. (2014). RML: A generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference* (Vol. 1184). Seoul, Korea: CEUR-WS.org.

Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of Data Integration*. (A. Doan, A. Halevy, & Z. Ives, Eds.). Boston: Morgan Kaufmann. http://doi.org/10.1016/B978-0-12-416044-6.00004-1

Dodds, L., & Davis, I. (2012). *A pattern catalogue for modelling, publishing, and consuming Linked Data*. Retrieved from http://patterns.dataincubator.org/book/linked-data-patterns.pdf

Fu, B., Noy, N. F., & Storey, M.-A. (2013). Indented Tree or Graph? A Usability Study of Ontology Visualization Techniques in the Context of Class Mapping Evaluation. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, … K. Janowicz (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8218 LNCS, pp. 117–134). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-41335-3_8

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. http://doi.org/10.1.1.101.7493

Hascoët, M., & Dragicevic, P. (2012). Interactive graph matching and visual comparison of graphs and clustered graphs. In *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12* (p. 522). New York, USA: ACM Press. http://doi.org/10.1145/2254556.2254654

Heyvaert, P., Dimou, A., Herregodts, A.-L., Verborgh, R., Schuurman, D., Mannens, E., & Van de Walle, R. (2016). RMLEditor: A Graph-Based Mapping Editor for Linked Data Mappings. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016* (Vol. 9088, pp. 709–723). Heraklion, Crete: Springer International Publishing. http://doi.org/10.1007/978-3-319-34129-3_43

Jiménez-Ruiz, E., & Cuenca Grau, B. (2011). LogMap: Logic-Based and Scalable Ontology Matching. In *Lecture Notes in Computer Science (including subseries Lecture Notes in*

Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 7031 LNCS, pp. 273–288). http://doi.org/10.1007/978-3-642-25073-6_18

Jiménez-Ruiz, E., Grau, B. C., Zhou, Y., & Horrocks, I. (2012). Large-scale interactive ontology matching: Algorithms and implementation. In *20th European Conference on Artificial Intelligence* (Vol. 242, pp. 444–449). Montpellier, France: IOS Press. http://doi.org/10.3233/978-1-61499-098-7-444

Jiménez-Ruiz, E., Kharlamov, E., Zheleznyakov, D., Horrocks, I., Pinkel, C., Skjæveland, M. G., … Mora, J. (2015). BootOX: Practical Mapping of RDBs to OWL 2. In *The Semantic Web - ISWC 2015: 14th International Semantic Web Conference* (pp. 113–132). Bethlehem, PA, USA: Springer International Publishing. http://doi.org/10.1007/978-3-319-25010-6_7

Knoblock, C. A., Szekely, P., Ambite, J. L., Goel, A., Gupta, S., Lerman, K., … Mallick, P. (2012). Semi-automatically mapping structured sources into the semantic web. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7295 LNCS, pp. 375–390). http://doi.org/10.1007/978-3-642-30284-8_32

Lambrix, P., Dragisic, Z., Ivanova, V., & Anslow, C. (2016). Visualization for Ontology Evolution. In *Proceedings of the International Workshop on Visualizations and User Interfaces for Ontologies and Linked Data co-located with 15th International Semantic Web Conference (ISWC 2016)*. Kobe, Japan: Springer.

Lanzenberger, M., Sampson, J., & Rester, M. (2009). Visualization in Ontology Tools. In *2009 International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 705–711). Fukuoka, Japan: IEEE. http://doi.org/10.1109/CISIS.2009.178

Lembo, D., Rosati, R., Ruzzi, M., Savo, D. F., & Tocci, E. (2014). Visualization and Management of Mappings in Ontology-based Data Access (Progress Report). In *Informal Proceedings of the 27th International Workshop on Description Logics* (pp. 595–607). Vienna, Austria: CEUR-WS. org.

Lohmann, S., Negru, S., Haag, F., & Ertl, T. (2014). VOWL 2: User-Oriented Visualization of Ontologies. In *Knowledge Engineering and Knowledge Management: 19th International Conference* (pp. 266–281). Linköping, Sweden: Springer International Publishing. http://doi.org/10.1007/978-3-319-13704-9_21

Madrazo, L., Massetti, M., Sicilia, Á., Wadel, G., & Ianni, M. (2015). SEíS: A semantic-based system for integrating building energy data. *Informes de La Construcción*, *67*(537). http://doi.org/10.3989/ic.13.048

Madrazo, L., Sicilia, Á., & Nemirovski, G. (2013). Shared Vocabularies to Support the Creation of Energy Urban Systems Models. In *4th Workshop organised by the EEB data models community ICT for Sustainable Places* (pp. 130–150). Nice, France: Publications Office of the European Union. http://doi.org/10.2759/40897

Nemirovski, G., Nolle, A., Sicilia, Á., Ballarini, I., & Corado, V. (2013). Data integration driven ontology design, case study smart city. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* (pp. 43–52). Madrid,: ACM Press. http://doi.org/10.1145/2479787.2479830

Nemirovski, G., Sicilia, Á., Galán, F., Massetti, M., & Madrazo, L. (2012). Ontological Representation of Knowledge Related to Building Energy-efficiency. In *Sixth International Conference on Advances in Semantic Processing* (pp. 20–27). Barcelona, Spain: IARIA XPS Press.

Neto, L. E. T., Vidal, V. M. P., Casanova, M. A., & Monteiro, J. M. (2013). R2RML by Assertion: A Semi-automatic Tool for Generating Customised R2RML Mappings. In *Extended Semantic Web Conference* (pp. 248–252). Montpellier, France: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-41242-4_33

Nolle, A., & Nemirovski, G. (2013). ELITE: An entailment-based federated query engine for complete and transparent semantic data integration. In *Informal Proceedings of the 26th International Workshop on Description Logics* (Vol. 1014, pp. 854–867). Ulm, Germany: CEUR-WS. org.

Pinkel, C., Binnig, C., Haase, P., Martin, C., Sengupta, K., & Trame, J. (2014). How to best find a partner? An evaluation of editing approaches to construct R2RML mappings. In *European Semantic Web Conference* (pp. 675–690). Anissaras, Crete, Greece: Springer International Publishing. http://doi.org/10.1007/978-3-319-07443-6_45

Pinkel, C., Binnig, C., Jimenez-Ruiz, E., Kharlamov, E., May, W., Nikolov, A., … Horrocks, I. (2016). RODI: Bench-marking Relational-to-Ontology Mapping Generation Quality. *Semantic Web Journal*, *in press*. Retrieved from http://www.semantic-web-journal.net/content/rodi-benchmarking-relational-ontology-mapping-generation-quality-0

Pinkel, C., Binnig, C., Jiménez-Ruiz, E., May, W., Ritze, D., Skjæveland, M. G., … Kharlamov, E. (2015). RODI: a benchmark for automatic mapping generation in relational-to-ontology data integration. In *Proceedings of the 12th European Semantic Web Conference on The Semantic Web. Latest Advances and New Domains* (pp. 21–37). Portoroz, Slovenia: Springer-Verlag. http://doi.org/10.1007/978-3-319-18818-8_2

Pinkel, C., Binnig, C., Kharlamov, E., & Haase, P. (2013). IncMap: Pay-as-you-go matching of relational schemata to OWL ontologies with IncMap. In *8th International Conference on Ontology Matching* (Vol. 1111, pp. 37–48). Sydney, Australia: CEUR-WS. org.

Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., & Rosati, R. (2008). Linking Data to Ontologies. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4900 LNCS, pp. 133–173). http://doi.org/10.1007/978-3-540-77688-8_5

Priyatna, F., Corcho, O., & Sequeda, J. (2014). Formalisation and experiences of R2RML-based SPARQL to SQL query translation using morph. In *Proceedings of the 23rd international conference on World wide web - WWW '14* (pp. 479–490). New York, USA: ACM Press. http://doi.org/10.1145/2566486.2567981

Priyatna, F., Villazón-Terrazas, B., Barrasa, J., & Schulte, J. (2011). ODEMapster. Retrieved June 5, 2016, from http://neon-toolkit.org/wiki/ODEMapster

Rodríguez-Muro, M., & Rezk, M. (2015). Efficient SPARQL-to-SQL with R2RML mappings. *Web Semantics: Science, Services and Agents on the World Wide Web*, *33*, 141–169. http://doi.org/10.1016/j.websem.2015.03.001

Savo, D. F., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Romagnoli, V., … Stella, G. (2010). Mastro at work: Experiences on ontology-based data access. In *Proceedings of the 2010 International Workshop on Description Logics* (Vol. 573, pp. 20–31). Waterloo, Ontario, Canada: CEUR-WS. org.

Sengupta, K., Haase, P., Schmidt, M., & Hitzler, P. (2013). Editing R2RML Mappings Made Easy. In *Proceedings of the ISWC 2013 Posters & Demonstrations Track a track within the 12th International Semantic Web Conference* (pp. 101–104). Sydney, Australia: CEUR-WS. org.

Sequeda, J. F., Garcia-Castro, A., Corcho, O., Miranker, D. P., & Tirmizi, S. H. (2009). Overcoming database heterogeneity to facilitate social networks: the Colombian displaced population as a case study. In *18th International World Wide Web Conference*. Madrid, Spain. Retrieved from http://www2009.eprints.org/207/

Sequeda, J. F., Tirmizi, S. H., & Miranker, D. P. (2008). A Bootstrapping Architecture for Integration of Relational Databases to the Semantic Web. In *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008)*. Karlsruhe, Germany: CEUR-WS. org.

Sheth, A. P., & Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, *22*(3), 183–236. http://doi.org/10.1145/96602.96604

Sicilia, Á., Madrazo, L., & Pleguezuelos, J. (2015). Integrating multiple data sources, domains and tools in urban energy models using semantic technologies. In *eWork and eBusiness in Architecture, Engineering and Construction - Proceedings of the 10th European Conference on Product and Process Modelling, ECPPM 2014* (pp. 837–844). CRC Press/Balkema.

Sicilia, Á., & Nemirovski, G. (2016). AutoMap4OBDA: Automated Generation of R2RML Mappings for OBDA. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings* (pp. 577–592). Bologna, Italy: Springer International Publishing. http://doi.org/10.1007/978-3-319-49004-5_37

Sicilia, Á., Nemirovski, G., Massetti, M., & Madrazo, L. (2015). The RÉPENER linked dataset. *Semantic Web*, *6*(2), 131–137. http://doi.org/10.3233/SW-130131

Sicilia, Á., Nemirovski, G., & Nolle, A. (2016). Map-On: A web-based editor for visual ontology mapping. *Semantic Web*, *in press*. Retrieved from http://www.semantic-web-journal.net/content/map-web-based-editor-visual-ontology-mapping-0

Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., & Gangemi, A. (2012). *Ontology Engineering in a Networked World*. (M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, & A. Gangemi, Eds.). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-24794-1

W3C. (2009). OWL 2 Web Ontology Language Document Overview. Retrieved from https://www.w3.org/TR/owl2-overview/

W3C. (2012). R2RML: RDB to RDF Mapping Language. Retrieved from http://www.w3.org/TR/r2rml/

W3C. (2013). SPARQL Query Language for RDF. http://doi.org/citeulike-article-id:2620569

W3C. (2014). Resource Description Framework (RDF): Concepts and Abstract Syntax. Retrieved from https://www.w3.org/TR/rdf11-concepts/

Wolters, M., Nemirovski, G., & Nolle, A. (2013). ClickOnA: An editor for DL-liteA based ontology design. In *Informal Proceedings of the 26th International Workshop on Description Logics* (Vol. 1014, pp. 377–389). Ulm, Germany: CEUR-WS. org.

Appendix

A. Abbreviations

ARC	Architecture, Representation, and Computation research group
CEN	European Committee for Standardization
CNIG	Geo-graphical Information National Institute
CO_2	Carbon Dioxide
CSV	Comma Separated Values
ETL	Extract, Transform, and Load
HVAC	Heating, Ventilation, and Air Conditioning
ICAEN	Institut Català d'Energia
ICT	Information and Communication Technologies
IRI	Internationalized Resource Identifier
ISO	International Standards Organization
JSON	JavaScript Object Notation
OBDA	Ontology-Based Data Access
OWL	Web Ontology Language
R2RML	RDB to RDF Mapping Language
RDF	Resource Description Framework
RML	RDF Mapping Language
SEIF	Semantic Energy Information Framework
SEíS	Semantic Energy Information System
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VOWL	Visual Notation for OWL Ontologies
W3C	World Wide Web Consortium
XML	Extensible Markup Language

B. Example of R2RML mappings generated by Short path strategy technique

The following mappings have been generated through the *Short path strategy technique* for the examples of Figure 11.

Mapping for issue 1:

```
<mapping1> a rr:TriplesMap;
     rr:logicalTable [  rr:sqlQuery "SELECT Buildings.id FROM
Buildings”;
     rr:subjectMap [    rr:template "…/building/{Buildings.id}";
                        rr:class ex:Building
                   ];

     rr:predicateObjectMap [
          rr:predicate  ex:hasRoom ;
          rr:objectMap [ rr:template "…/room/{Buildings.id}" ]
                   ];
     .

<mapping2> a rr:TriplesMap;
     rr:logicalTable [  rr:sqlQuery "SELECT Buildings.id FROM
Buildings”;
     rr:subjectMap [    rr:template "…/room/{Buildings.id}";
                        rr:class ex:Room
                   ];

     rr:predicateObjectMap [
          rr:predicate  ex:hasWall ;
          rr:objectMap [ rr:template "…/wall/{Buildings.id}" ]
                   ];
     .

<mapping3> a rr:TriplesMap;
     rr:logicalTable [  rr:sqlQuery "SELECT Buildings.id, Windows.id
                     FROM Buildings JOIN Windows ON Buildings.fkWindow
                     = Windows.id”;
     rr:subjectMap [    rr:template "…/wall/{Buildings.id}";
                        rr:class ex:Wall
                   ];

     rr:predicateObjectMap [
          rr:predicate  ex:hasWindow ;
          rr:objectMap [ rr:template "…/wall/{Windows.id}" ]
                   ];
     .

<mapping4> a rr:TriplesMap;
     rr:logicalTable [  rr:sqlQuery "SELECT Windows.id FROM Windows”;
     rr:subjectMap [    rr:template "…/window/{ Windows.id}";
                        rr:class ex:Windows
                   ];
     .
```

Mappings for issue 2:

```
<mapping1> a rr:TriplesMap;
     rr:logicalTable [  rr:sqlQuery "SELECT Buildings.id, Cities.id FROM
                        Buildings JOIN Blocks ON Buildings.fkBlock =
                        Blocks.id JOIN Districts ON Blocks.fkDistrict =
                        District.id JOIN Cities ON District.fkDistrict =
                        Cities.id”;
     rr:subjectMap [    rr:template "…/building/{Buildings.id}";
                        rr:class ex:Wall
                   ];

     rr:predicateObjectMap [
          rr:predicate  ex:hasCity ;
          rr:objectMap [ rr:template "…/city/{Cities.id}" ]
                   ];
     .
<mapping2> a rr:TriplesMap;
     rr:logicalTable [  rr:sqlQuery "SELECT Cities.id FROM Cities”;
     rr:subjectMap [    rr:template "…/city/{ Cities.id}";
                        rr:class ex:City
                   ];
     .
```

C. Scientific contributions of the doctoral student

Madrazo, L., Massetti, M., **Sicilia, Á.**, Wadel, G., & Ianni, M. (2015). SEíS: A semantic-based system for integrating building energy data. *Informes de La Construcción*, *67*(537). http://doi.org/10.3989/ic.13.048

This article describes the methodology applied to create the semantic energy information system of the RÉPENER research project using Semantic Web technologies. The PhD Student devised and implemented the technological solution for managing energy-related data. This included the participation in the development of the project's ontology.

Sicilia, Á., Nemirovski, G., Massetti, M., & Madrazo, L. (2015). The RÉPENER linked dataset. *Semantic Web*, *6*(2), 131–137. http://doi.org/10.3233/SW-130131

This article describes the RÉPENER linked dataset which was one of the outcomes of the RÉPENER research project. Data from the Spanish territory regarding energy certification, building monitoring, and geographical data had been integrated using Semantic Web technologies. The work of the PhD student has been to devise and implement the tools and methods to integrate the different data sources.

Madrazo, L., **Sicilia, Á.**, & Nemirovski, G. (2013). Shared Vocabularies to Support the Creation of Energy Urban Systems Models. In *4th Workshop organised by the EEB data models community ICT for Sustainable Places* (pp. 130–150). Nice, France: Publications Office of the European Union. http://doi.org/10.2759/40897

This article describes the methodology and development of the semantic energy information framework of the SEMANCO research project which facilitates the link between the tools and the energy-related data to support decision making in energy efficient urban planning. The PhD student participated in the design of the methodology to capture experts' knowledge. Moreover, the PhD student participated in the development of the semantic energy information framework.

Nemirovski, G., Nolle, A., **Sicilia, Á.**, Ballarini, I., & Corado, V. (2013). Data integration driven ontology design, case study smart city. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* (p. 43-52). Madrid, Spain: ACM Press. http://doi.org/10.1145/2479787.2479830

This article describes 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. The contribution of the PhD student has been to describe the case study of weather domain. He participated in the definition of the methodology with special focus on the steps 3. Data Sources' Vocabularies Mappings and 5. Mapping Data sources.*

Sicilia, Á., & Nemirovski, G. (2016). AutoMap4OBDA: Automated Generation of R2RML Mappings for OBDA. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings* (pp. 577–592). Bologna, Italy: Springer International Publishing. http://doi.org/10.1007/978-3-319-49004-5_37

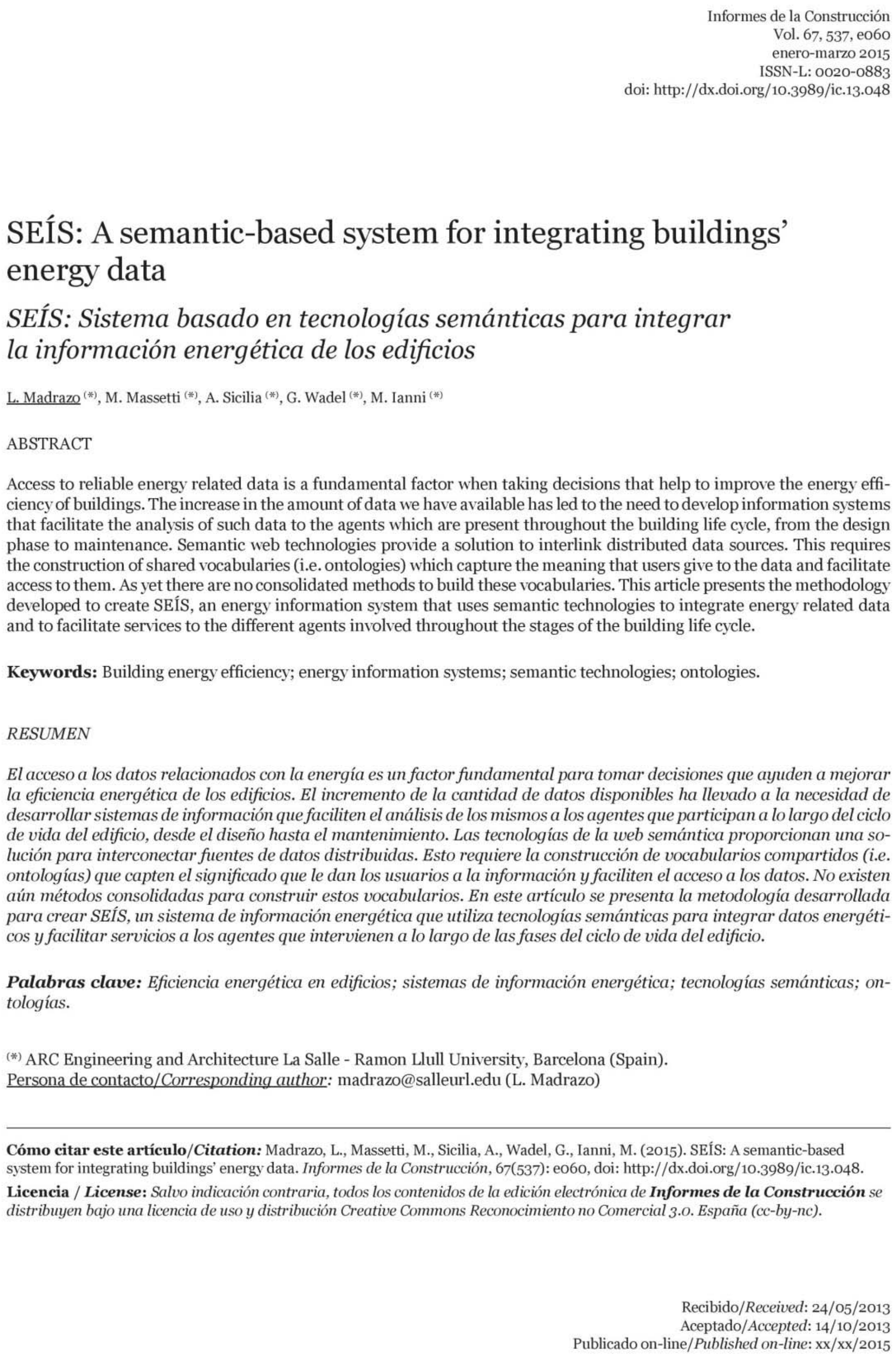
This article describes the techniques developed for generating automatically R2RML mappings between a relational source and an ontology. In the article is presented AutoMap4OBDA, a system which automatically generates R2RML mappings based on the intensive use of relational source contents and features of the target ontology. The PhD student has devised and implemented the different techniques and he has integrated them in the AutoMap4OBDA tool. The PhD student has carried out the evaluation with the RODI benchmark suite.

Sicilia, Á., Nemirovski, G., & Nolle, A. (2016). Map-On: A web-based editor for visual ontology mapping. *Semantic Web Journal*, –in press. Retrieved from http://www.semantic-web-journal.net/content/map-web-based-editor-visual-ontology-mapping-0

This article presents Map-On, a web-based editor for visual ontology mapping. The Map-On editor provides a graphical environment for the ontology mapping creation using an interactive graph layout. A point-and-click interface simplifies the map-ping creation process. The editor automatically generates a R2RML document based on user inputs, particularly producing IRI patterns and SQL queries. The work of the PhD student has been to devise and implement the Map-On tool. Moreover, the visual representation of the mappings used in Map-On have been devised and implemented by the PhD student. Moreover, he has carried out the user study to evaluate the usability of the tool.

Publications

SEíS: A semantic-based system for integrating building energy data. *Informes de la Construcción* , 2015



Informes de la Construcción
Vol. 67, 537, e060
enero-marzo 2015
ISSN-L: 0020-0883
doi: http://dx.doi.org/10.3989/ic.13.048

SEÍS: A semantic-based system for integrating buildings' energy data

SEÍS: Sistema basado en tecnologías semánticas para integrar la información energética de los edificios

L. Madrazo [*], M. Massetti [*], A. Sicilia [*], G. Wadel [*], M. Ianni [*]

ABSTRACT

Access to reliable energy related data is a fundamental factor when taking decisions that help to improve the energy efficiency of buildings. The increase in the amount of data we have available has led to the need to develop information systems that facilitate the analysis of such data to the agents which are present throughout the building life cycle, from the design phase to maintenance. Semantic web technologies provide a solution to interlink distributed data sources. This requires the construction of shared vocabularies (i.e. ontologies) which capture the meaning that users give to the data and facilitate access to them. As yet there are no consolidated methods to build these vocabularies. This article presents the methodology developed to create SEÍS, an energy information system that uses semantic technologies to integrate energy related data and to facilitate services to the different agents involved throughout the stages of the building life cycle.

Keywords: Building energy efficiency; energy information systems; semantic technologies; ontologies.

RESUMEN

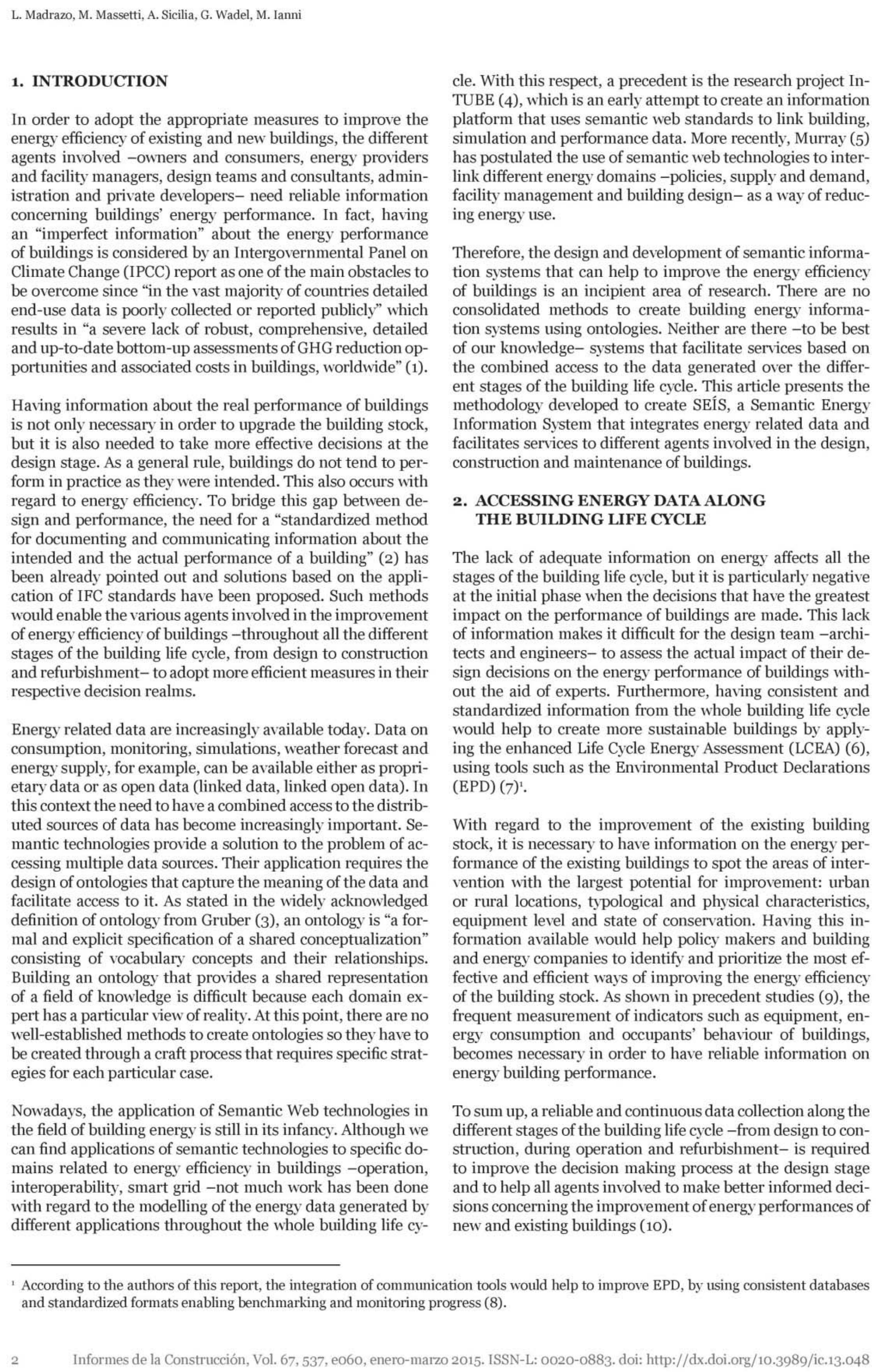
El acceso a los datos relacionados con la energía es un factor fundamental para tomar decisiones que ayuden a mejorar la eficiencia energética de los edificios. El incremento de la cantidad de datos disponibles ha llevado a la necesidad de desarrollar sistemas de información que faciliten el análisis de los mismos a los agentes que participan a lo largo del ciclo de vida del edificio, desde el diseño hasta el mantenimiento. Las tecnologías de la web semántica proporcionan una solución para interconectar fuentes de datos distribuidas. Esto requiere la construcción de vocabularios compartidos (i.e. ontologías) que capten el significado que le dan los usuarios a la información y faciliten el acceso a los datos. No existen aún métodos consolidadas para construir estos vocabularios. En este artículo se presenta la metodología desarrollada para crear SEÍS, un sistema de información energética que utiliza tecnologías semánticas para integrar datos energéticos y facilitar servicios a los agentes que intervienen a lo largo de las fases del ciclo de vida del edificio.

Palabras clave: *Eficiencia energética en edificios; sistemas de información energética; tecnologías semánticas; ontologías.*

[*] ARC Engineering and Architecture La Salle - Ramon Llull University, Barcelona (Spain).
Persona de contacto/*Corresponding author:* madrazo@salleurl.edu (L. Madrazo)

Cómo citar este artículo/*Citation:* Madrazo, L., Massetti, M., Sicilia, A., Wadel, G., Ianni, M. (2015). SEÍS: A semantic-based system for integrating buildings' energy data. *Informes de la Construcción*, 67(537): e060, doi: http://dx.doi.org/10.3989/ic.13.048.

Recibido/*Received*: 24/05/2013
Aceptado/*Accepted*: 14/10/2013
Publicado on-line/*Published on-line*: xx/xx/2015



1. INTRODUCTION

In order to adopt the appropriate measures to improve the energy efficiency of existing and new buildings, the different agents involved –owners and consumers, energy providers and facility managers, design teams and consultants, administration and private developers– need reliable information concerning buildings' energy performance. In fact, having an "imperfect information" about the energy performance of buildings is considered by an Intergovernmental Panel on Climate Change (IPCC) report as one of the main obstacles to be overcome since "in the vast majority of countries detailed end-use data is poorly collected or reported publicly" which results in "a severe lack of robust, comprehensive, detailed and up-to-date bottom-up assessments of GHG reduction opportunities and associated costs in buildings, worldwide" (1).

Having information about the real performance of buildings is not only necessary in order to upgrade the building stock, but it is also needed to take more effective decisions at the design stage. As a general rule, buildings do not tend to perform in practice as they were intended. This also occurs with regard to energy efficiency. To bridge this gap between design and performance, the need for a "standardized method for documenting and communicating information about the intended and the actual performance of a building" (2) has been already pointed out and solutions based on the application of IFC standards have been proposed. Such methods would enable the various agents involved in the improvement of energy efficiency of buildings –throughout all the different stages of the building life cycle, from design to construction and refurbishment– to adopt more efficient measures in their respective decision realms.

Energy related data are increasingly available today. Data on consumption, monitoring, simulations, weather forecast and energy supply, for example, can be available either as proprietary data or as open data (linked data, linked open data). In this context the need to have a combined access to the distributed sources of data has become increasingly important. Semantic technologies provide a solution to the problem of accessing multiple data sources. Their application requires the design of ontologies that capture the meaning of the data and facilitate access to it. As stated in the widely acknowledged definition of ontology from Gruber (3), an ontology is "a formal and explicit specification of a shared conceptualization" consisting of vocabulary concepts and their relationships. Building an ontology that provides a shared representation of a field of knowledge is difficult because each domain expert has a particular view of reality. At this point, there are no well-established methods to create ontologies so they have to be created through a craft process that requires specific strategies for each particular case.

Nowadays, the application of Semantic Web technologies in the field of building energy is still in its infancy. Although we can find applications of semantic technologies to specific domains related to energy efficiency in buildings –operation, interoperability, smart grid –not much work has been done with regard to the modelling of the energy data generated by different applications throughout the whole building life cycle. With this respect, a precedent is the research project InTUBE (4), which is an early attempt to create an information platform that uses semantic web standards to link building, simulation and performance data. More recently, Murray (5) has postulated the use of semantic web technologies to interlink different energy domains –policies, supply and demand, facility management and building design– as a way of reducing energy use.

Therefore, the design and development of semantic information systems that can help to improve the energy efficiency of buildings is an incipient area of research. There are no consolidated methods to create building energy information systems using ontologies. Neither are there –to be best of our knowledge– systems that facilitate services based on the combined access to the data generated over the different stages of the building life cycle. This article presents the methodology developed to create SEÍS, a Semantic Energy Information System that integrates energy related data and facilitates services to different agents involved in the design, construction and maintenance of buildings.

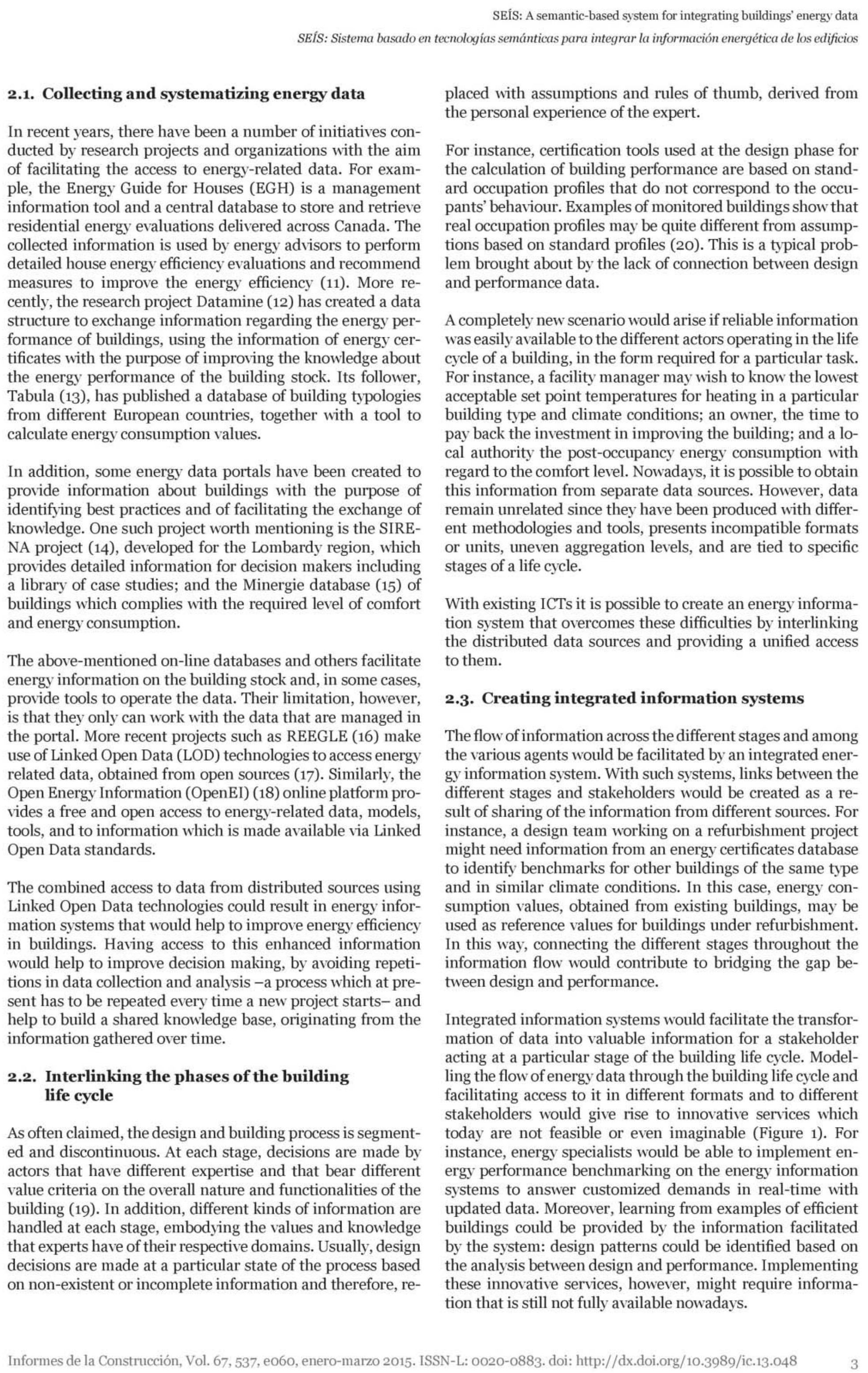
2. ACCESSING ENERGY DATA ALONG THE BUILDING LIFE CYCLE

The lack of adequate information on energy affects all the stages of the building life cycle, but it is particularly negative at the initial phase when the decisions that have the greatest impact on the performance of buildings are made. This lack of information makes it difficult for the design team –architects and engineers– to assess the actual impact of their design decisions on the energy performance of buildings without the aid of experts. Furthermore, having consistent and standardized information from the whole building life cycle would help to create more sustainable buildings by applying the enhanced Life Cycle Energy Assessment (LCEA) (6), using tools such as the Environmental Product Declarations (EPD) (7)[1].

With regard to the improvement of the existing building stock, it is necessary to have information on the energy performance of the existing buildings to spot the areas of intervention with the largest potential for improvement: urban or rural locations, typological and physical characteristics, equipment level and state of conservation. Having this information available would help policy makers and building and energy companies to identify and prioritize the most effective and efficient ways of improving the energy efficiency of the building stock. As shown in precedent studies (9), the frequent measurement of indicators such as equipment, energy consumption and occupants' behaviour of buildings, becomes necessary in order to have reliable information on energy building performance.

To sum up, a reliable and continuous data collection along the different stages of the building life cycle –from design to construction, during operation and refurbishment– is required to improve the decision making process at the design stage and to help all agents involved to make better informed decisions concerning the improvement of energy performances of new and existing buildings (10).

[1] According to the authors of this report, the integration of communication tools would help to improve EPD, by using consistent databases and standardized formats enabling benchmarking and monitoring progress (8).



2.1. Collecting and systematizing energy data

In recent years, there have been a number of initiatives conducted by research projects and organizations with the aim of facilitating the access to energy-related data. For example, the Energy Guide for Houses (EGH) is a management information tool and a central database to store and retrieve residential energy evaluations delivered across Canada. The collected information is used by energy advisors to perform detailed house energy efficiency evaluations and recommend measures to improve the energy efficiency (11). More recently, the research project Datamine (12) has created a data structure to exchange information regarding the energy performance of buildings, using the information of energy certificates with the purpose of improving the knowledge about the energy performance of the building stock. Its follower, Tabula (13), has published a database of building typologies from different European countries, together with a tool to calculate energy consumption values.

In addition, some energy data portals have been created to provide information about buildings with the purpose of identifying best practices and of facilitating the exchange of knowledge. One such project worth mentioning is the SIRENA project (14), developed for the Lombardy region, which provides detailed information for decision makers including a library of case studies; and the Minergie database (15) of buildings which complies with the required level of comfort and energy consumption.

The above-mentioned on-line databases and others facilitate energy information on the building stock and, in some cases, provide tools to operate the data. Their limitation, however, is that they only can work with the data that are managed in the portal. More recent projects such as REEGLE (16) make use of Linked Open Data (LOD) technologies to access energy related data, obtained from open sources (17). Similarly, the Open Energy Information (OpenEI) (18) online platform provides a free and open access to energy-related data, models, tools, and to information which is made available via Linked Open Data standards.

The combined access to data from distributed sources using Linked Open Data technologies could result in energy information systems that would help to improve energy efficiency in buildings. Having access to this enhanced information would help to improve decision making, by avoiding repetitions in data collection and analysis –a process which at present has to be repeated every time a new project starts– and help to build a shared knowledge base, originating from the information gathered over time.

2.2. Interlinking the phases of the building life cycle

As often claimed, the design and building process is segmented and discontinuous. At each stage, decisions are made by actors that have different expertise and that bear different value criteria on the overall nature and functionalities of the building (19). In addition, different kinds of information are handled at each stage, embodying the values and knowledge that experts have of their respective domains. Usually, design decisions are made at a particular state of the process based on non-existent or incomplete information and therefore, replaced with assumptions and rules of thumb, derived from the personal experience of the expert.

For instance, certification tools used at the design phase for the calculation of building performance are based on standard occupation profiles that do not correspond to the occupants' behaviour. Examples of monitored buildings show that real occupation profiles may be quite different from assumptions based on standard profiles (20). This is a typical problem brought about by the lack of connection between design and performance data.

A completely new scenario would arise if reliable information was easily available to the different actors operating in the life cycle of a building, in the form required for a particular task. For instance, a facility manager may wish to know the lowest acceptable set point temperatures for heating in a particular building type and climate conditions; an owner, the time to pay back the investment in improving the building; and a local authority the post-occupancy energy consumption with regard to the comfort level. Nowadays, it is possible to obtain this information from separate data sources. However, data remain unrelated since they have been produced with different methodologies and tools, presents incompatible formats or units, uneven aggregation levels, and are tied to specific stages of a life cycle.

With existing ICTs it is possible to create an energy information system that overcomes these difficulties by interlinking the distributed data sources and providing a unified access to them.

2.3. Creating integrated information systems

The flow of information across the different stages and among the various agents would be facilitated by an integrated energy information system. With such systems, links between the different stages and stakeholders would be created as a result of sharing of the information from different sources. For instance, a design team working on a refurbishment project might need information from an energy certificates database to identify benchmarks for other buildings of the same type and in similar climate conditions. In this case, energy consumption values, obtained from existing buildings, may be used as reference values for buildings under refurbishment. In this way, connecting the different stages throughout the information flow would contribute to bridging the gap between design and performance.

Integrated information systems would facilitate the transformation of data into valuable information for a stakeholder acting at a particular stage of the building life cycle. Modelling the flow of energy data through the building life cycle and facilitating access to it in different formats and to different stakeholders would give rise to innovative services which today are not feasible or even imaginable (Figure 1). For instance, energy specialists would be able to implement energy performance benchmarking on the energy information systems to answer customized demands in real-time with updated data. Moreover, learning from examples of efficient buildings could be provided by the information facilitated by the system: design patterns could be identified based on the analysis between design and performance. Implementing these innovative services, however, might require information that is still not fully available nowadays.

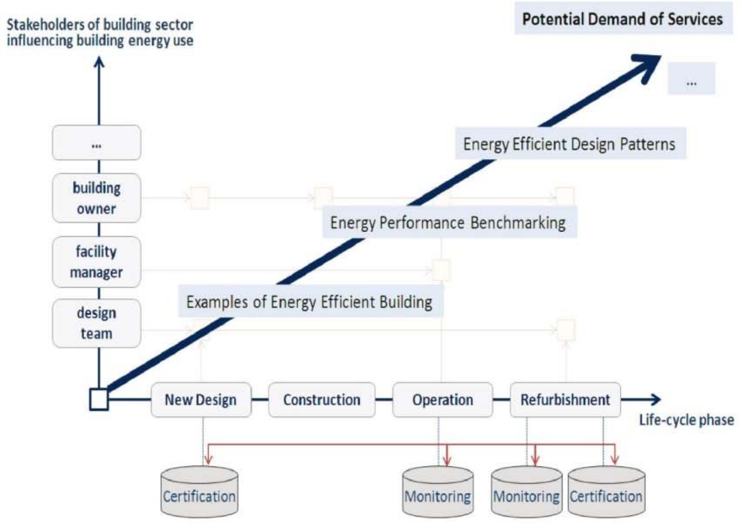

Stakeholders of building sector influencing building energy use
Potential Demand of Services
...
...
building owner
facility manager
design team
Energy Efficient Design Patterns
Energy Performance Benchmarking
Examples of Energy Efficient Building
New Design
Construction
Operation
Refurbishment
Life-cycle phase
Certification
Monitoring
Monitoring
Certification

Figure 1. Innovative services connecting data sources and stakeholders.

Today, different barriers hamper the creation of such integrated energy information systems. Some of them have to do with the division of design knowledge into specialized compartments, whereby each one corresponds to a particular professional skill. The lack of connection between stages –from design to operation– is also due to this knowledge specialization. Other kinds of difficulties, however, stem from the information itself: the difficulty to access energy related information in the appropriate format in the moment that it is needed.

3. CREATING A SEMANTIC ENERGY INFORMATION SYSTEM

In order to design and build an integrated energy system as envisioned in the previous section we have turned to semantic technologies. These technologies facilitate the interlinkage between multiple data sources based on a common vocabulary of terms and relationships known as ontologies. The following section describes the process followed to design and implement SEÍS, a semantic energy information system to provide access to building energy data.

The process to build the system has developed along two parallel tracks: a) A requirements capture process to identify the needs of the potential users of the system in terms of data and services. This process was carried out by means of use cases that encapsulated data, services and users in a particular application case. b) The creation of an energy model using ontologies that encompass the different data sources identified in the use cases. The energy model contains the terms of the shared vocabulary and their relationships (Figure 2).

3.1. Capturing users' requirements

At the start, some of the representative stakeholders involved at different stages of the building life cycle were identified. Their selection was guided by the experience and knowledge of the authors. They were represented by an architecture design team (Frutos – Sanmartin Arquitectos), a public administration department in charge of the building energy certification (the Catalan Institute of Energy, ICAEN), and a research group involved in building energy assessment, monitoring and management (Beegroup, CIMNE). This limited number of stakeholders was considered representative enough to carry out the process of knowledge and information capturing which was the first step in the construction of an ontology spanning through several stages of the building lifecycle.

In a series of interviews, we presented the selected stakeholders with a vision of the functionalities of the system to be developed. They were asked about the data they needed to perform their activities and the difficulties they had accessing it. Based on their feedback and the previous knowledge of the authors, we elaborated Table 1 with a more detailed list of users involved in the building life cycle, the data they required at each stage, and the actions they performed.

The table helped to identify the gaps that currently impede the flow of information across the building life cycle in which various activities are performed by different actors and in different stages of the building life cycle. This systematization of activities and actors prepared the ground for the next step of the knowledge capturing process that was to model use cases.

Modelling use cases

In order to proceed with the system development, some of the most relevant actors, phases and activities which had an impact on the decisions concerning the building energy performance during the building life cycle were identified and their interactions encapsulated in use cases.

In the context of this research, a use case is a conceptual model that relates users, data and services at a particular stage of

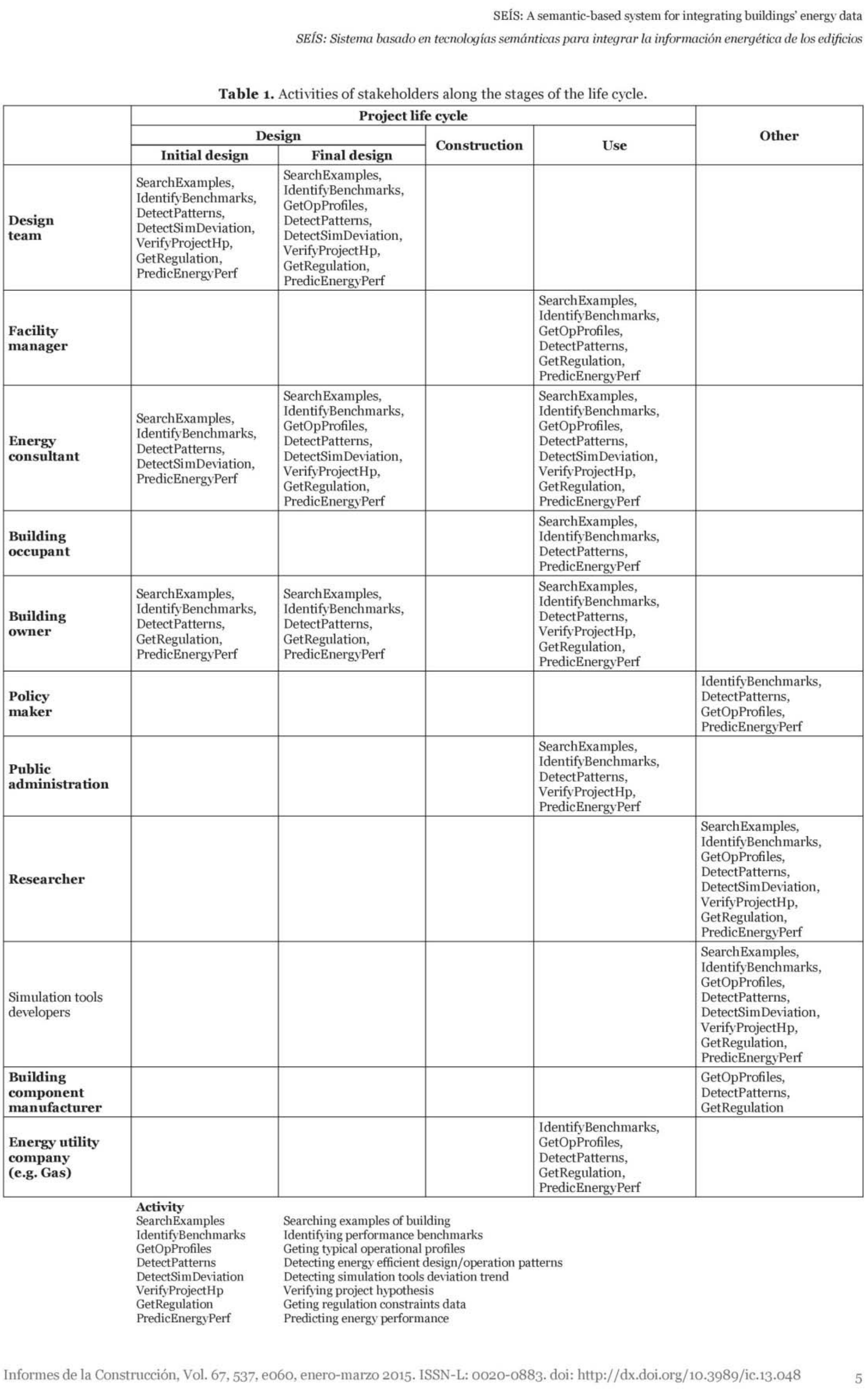


Table 1. Activities of stakeholders along the stages of the life cycle.

	Project life cycle				**Other**
	Design		**Construction**	**Use**	
	Initial design	**Final design**			
Design team	SearchExamples, IdentifyBenchmarks, DetectPatterns, DetectSimDeviation, VerifyProjectHp, GetRegulation, PredicEnergyPerf	SearchExamples, IdentifyBenchmarks, GetOpProfiles, DetectPatterns, DetectSimDeviation, VerifyProjectHp, GetRegulation, PredicEnergyPerf			
Facility manager				SearchExamples, IdentifyBenchmarks, GetOpProfiles, DetectPatterns, GetRegulation, PredicEnergyPerf	
Energy consultant	SearchExamples, IdentifyBenchmarks, DetectPatterns, DetectSimDeviation, PredicEnergyPerf	SearchExamples, IdentifyBenchmarks, GetOpProfiles, DetectPatterns, DetectSimDeviation, VerifyProjectHp, GetRegulation, PredicEnergyPerf		SearchExamples, IdentifyBenchmarks, GetOpProfiles, DetectPatterns, DetectSimDeviation, VerifyProjectHp, GetRegulation, PredicEnergyPerf	
Building occupant				SearchExamples, IdentifyBenchmarks, DetectPatterns, PredicEnergyPerf	
Building owner	SearchExamples, IdentifyBenchmarks, DetectPatterns, GetRegulation, PredicEnergyPerf	SearchExamples, IdentifyBenchmarks, DetectPatterns, GetRegulation, PredicEnergyPerf		SearchExamples, IdentifyBenchmarks, DetectPatterns, VerifyProjectHp, GetRegulation, PredicEnergyPerf	
Policy maker					IdentifyBenchmarks, DetectPatterns, GetOpProfiles, PredicEnergyPerf
Public administration				SearchExamples, IdentifyBenchmarks, DetectPatterns, VerifyProjectHp, PredicEnergyPerf	
Researcher					SearchExamples, IdentifyBenchmarks, GetOpProfiles, DetectPatterns, DetectSimDeviation, VerifyProjectHp, GetRegulation, PredicEnergyPerf
Simulation tools developers					SearchExamples, IdentifyBenchmarks, GetOpProfiles, DetectPatterns, DetectSimDeviation, VerifyProjectHp, GetRegulation, PredicEnergyPerf
Building component manufacturer					GetOpProfiles, DetectPatterns, GetRegulation
Energy utility company (e.g. Gas)				IdentifyBenchmarks, GetOpProfiles, DetectPatterns, GetRegulation, PredicEnergyPerf	

Activity	
SearchExamples	Searching examples of building
IdentifyBenchmarks	Identifying performance benchmarks
GetOpProfiles	Geting typical operational profiles
DetectPatterns	Detecting energy efficient design/operation patterns
DetectSimDeviation	Detecting simulation tools deviation trend
VerifyProjectHp	Verifying project hypothesis
GetRegulation	Geting regulation constraints data
PredicEnergyPerf	Predicting energy performance

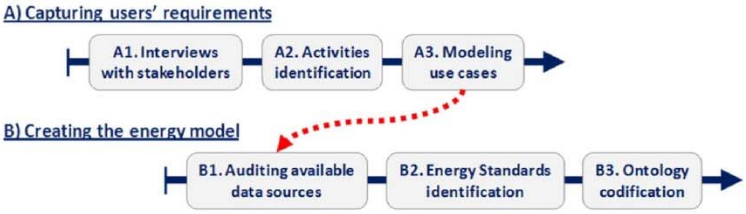

A) Capturing users' requirements
A1. Interviews with stakeholders
A2. Activities identification
A3. Modeling use cases
B) Creating the energy model
B1. Auditing available data sources
B2. Energy Standards identification
B3. Ontology codification

Figure 2. Development process of SEÍS.

the working process. The purpose of a use case is twofold: to define the specifications of the information system that is going to be developed and to create a description of the energy model formalized as an ontology.

Four different use cases have been modelled following the same procedure:

- definition of the goals of the user, and the activities to achieve them;
- identification of the information that the user needs to take actions to reduce energy use;
- identification of the information sources which are used in current practice;
- association of the information with the different stages of the building life cycle;
- detection of the obstacles and limitations that prevent access to the information.
- creation of new services which facilitate the information in the appropriate form;
- identification of the data that new services require.

The scope and characteristics of each use case are summarized as follows:

Use case 1. A design team working at the initial design phase of a new building

The user's profile is a small design team without advanced knowledge on energy efficiency. Their task is to design a low energy building. At the early design stage, they need to set energy performance goals and explore possible design solutions. To do this, they need:

- to look for performance benchmarks to set design goals in terms of energy efficiency
- to look for successful precedents which might serve as models to follow, indicating possible design solutions for similar design problems
- to understand the impact of their design decisions in terms of energy performance.

Typically, a design team obtains this information from several sources such as previous projects, databases, web pages, specialized journals, research and technical reports, consultants and colleagues. These information sources are heterogeneous and disperse and, therefore, difficult to reach. Furthermore, without the help of energy consultants, designers cannot adequately interpret some of the information they obtain. However, having access to the information and analysing it properly absorbs substantial resources from the project in terms of time and budget, as a whole.

Three services supported by the energy information system would help to overcome the obstacles and the limitations of the design team at the design stage:

- to provide examples of buildings for new building design
- to facilitate performance benchmarks for new building design
- to identify energy efficient design patterns

Use case 2. A facilities manager responsible for the operation of an existing building

The user is a facilities manager without advanced knowledge on how to operate a building and its equipment efficiently, in terms of energy performance. The goal is to improve the energy performance of a building, which means:

- to look for performance benchmarks of comparable buildings with similar building equipment.
- to look for precedents that might help to make decisions concerning changes in current operational settings and maintenance interventions.
- to estimate the impact of the changes made in current operational settings and maintenance interventions.

As in the previous case, nowadays a facility manager would have this information from several heterogeneous and disperse sources. The lack of a consolidated experience in energy- efficient operations of the facility manager would prevent him from identifying and exploiting the information. In particular, it is difficult to obtain energy monitoring data from existing buildings, information that may be especially relevant for a facility manager.

Three services would be provided by the energy information system to overcome these obstacles:

- to provide examples of buildings for building operation
- to facilitate performance benchmarks for building operation
- to identify energy efficient operational patterns

Use case 3. A building owner involved in the renovation of a building

Users are building owners who can be private investors or public administrations. Typically, their knowledge of building energy efficiency would be lower than the one of users in previous cases. The goal of the owner is to improve the performance of a building by renovating it. To achieve his goal, it is necessary:

- to look for performance benchmarks to have an idea of possible improvements

- to look for precedents of building renovations which can serve as models
- to be aware of the impact of the different options of the renovation of the building, in particular of the costs.

Nowadays, it is difficult to identify buildings that require renovation plans and to define appropriate actions to improve them. . The missing information would be retrieved from the information system with these services:

- to provide examples of renovated buildings
- to facilitate performance benchmarks for building renovation
- to identify energy efficient design patterns

Use case 4. An energy consultant performing simulations at the design phase

The user is an energy consultant working at the design stage who has a better knowledge of building performance and systems compared to the users in previous cases. The goal is to provide the design team with guidelines and recommendations in order to achieve a low energy building. The consultant participates in the refinement of energy performance objectives, proposes strategies and detailed solutions, and informs the design team of the energy performance of the design solution supported by an energy simulation software. The fulfilment of these activities requires:

- to look for performance benchmarks to set design objectives and to compare design alternatives.
- to look for precedents and explain the reasons for their good performance to the design team.
- to study and explain the impact that design variables have on the energy performance of the building to the design team.

Although experts have easier access to the different information sources they need to carry out their consultancy, they still have to access large amounts of information with high level of detail to derive some conclusions from it. This activity is very time-consuming and constitutes a substantial obstacle for the consultancy business. These four services provided by the information system would help energy consultants in their task:

- to upload building energy simulations
- to provide examples of buildings for new building design
- to facilitate performance benchmarks for new building design
- to identify energy efficient design patterns

Based on these four use cases, a prototype of an energy information system has been developed using semantic web technologies.

3.2. Creating an energy model

To interlink the various data sources identified in the use cases, it is necessary to create a shared vocabulary that will enable services to identify the data that different users need and in the format they require in a particular context. Each of the available data sources and energy parameters have been analysed and classified and the relationships between them identified. This set of definitions constitutes an energy model that is formalized in a later stage as an ontology (21).

Creating a shared vocabulary requires the usage of standardised terminologies and of certain agreed definitions that facilitate the understanding among users of the vocabulary. For the SEÍS energy model, we have used standard definitions proposed by previous research projects like Datamine (12) and by international ISO CEN standards (e.g. ISO 13790:2008). Energy experts and ontology engineers have done this work collaboratively. The result has been an ontology that describes entities and relations, data types and units from all data sources (22).

The SEÍS information system uses semantic technologies to integrate various data sources from different applications operating at various stages of the building life cycle (Figure3). The data sources encompass energy certificates, building descriptions, simulation outcomes, energy monitoring, and climate data:

- Building energy certificates and simulation data from design and refurbishment phases collected by the Catalan Energy Institute (ICAEN). Data include building energy rating, consumptions, types of mechanical systems and geometric characteristics. From 1800 certificates available, we have included 200 certificates that contain simulation outputs in the dataset. Currently, relevant attributes such as consumptions and emissions are not included in this database. For these, approximated values have been calculated from studies of the Spanish building sector (9), and from standard values of national regulations (23) and European standards (24).
- Building monitoring data, such as those provided by the company Leako, which maintains a database of thermal consumptions for heating and hot water, water consumption and temperature for several buildings. This database does not include building envelope and equipment data.
- Geographical data collected by public institutes such as the Geographical Information National Institute (CNIG)(25) which are published in the Spanish gazetteer afterwards. The data include population, areas, elevation, or Universal Transverse Mercator (UTM) coordinate. Climate zones –fundamental for buildings energy performance analysis– were derived from the classification of Spanish Building Code (26).

4. INTERACTING WITH A SEMANTIC ENERGY INFORMATION SYSTEM

The semantically modelled energy related data and the services that they operate with are accessible to different user profiles through an on-line application (www.seis-system.org) which has two front-end interfaces: a data portal where the integrated sources are shared with third-parties and the information system front-end which gives access to the energy services.

In this application, users with different profiles (Design Team, Facilities Manager, Building Owner, Energy consultant) can retrieve and upload data in interaction with the system and they can invoke services that operate with the data. This represents the first step in the creation of a more comprehensive energy system with additional data and services.

At the start, users select their profile. In this way, they can have access to the implemented services that suit their activity level. Once a service is selected, the user provides further

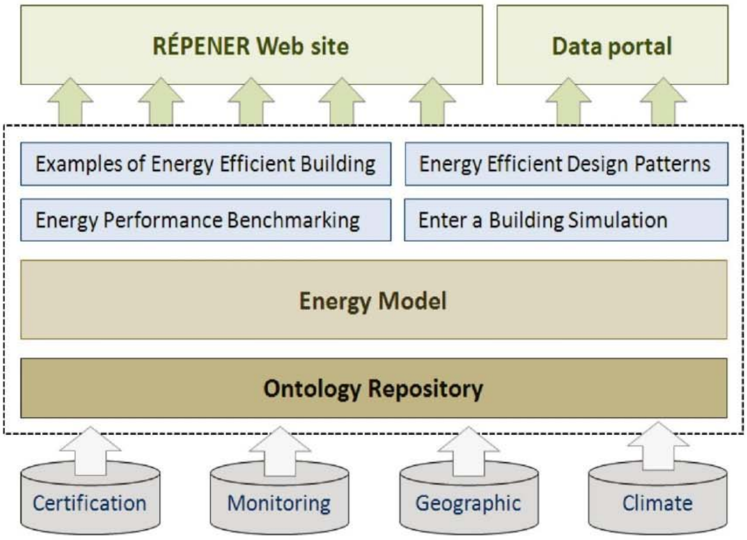

RÉPENER Web site
Data portal
Examples of Energy Efficient Building
Energy Efficient Design Patterns
Energy Performance Benchmarking
Enter a Building Simulation
Energy Model
Ontology Repository
Certification
Monitoring
Geographic
Climate

Figure 3. SEÍS information system structure.

inputs (e.g. use and location) which are needed to identify the appropriate service. Explanations and comments guide users in their navigation through the information system. In this way, they can progressively reach the information they are looking for.

Based on the input data provided by the user, each service generates an output data from the available data sources. This background process requires: first, the transformation of the row data that the service requires from the original sources into a neutral standard (e.g. *use of first heat generator* – standard attribute of the energy model – may be obtained from the ICAEN data in function of specific data from this source: *hot water system type*; *heating system type* and *joint generation of heating and hot water*), and secondly, the triggering of specific processes to produce the expected output (e.g. a set of projects, a benchmark).

During these processes, SEíS services use the energy model to access the data they require. So far, four services have been implemented which suit the needs identified in the use cases described in the previous section. These services are constrained by the data sources available. In the future, new services can be added thanks to the open structure of the system.

4.1. Implemented services

The services actually implemented in the system are described next:

Service 1. Examples of energy efficient buildings

This service searches for examples of efficient buildings based on the inputs Location and Use. A user may prioritise the Energy Uses and the Performance Indicators that are relevant for the search. The outcome of the search is a ranking of buildings organized according to two performance indicators: Energy (heating and cooling demand, total primary energy, CO2 emissions) and Indoor Space (time above and below comfort). The user can interact with the output by selecting a building to obtain specific information about it. This information is obtained from different databases and organized in four categories: building properties, performance, operation, and outdoor environment.

Service 2. Performance benchmarks

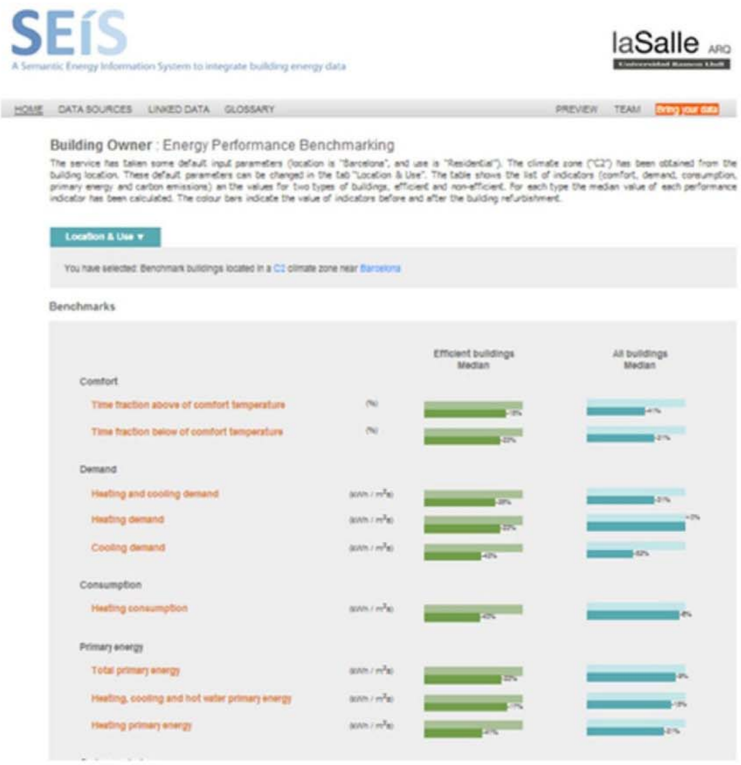
This service elaborates benchmarks based on the inputs Location and Use. A list of indicators (Comfort, Demand, Consumption, Primary energy and Carbon emissions) is displayed. Values are calculated for two types of building: "most efficient ones" and "all buildings".

If the user is a building owner, the benchmarking service furnishes insights about to improve the performance in buildings that have been renovated. For both groups (most efficient and all buildings) the median value of each performance indicator is calculated. Bar graphs show each indicator values of before and after the refurbishment (Figure 4).

Service 3. Energy efficient design patterns

This service helps designers to identify patterns for energy efficient design that are derived from the data of the buildings accessed by the information system.

The inputs required by the service include Location and Use. In addition, the user selects an Energy Performance indicator and the service determines the group of most efficient buildings related to it. The differences between most efficient buildings and other buildings are shown, and if a design variable is particularly relevant is highlighted. This way, a user of the system can easily identify which design variables have a greater impact on the design of an efficient building.



Building Owner : Energy Performance Benchmarking

The service has taken some default input parameters (location is "Barcelona", and use is "Residential"). The climate zone ("C2") has been obtained from the building location. These default parameters can be changed in the tab "Location & Use". The table shows the list of indicators (comfort, demand, consumption, primary energy and carbon emissions) an the values for two types of buildings, efficient and non-efficient. For each type the median value of each performance indicator has been calculated. The colour bars indicate the value of indicators before and after the building refurbishment.

You have selected: Benchmark buildings located in a C2 climate zone near Barcelona

Benchmarks

		Efficient buildings Median	All buildings Median
Comfort			
Time fraction above of comfort temperature	(%)		
Time fraction below of comfort temperature	(%)		
Demand			
Heating and cooling demand	(kWh / m²a)		
Heating demand	(kWh / m²a)		
Cooling demand	(kWh / m²a)		
Consumption			
Heating consumption	(kWh / m²a)		
Primary energy			
Total primary energy	(kWh / m²a)		
Heating, cooling and hot water primary energy	(kWh / m²a)		
Heating primary energy	(kWh / m²a)		

Figure 4. Performance benchmarks requested by a building owner.

Service 4. Building simulation results

An energy consultant may use this service to upload the data of a building energy simulation. The information to upload is divided in five categories: Project Data, Building Properties, Outdoor Environment, Operation and Performance. The data uploaded will be assigned to the terms of the ontology thus ensuring the compatibility of the new data with the existing data.

Having uploaded the data of the building simulation, an energy consultant may rank it with other buildings available in the system, by accessing the service Examples of Energy Efficient Buildings. Alternatively, a user can compare the results of the building simulation with benchmarks by accessing the service Energy Performance Benchmarking.

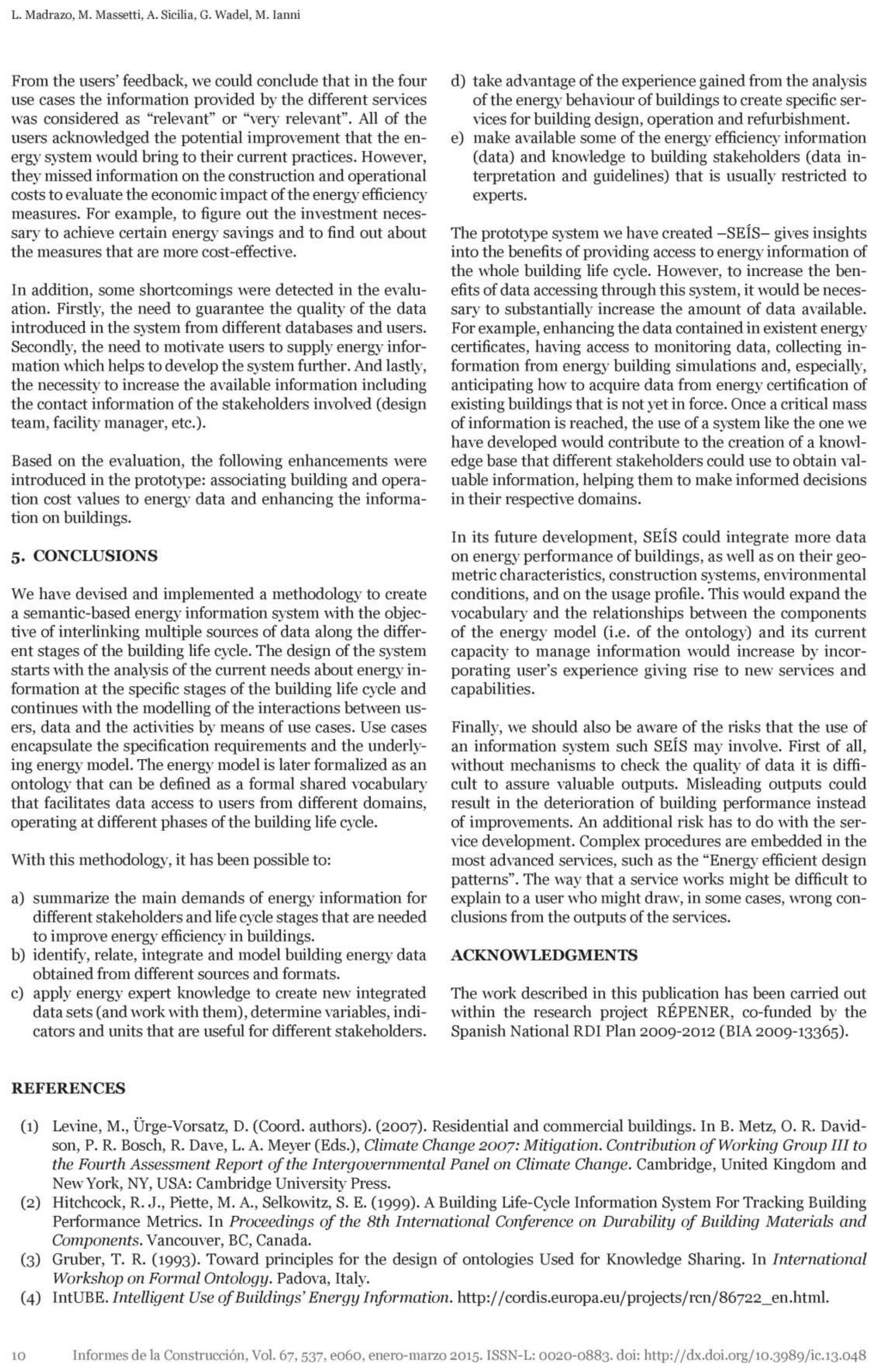
While the benchmarks' calculation of all selected buildings is based on median values, that of the subset of efficient buildings is determined as follows. All selected buildings are ranked according to weighted values from different performance indicators. Then the top 30% of the ranking are considered to be energy efficient buildings. In this case, we are applying the same rate as the one derived from the building energy certificates performed in Spain until September 2012. However, if this rate improves in the future (for example, if the objectives of Nearly-Zero Energy Buildings are reached) this percentage might be changed in the system.

4. VALIDATION OF THE SYSTEM

The information system has been tested with potential users in order to verify its value. The objectives of this validation process were to verify the strengths and weaknesses of the prototype and to identify the missing functionalities and future improvements. Six users representing the typical user profiles were invited to a working session:

a) an architect involved in sustainable building design.
b) an energy consultant specialized in building energy simulation.
c) a public officer of the Catalan Institute of Energy, responsible for energy certificates and promoting energy efficiency in buildings.
d) a public officer of the Catalan Housing Agency involved in the development of related projects and the management of social housing.
e) a facility manager responsible for large tertiary buildings.
f) a technician from the office of building permits of a local authority.

In a joint introductory session, the information system was presented to the users and its purpose and functionalities discussed. Afterwards, they had an initial opportunity to interact with the interface to test its usability. Later, separate interviews were conducted to elaborate conclusions on individual basis.



From the users' feedback, we could conclude that in the four use cases the information provided by the different services was considered as "relevant" or "very relevant". All of the users acknowledged the potential improvement that the energy system would bring to their current practices. However, they missed information on the construction and operational costs to evaluate the economic impact of the energy efficiency measures. For example, to figure out the investment necessary to achieve certain energy savings and to find out about the measures that are more cost-effective.

In addition, some shortcomings were detected in the evaluation. Firstly, the need to guarantee the quality of the data introduced in the system from different databases and users. Secondly, the need to motivate users to supply energy information which helps to develop the system further. And lastly, the necessity to increase the available information including the contact information of the stakeholders involved (design team, facility manager, etc.).

Based on the evaluation, the following enhancements were introduced in the prototype: associating building and operation cost values to energy data and enhancing the information on buildings.

5. CONCLUSIONS

We have devised and implemented a methodology to create a semantic-based energy information system with the objective of interlinking multiple sources of data along the different stages of the building life cycle. The design of the system starts with the analysis of the current needs about energy information at the specific stages of the building life cycle and continues with the modelling of the interactions between users, data and the activities by means of use cases. Use cases encapsulate the specification requirements and the underlying energy model. The energy model is later formalized as an ontology that can be defined as a formal shared vocabulary that facilitates data access to users from different domains, operating at different phases of the building life cycle.

With this methodology, it has been possible to:

a) summarize the main demands of energy information for different stakeholders and life cycle stages that are needed to improve energy efficiency in buildings.
b) identify, relate, integrate and model building energy data obtained from different sources and formats.
c) apply energy expert knowledge to create new integrated data sets (and work with them), determine variables, indicators and units that are useful for different stakeholders.
d) take advantage of the experience gained from the analysis of the energy behaviour of buildings to create specific services for building design, operation and refurbishment.
e) make available some of the energy efficiency information (data) and knowledge to building stakeholders (data interpretation and guidelines) that is usually restricted to experts.

The prototype system we have created –SEÍS– gives insights into the benefits of providing access to energy information of the whole building life cycle. However, to increase the benefits of data accessing through this system, it would be necessary to substantially increase the amount of data available. For example, enhancing the data contained in existent energy certificates, having access to monitoring data, collecting information from energy building simulations and, especially, anticipating how to acquire data from energy certification of existing buildings that is not yet in force. Once a critical mass of information is reached, the use of a system like the one we have developed would contribute to the creation of a knowledge base that different stakeholders could use to obtain valuable information, helping them to make informed decisions in their respective domains.

In its future development, SEÍS could integrate more data on energy performance of buildings, as well as on their geometric characteristics, construction systems, environmental conditions, and on the usage profile. This would expand the vocabulary and the relationships between the components of the energy model (i.e. of the ontology) and its current capacity to manage information would increase by incorporating user's experience giving rise to new services and capabilities.

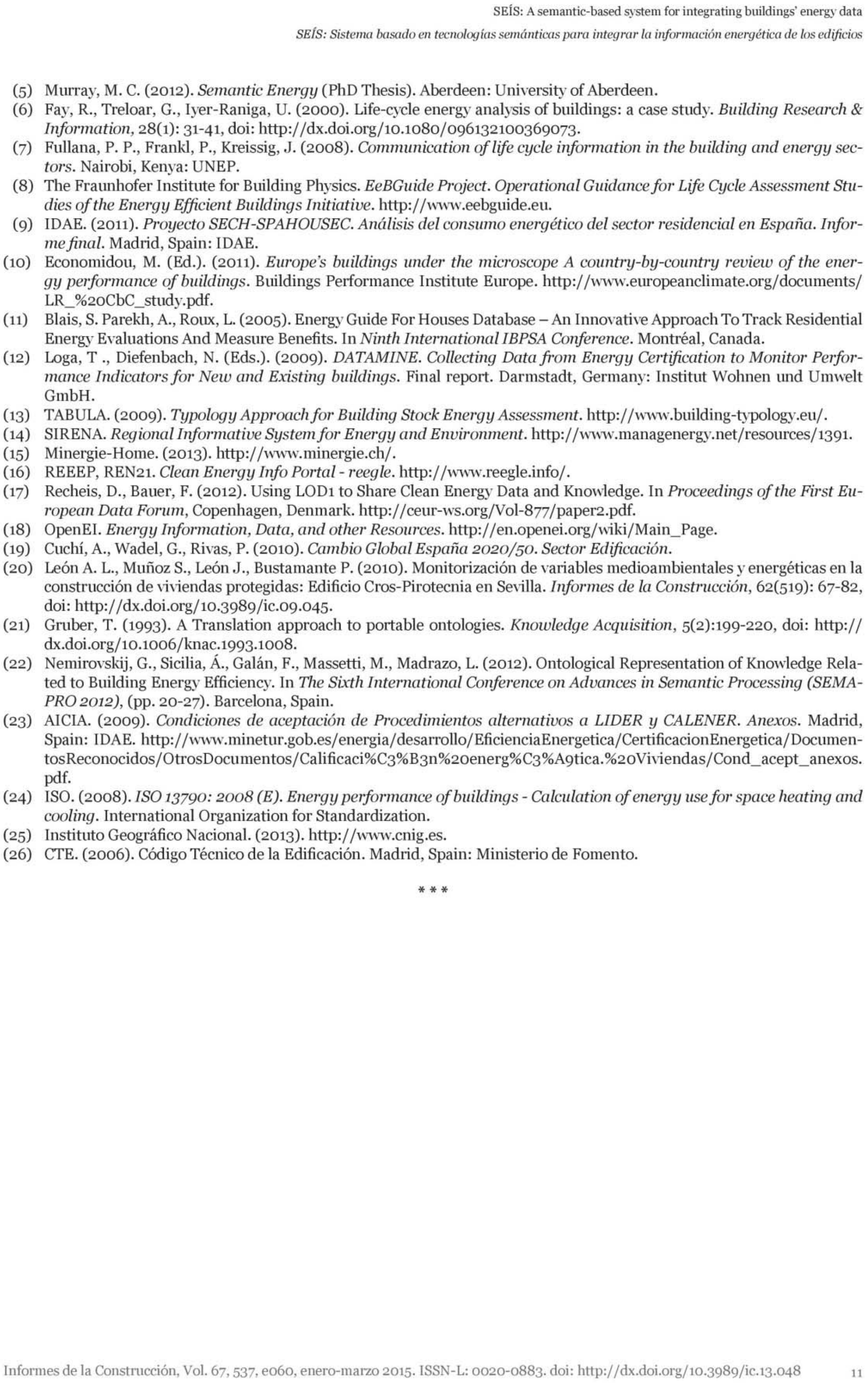
Finally, we should also be aware of the risks that the use of an information system such SEÍS may involve. First of all, without mechanisms to check the quality of data it is difficult to assure valuable outputs. Misleading outputs could result in the deterioration of building performance instead of improvements. An additional risk has to do with the service development. Complex procedures are embedded in the most advanced services, such as the "Energy efficient design patterns". The way that a service works might be difficult to explain to a user who might draw, in some cases, wrong conclusions from the outputs of the services.

ACKNOWLEDGMENTS

The work described in this publication has been carried out within the research project RÉPENER, co-funded by the Spanish National RDI Plan 2009-2012 (BIA 2009-13365).

REFERENCES

(1) Levine, M., Ürge-Vorsatz, D. (Coord. authors). (2007). Residential and commercial buildings. In B. Metz, O. R. Davidson, P. R. Bosch, R. Dave, L. A. Meyer (Eds.), *Climate Change 2007: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.
(2) Hitchcock, R. J., Piette, M. A., Selkowitz, S. E. (1999). A Building Life-Cycle Information System For Tracking Building Performance Metrics. In *Proceedings of the 8th International Conference on Durability of Building Materials and Components*. Vancouver, BC, Canada.
(3) Gruber, T. R. (1993). Toward principles for the design of ontologies Used for Knowledge Sharing. In *International Workshop on Formal Ontology*. Padova, Italy.
(4) IntUBE. *Intelligent Use of Buildings' Energy Information*. http://cordis.europa.eu/projects/rcn/86722_en.html.



(5) Murray, M. C. (2012). *Semantic Energy* (PhD Thesis). Aberdeen: University of Aberdeen.

(6) Fay, R., Treloar, G., Iyer-Raniga, U. (2000). Life-cycle energy analysis of buildings: a case study. *Building Research & Information,* 28(1): 31-41, doi: http://dx.doi.org/10.1080/096132100369073.

(7) Fullana, P. P., Frankl, P., Kreissig, J. (2008). *Communication of life cycle information in the building and energy sectors*. Nairobi, Kenya: UNEP.

(8) The Fraunhofer Institute for Building Physics. *EeBGuide Project*. *Operational Guidance for Life Cycle Assessment Studies of the Energy Efficient Buildings Initiative*. http://www.eebguide.eu.

(9) IDAE. (2011). *Proyecto SECH-SPAHOUSEC. Análisis del consumo energético del sector residencial en España. Informe final*. Madrid, Spain: IDAE.

(10) Economidou, M. (Ed.). (2011). *Europe's buildings under the microscope A country-by-country review of the energy performance of buildings*. Buildings Performance Institute Europe. http://www.europeanclimate.org/documents/LR_%20CbC_study.pdf.

(11) Blais, S. Parekh, A., Roux, L. (2005). Energy Guide For Houses Database – An Innovative Approach To Track Residential Energy Evaluations And Measure Benefits. In *Ninth International IBPSA Conference*. Montréal, Canada.

(12) Loga, T ., Diefenbach, N. (Eds.). (2009). *DATAMINE. Collecting Data from Energy Certification to Monitor Performance Indicators for New and Existing buildings*. Final report. Darmstadt, Germany: Institut Wohnen und Umwelt GmbH.

(13) TABULA. (2009). *Typology Approach for Building Stock Energy Assessment*. http://www.building-typology.eu/.

(14) SIRENA. *Regional Informative System for Energy and Environment*. http://www.managenergy.net/resources/1391.

(15) Minergie-Home. (2013). http://www.minergie.ch/.

(16) REEEP, REN21. *Clean Energy Info Portal - reegle*. http://www.reegle.info/.

(17) Recheis, D., Bauer, F. (2012). Using LOD1 to Share Clean Energy Data and Knowledge. In *Proceedings of the First European Data Forum*, Copenhagen, Denmark. http://ceur-ws.org/Vol-877/paper2.pdf.

(18) OpenEI. *Energy Information, Data, and other Resources*. http://en.openei.org/wiki/Main_Page.

(19) Cuchí, A., Wadel, G., Rivas, P. (2010). *Cambio Global España 2020/50. Sector Edificación.*

(20) León A. L., Muñoz S., León J., Bustamante P. (2010). Monitorización de variables medioambientales y energéticas en la construcción de viviendas protegidas: Edificio Cros-Pirotecnia en Sevilla. *Informes de la Construcción*, 62(519): 67-82, doi: http://dx.doi.org/10.3989/ic.09.045.

(21) Gruber, T. (1993). A Translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, doi: http://dx.doi.org/10.1006/knac.1993.1008.

(22) Nemirovskij, G., Sicilia, Á., Galán, F., Massetti, M., Madrazo, L. (2012). Ontological Representation of Knowledge Related to Building Energy Efficiency. In *The Sixth International Conference on Advances in Semantic Processing (SEMAPRO 2012)*, (pp. 20-27). Barcelona, Spain.

(23) AICIA. (2009). *Condiciones de aceptación de Procedimientos alternativos a LIDER y CALENER*. *Anexos*. Madrid, Spain: IDAE. http://www.minetur.gob.es/energia/desarrollo/EficienciaEnergetica/CertificacionEnergetica/DocumentosReconocidos/OtrosDocumentos/Calificaci%C3%B3n%20energ%C3%A9tica.%20Viviendas/Cond_acept_anexos.pdf.

(24) ISO. (2008). *ISO 13790: 2008 (E). Energy performance of buildings - Calculation of energy use for space heating and cooling*. International Organization for Standardization.

(25) Instituto Geográfico Nacional. (2013). http://www.cnig.es.

(26) CTE. (2006). Código Técnico de la Edificación. Madrid, Spain: Ministerio de Fomento.

* * *

The RÉPENER linked dataset. *Semantic Web*, 2016

The RÉPENER Linked Dataset

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. 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).

[*] Corresponding author.

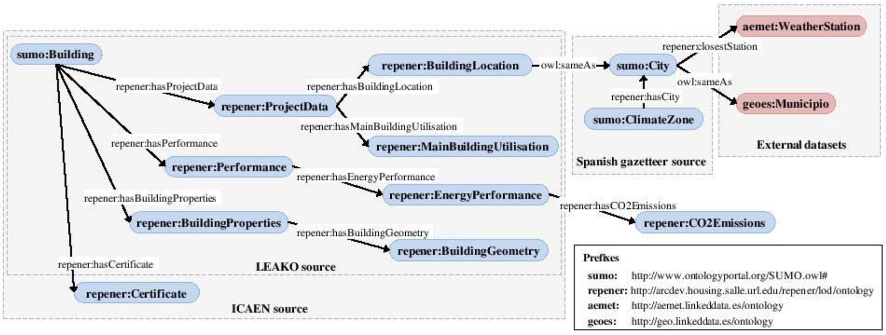

sumo:Building
repener:hasProjectData
repener:ProjectData
repener:hasBuildingLocation
repener:BuildingLocation
owl:sameAs
sumo:City
repener:closestStation
aemet:WeatherStation
owl:sameAs
geoes:Municipio
repener:hasCity
sumo:ClimateZone
Spanish gazetteer source
External datasets
repener:hasMainBuildingUtilisation
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
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.

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 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

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

[4] http://www.cnig.es

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.

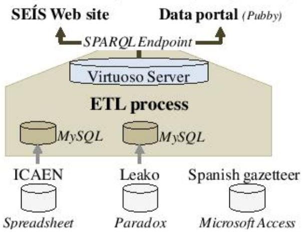


Fig. 2. Components of the RÉPENER dataset creation process.

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 understand 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 al-

[5] http://virtuoso.openlinksw.com

[6] http://arcdev.housing.salle.url.edu/repener

ways 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.

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.

[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

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

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 options 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

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

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 building 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

[13] www.reegle.info

[14] www.openei.org

[15] http://stats.lod2.eu/stats

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. eWork and eBusiness in Architecture, Engineering and Construction. 2012.

[2] IDAE, Energy consumption Analysis in the Spanish residential sector. Final Report. Available at: http://www.idae.es/index.php/mod.documentos/mem.descarga?file=/documentos_Informe_SPAHOUSEC_ACC_f68291a3.pdf (accessed September 6, 2013).

[3] ISO 13790: 2008 (E), 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. The Sixth International Conference on Advances in Semantic Processing (SEMAPRO) Barcelona, 2012.

[5] V. Corrado, S.P. Corgnati, and M. Garbino. "Energy Consumption Data Collection with DATAMINE". Energy, Climate and Indoor Comfort in Mediterranean Countries (Climamed). Aicarr, 2007.

[6] I. Niles and A. Pease. Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, 2001.

[7] C. Bizer and R. Cyganiak. D2RQ – Lessons learned. Position paper at the W3C Workshop on RDF Access to Relational Databases, 2007.

[8] L. Dodds and I. Davis. Linked Data Patterns. A pattern catalogue for modelling, publishing, and consuming Linked Data. Available at: http://patterns.dataincubator.org/book/linked-data-patterns.pdf (accessed September 6, 2013).

[9] J. Volz, C. Bizer, M. Gaedke and G. Kobilarov. Silk – A Link Discovery Framework for the Web of Data. Proceedings of the 2nd Workshop about Linked Data on the Web, 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, ISEMANTICS 6th International Conference on Semantic Systems, 2010.

[11] F. Bauer, D. Recheis, and M. Kaltenböck. data.reegle.info – A New Key Portal for Open Energy Data. In Environmental Software Systems.Frameworks of eEnvironment, IFIP Advances in Information and Communication Technology, Volume 359/2011, 2011.