READY4SmartCities - ICT Roadmap and Data Interoperability for Energy Systems in Smart Cities

Deliverable D4.1: Requirements and guidelines for energy data generation

Document Details

Delivery date:	*M8*
Lead Beneficiary:	*Universidad Politécnica de Madrid*
Dissemination Level (*):	*PU*
Version:	*1*
Preparation Date:	*26/05/2014*
Prepared by:	*Filip Radulovic (UPM), Raúl García-Castro (UPM), María Poveda-Villalón (UPM), Matthias Weise (AEC3), Thanasis Tryferidis (CERTH)*
Reviewed by:	*Simon Robinson and Strahil Birov (EMP) and Anna Osello (Polito)*
Approved by:	*Christian Mastrodonato (D'APPOLONIA), Asunción Gómez Pérez (UPM)*

(*) Only one choice between:

- PU = Public
- PP = Restricted to other programme participants (including the Commission Services)
- RE = Restricted to a group specified by the consortium (including the Commission Services)
- CO = Confidential, only for members of the consortium (including the Commission Services)

Project Contractual Details

Project Title:	ICT Roadmap and Data Interoperability for Energy Systems in Smart Cities
Project Acronym:	READY4SmartCities
Grant Agreement No.:	608711
Project Start Date:	2013-10-01
Project End Date:	2015-09-30
Duration:	24 months
Project Officer:	Rogelio Segovia

Revision History

Date	Author	Partner	Content	Ver.
26/2/022014	Filip Radulovic	UPM	Deliverable structure	0.1
28/02/2014	Filip Radulovic	UPM	Added Section 3.1	0.2
04/03/2014	Filip Radulovic	UPM	Added Section 3.2	0.3
7/03/2014	Filip Radulovic	UPM	Added Section 3.3	0.4
11/03/2014	Filip Radulovic	UPM	Minor structure update	0.5
14/03/2014	Filip Radulovic	UPM	Added Section 3.4	0.6
18/03/2014	Filip Radulovic	UPM	Added Section 3.5	0.7
20/03/2014	Raúl García-Castro	UPM	Document updated	0.8
25/03/2014	Filip Radulovic	UPM	Minor updates	0.9
28/03/2014	Filip Radulovic	UPM	Added Section 3.6	0.91
31/03/2014	María Poveda-Villalón	UPM	Updated Section 3.6	0.92
31/03/2014	Filip Radulovic	UPM	Added Section 3.7	0.93
04/04/2014	Filip Radulovic	UPM	Added Section 3.8	0.94
08/04/2014	Raúl García-Castro	UPM	Document updated	0.95
10/04/2014	Matthias Weise	AEC3	Provided input for Chapter 2	0.96
23/04/2014	Thanasis Tryferidis	CERTH	Added Chapter 2 + ANNEXS	0.97
02/05/2014	Matthias Weise	AEC3	Added IFC (building) example	0.98
05/052014	Filip Radulovic	UPM	Added Introduction and Conclusions	0.99
05/052014	Filip Radulovic	UPM	Minor updates. First version of the document	1.0
17/05/2014	Filip Radulovic	UPM	Implemented reviewer comments (Simon Robinson and Strahil Birov – EMP)	1.2
23/05/2014	Thanasis Tryferidis	CERTH	Updated Chapter 2	1.4
26/05/2014	Filip Radulovic, Raúl García-Castro	UPM	Implemented reviewer comments (Anna Osello – Polito) Second version of the document	2.0

Statement of financial support:
The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. FP7-SMARTCITIES 2013-608711

Executive Summary

One of the main tasks in the READY4SmartCity project is to enable a new energy data ecosystem that will support the interoperability and exploitation of data by adopting Semantic Web standards and technologies. Particularly, work package 4 promises to deliver requirements and guidelines on how to generate Linked Data, which is the key paradigm in the Semantic Web movement.

This deliverable presents the requirements for the generation of Linked Data in the energy domain (Chapter 2), which are derived from a survey conducted with various stakeholders. The survey investigated companies in the energy domain, gathering information about energy data and the characteristics of such data.

This deliverable also contains a set of guidelines for generating Linked Data in the energy domain (Chapter 3), which are developed having in mind the previously-mentioned requirements. The guidelines contain two examples of Linked Data generation, giving a more detailed view and helping the interested parties to better understand the generation process. One example is related to energy consumption in social housing, while the other example is related to data coming from building information models.

These guidelines help organizations in the energy domain from both public and private sectors in generating Linked Data from already-existing data, by providing detailed descriptions of each task in the generation process. Furthermore, the examples provided within the guidelines help developers from different organizations to gain better insight into the process of Linked Data generation, thus ensuring the highest quality of the outputs of the process.

For each of the tasks in the guidelines that can be supported by some software tool (either manually or automatically), we provide a list of potential tools to be used. This way, the document also provides a catalogue of the different technologies that can be used in the Linked Data generation process.

Glossary

Dataset	A collection of RDF data, comprising one or more RDF graphs that is published, maintained, or aggregated by a single provider. In SPARQL, an RDF Dataset represents a collection of RDF graphs over which a query may be performed.
Linked Data	A pattern for hyperlinking machine-readable data sets to each other using Semantic Web techniques, especially via the use of RDF and URIs. Enables distributed SPARQL queries of the data sets and a browsing or discovery approach to finding information (as compared to a search strategy). Linked Data is intended for access by both humans and machines. Linked Data uses the RDF family of standards for data interchange (e.g., RDF/XML, RDFa, Turtle) and query (SPARQL).
Ontology	A formal model that allows knowledge to be represented for a specific domain. An ontology describes the types of things that exist (classes), the relationships between them (properties) and the logical ways those classes and properties can be used together (axioms).
OWL	Web Ontology Language (OWL) is a family of knowledge representation and vocabulary description languages for authoring ontologies, based on RDF and standardized by the W3C.
RDF	Resource Description Framework (RDF) is a family of international standards for data interchange on the Web produced by W3C. RDF is based on the idea of identifying things using Web identifiers or HTTP URIs, and describing resources in terms of simple properties and property values.
SPARQL	SPARQL Protocol and RDF Query Language (SPARQL) defines a query language for RDF data, analogous to the Structured Query Language (SQL) for relational databases. It is a family of standards of the World Wide Web Consortium.
URI	A global identifier standardized by joint action of the World Wide Web Consortium and Internet Engineering Task Force. A Uniform Resource Identifier (URI) may or may not be resolvable on the Web. URIs can be used to uniquely identify virtually anything including a physical building or more abstract concepts such as colors.

Table of Contents

1 Introduction 7
1.1 Document structure 7
1.2 Contribution of partners 7
2 Requirements 8
2.1 Requirements elicitation technique 8
2.2 Stakeholders engaged 8
2.3 Technical specifications 9
2.4 Evaluation of results 9
2.4.1 General statistics 9
2.4.2 Data characteristics 11
2.4.3 Data access and legal issues 12
2.4.4 Quality and content of data 12
2.4.5 Organisational issues and motivation of participants 15
2.4.6 Conclusion and recommendation 16
3 Guidelines 17
3.1 Select data source 18
3.1.1 Energy consumption example 18
3.1.2 Building example 19
3.2 Obtain access to data source 20
3.2.1 Energy consumption example 20
3.2.2 Building example 20
3.3 Analyse licensing of the data source 21
3.3.1 Energy consumption example 22
3.3.2 Building example 23
3.4 Analyse data source 23
3.4.1 Energy consumption example 24
3.4.2 Building example 25
3.5 Define resource naming strategy 25
3.5.1 Energy consumption example 27
3.5.2 Building example 27
3.6 Develop ontology 28
3.6.1 Energy consumption example 31

3.6.2 Building example 39
3.7 Transform data source 40
3.7.1 Energy consumption example 43
3.7.2 Building example 47
3.8 Link with other datasets 47
3.8.1 Energy consumption example 49
3.8.2 Building example 49
4 Conclusions 51
5 References 52
ANNEX I - Questionnaire 53
ANNEX II - Survey replies 62
ANNEX III - Report for the invitation to the survey 67

1 Introduction

One of the main tasks in the READY4SmartCities project is to enable a new energy data ecosystem that will support interoperability and exploitation of data in the context of Smart Cities by adopting Semantic Web standards and technologies. Such data ecosystem will accommodate cross-domain data and will allow the exploitation of such data by identifying the set of relevant ontologies and the requirements and guidelines on how to generate, publish, and exploit data described according to these ontologies, i.e., Linked Data.

Work package 4 of the READY4SmartCities project promises to deliver requirements and guidelines on how to generate Linked Data in the energy domain, which is the cornerstone paradigm of the Semantic Web. To this date, a large number of both private and public companies and institutions from various domains have transformed their data into Linked Data, or have done so with data not coming from their institutions. However, this is not the case in the energy domain - the number of companies that have their data in the Linked Data form is rather low.

The purpose of this document is to present the requirements for the generation of Linked Data in the energy domain. Such requirements were compiled from a survey that was distributed to relevant stakeholders and that contains questions related to energy data and their characteristics (e.g., data formats and licenses, among others). The stakeholders that participated in the survey were people from organizations that provide energy-related data.

This document also presents a set of guidelines for the generation of Linked Data, which is an intensive engineering process that requires high attention in order to ensure the high quality of the produced data. These guidelines were developed having in mind the previously-mentioned requirements and consist of several consecutive tasks. For each task in the guidelines important information is provided, besides its detailed description, such as necessary inputs, desired outputs, different alternatives, a list of tools that help in achieving the task, and a list of tips for achieving better quality.

Along with these guidelines, two examples of Linked Data generation are presented. The first example is related to energy consumption in social housing. This example relies on the data from BECA project (Section 3.1), which provided us with data about energy consumption in households from several pilot sites in Torino, Italy. These data are originally stored in an Excel spreadsheet and have been transformed into Linked Data following the guidelines.

The second example of the guidelines is related to data from building information models. This example relies on data from various phases of the building lifecycle (design, construction, operation, demolition, and recycling).

1.1 Document structure

This deliverable is structured as follows. Chapter 2 presents the requirements for Linked Data generation in the energy domain, while Chapter 3 presents the guidelines for the generation of Linked Data, together with two examples. Finally, Chapter 4 draws some conclusions and defines future steps to be performed.

1.2 Contribution of partners

The following list states which partners have contributed to the different chapters of the deliverable.

- Introduction and conclusions. UPM
- Requirements. CERTH, AEC3, UPM
- Guidelines. UPM

2 Requirements

2.1 Requirements elicitation technique

In order to fully comprehend and analyse the user needs and the system requirements regarding the READY4SmartCities project, an elicitation technique was adopted. One of the most widely used elicitation technique [1] that provides a scientific approach through statistical analysis, while being rather simple to understand, are surveys through questionnaires. Thus, we created in work package 4 an online questionnaire (ANNEX I) that was distributed to various stakeholders acquiring in that manner valuable information.

The goal of this document is to provide a set of guidelines for Linked Data generation in energy domain. To this end, the questionnaires described in this chapter were developed with the aim of discovering which requirements and restrictions current energy-related data providers have in order to support those necessities through the guidelines.

At this point, the stakeholder's participation would aid to further elaborate the guidelines providing a more solid and accurate foundation taking into account actual needs and well-grounded considerations. The stakeholders that were approached were people with technical profile from organizations that provide energy-related data. Furthermore the survey's orientation was such that it surpassed the initial target population and it was widely available to any interested party.

2.2 Stakeholders engaged

The target population for the READY4SmartCities survey consisted primarily of stakeholders having access or connected somehow to energy-related data. Such stakeholders were reached through various channels as listed below:

- Mailing list of relevant partners/projects – each partner from the READY4SmartCities consortium shared a number of their partners from other projects based on their background and their relevance to the survey. The mailing list created counted more than 1000 people and was used to introduce the READY4SmartCities project and to invite interested people to fill in the survey.
- eeSemantics wiki[2] - CERTH partner is responsible for the maintenance of the eeSemantics wiki, forum and document library on Semantic Interoperability of Energy Efficiency ICT Tools for eeBuildings and beyond and therefore has access to the whole member list of relevant stakeholders (counting more than 500 members). An introduction to the READY4SmartCities project and concept was sent, followed by an invitation to participate in the survey, by both a post in the Forum and an email sent to the mailing list.
- READY4SmartCities Portal – the survey was made available and promoted on the first page of the READY4SmartCities website[3] and was posted on the website's newsletter.
- Social Networks – the questionnaire invitation was published through the READY4SmartCities project's social networks, namely LinkedIn and Twitter, early established in the project.

[1] J. A. Goguen, C. Linde, (1998). *Techniques for Requirements Elicitation*. In Proc. Requirements Eng. IEEE Computer Society, 152-161

[2] https://webgate.ec.europa.eu/fpfis/wikis/display/eeSemantics/Home

[3] http://www.ready4smartcities.eu/web/guest/guidelines-for-energy-data

- 4th VoCamp Participants – during the 4th VoCamp at Barcelona, participants with high relevance to energy-related data were approached and were requested to dedicate some time to answer the survey.
- Individual / personal contacts – the READY4SmartCities consortium consists of organizations with highly experienced and well-connected individuals. In some cases, relevant organizations were contacted after personal initiative.

2.3 Technical specifications

The READY4SmartCities questionnaire was created using Google Drive Documents, and more specifically Google Drive Forms. The reason as to why Google Drive Documents were selected lies with the fact that it's a well-structured, easy to handle and widely-known online application.

The questionnaire consisted of 22 questions structured in 8 pages, 4 spaces for further explanation and 1 final page for an agreement statement[4]. The questionnaire was available to all interested parties for nearly two months, from 3 February to 28 March 2014. After that time period, the results were collected, analysed and are presented in the sections below. The results do not contain any personal information and the confidentiality of the answers is preserved.

Furthermore, an online mailing platform was used in order to send the survey to more than 1000 individuals/organisations[5], which further provided rich insights and statistics on the total impact of the emails sent. A report with some interesting statistics, such as open/click rate, countries reached, etc., is presented in ANNEX III.

2.4 Evaluation of results

2.4.1 General statistics

The questionnaire invitation has been sent through the established READY4SmartCities network to more than 1500 people. From this, a total of 22 fully answered questionnaires have been received until 28th of May 2014, which means a return rate of around 1,3%. Thus the answers that we collected cannot be considered as representative of the whole population, and it should be clear that answers give only a first rough picture of the current situation. On the other hand, our approach for disseminating the survey was quite broad, which also entails that we covered profiles quite beyond the intended audience of the questionnaire (i.e., people with technical profile from organizations that provide energy-related data).

Answers are coming from 15 different countries whereas 1 is from outside of Europe. There is a good balance between non-profit (8), public (7) and private (7) organizations (Figure 1).

The targeted audience is, as expected, dominated by ICT, academia and the energy sector. It is worth mentioning that no answers have been received from organizations related to the Architecture Engineering and Construction (AEC) domain (Figure 2).

[4] https://docs.google.com/forms/d/1QuVr2bzKWqS2fLsYuElkE4z8l1g_jGuV672t5CkyqaY/viewform?usp=send_form

[5] http://mailchimp.com/

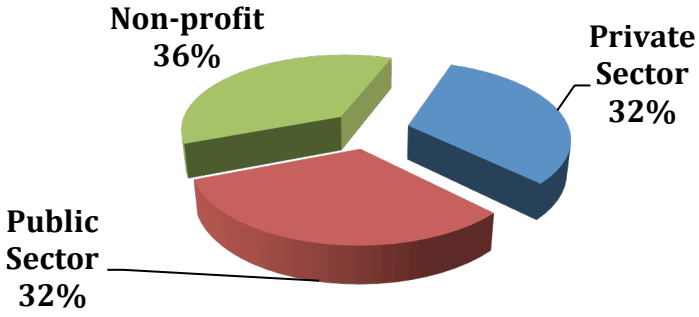


Figure 1. What sector does your company belong to?

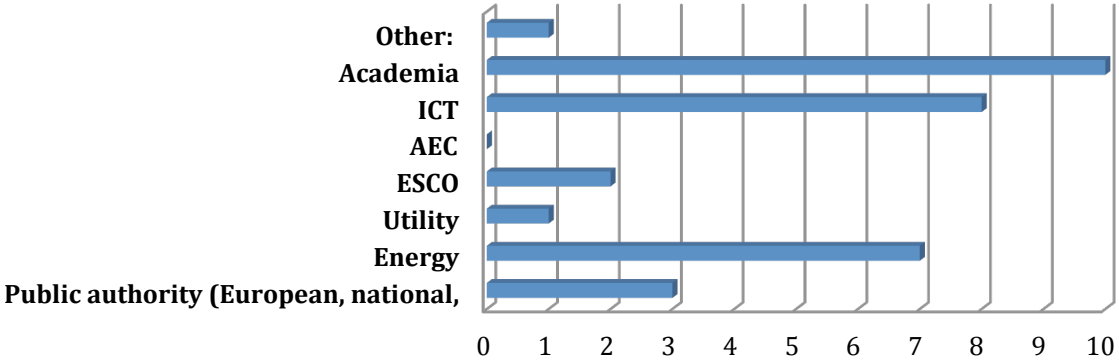


Figure 2. Which of the following categories best describes the organisation you primarily work in (regardless of your actual position)?

The Linked Data paradigm is known (9) or even has been used (5) by two third of the persons who answered the questionnaire (Figure 3). This is an unexpected high rate of Linked Data awareness, which is probably due to the fact that there is a higher willingness of this group of people to contribute to such survey. It may also explain why there is no participation of the AEC sector.

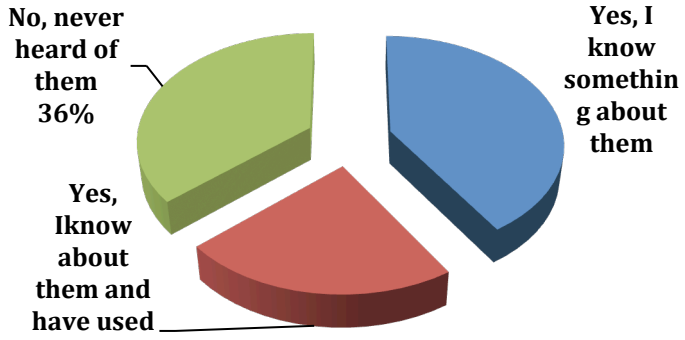


Figure 3. Are you familiar with the Linked Data paradigm and its related technologies?

Before starting to draw further conclusions from the received answers it should be mentioned that there are a lot of questions that allow multiple answers. Therefore, it is not always possible to check how single answers relate to each other, which would be necessary to assess the situation of a single data set. Nevertheless, the steps of the data generation process as discussed in WP4 are used as a baseline to evaluate results.

2.4.2 Data characteristics

Questions of the "Data Characteristics" section directly relate to technical aspects, namely the availability of a formal ontology and an adequate data representation. 6 out of 19 answers mention that they use either RDF-S or OWL as a formal representation of the ontology (Figure 4). All of them except one use RDF for data representation and thus already fulfil the main technical requirements of the Linked Data paradigm.

The majority of remaining answers use a structured and formalized data model, either as SQL table structure (11) or XML Schema definitions (5). It is interesting to note that SQL is mentioned three times together with "Unstructured" or "Not formalized" data model. In total there are 10 answers having an unstructured or not formalized data model out of 39 options given by 22 surveys. Accordingly, about 75% do have some sort of formalized data structure, where SQL and XSD seem to have the highest relevance. In the private sector 5 from 12 options include SQL, whereas only 1 is using XSD.

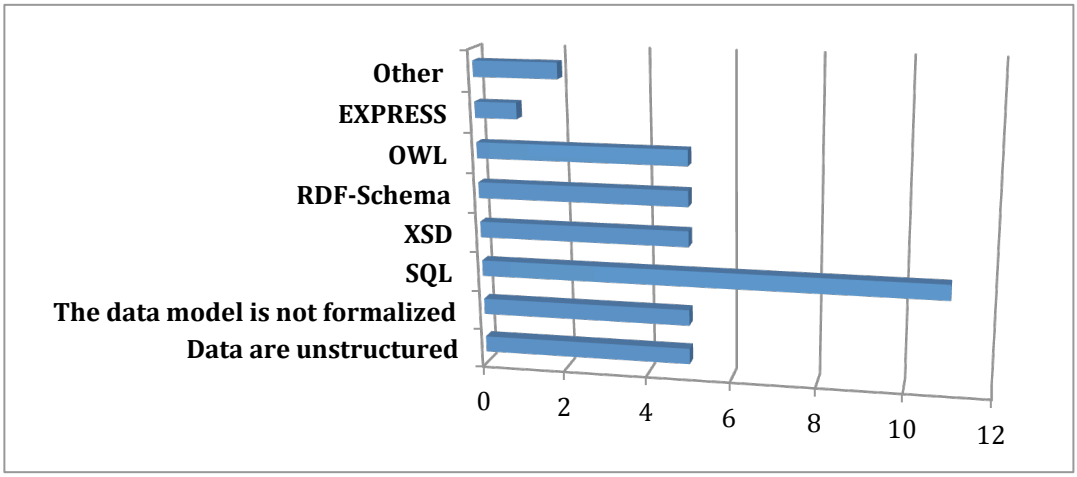


Figure 4. In what format are the data model used with the data?

As expected from the answers about the data model, most data sets are captured in relational databases (12), followed by the CSV (12), XLS (10), and XML (7) file formats. 67 options have been selected in total by 22 survey answers. 4 already use RDF and 14 are using unstructured data formats like PDF, DOC, RTF, Plain text or HTML (Figure 5).

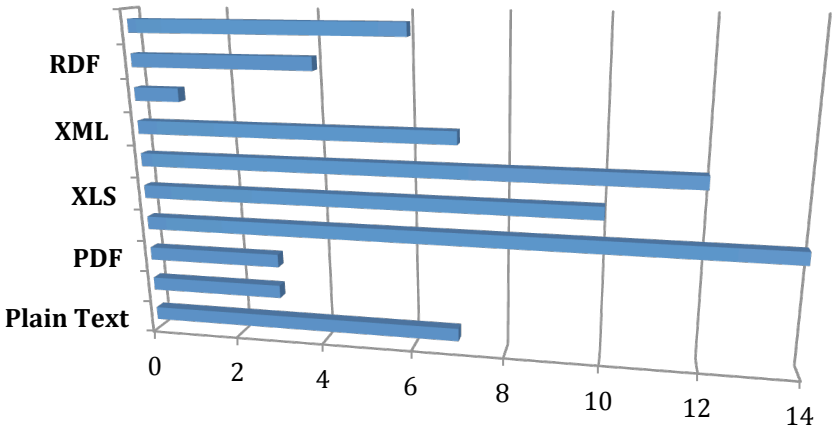


Figure 5. In what format are the data stored?

Meta data (data about the data) is available in over 60% of the cases, which may help to clarify data ownership and to track data changes.

2.4.3 Data access and legal issues

There are a couple of questions related to data access and legal issues. However, the received answers do not allow deriving a clear picture of the current situation. Some answers are not expected in given combinations, like for instance that data are made available to other organisations without being the rightsholder or having clarified the terms of use. Also, there are a lot of answers where the legal status is unknown. Only 7 (32%) organisations are the rightsholder of the data (Figure 6) and only 6 (32%) of them clarify the terms of use. No answers refer to a known license agreement like Creative Commons, GNU or MIT (Figure 7). Thus, it seems that legal issues are not yet thoroughly defined in most cases. The high rate of "unknown" answers might be due to the fact that the survey has been answered by technical experts. However, the legal status of data, in particular if published as Linked Data, is fundamental information that should be clear to everybody.

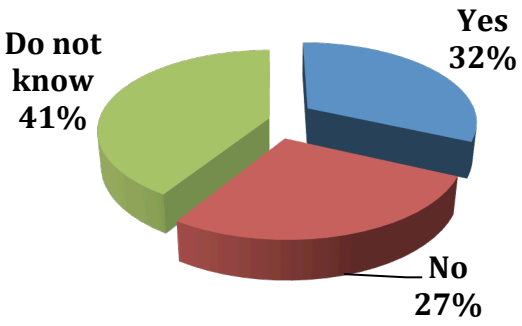


Figure 6. Is your organization the rightsholder of the data

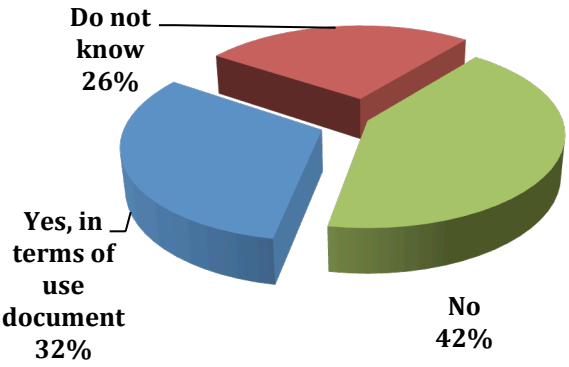


Figure 7. Are the terms of use of the data specified?

2.4.4 Quality and content of data

There are three answers where pure static data are stored. In most cases there will be dynamic (frequently updated, 7) or streaming data (2), or a mix of static, dynamic and streaming data (10) (Figure 8). As it is apparent the most often used method is the dynamic one.

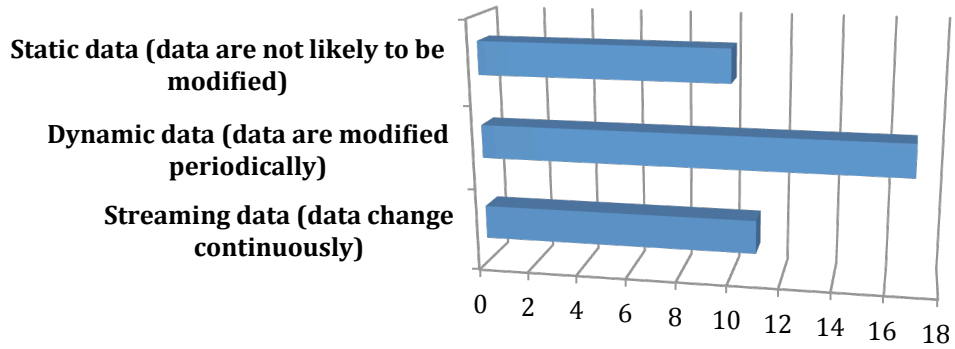


Figure 8. How often are the data updated?

In 14 cases some sort of data quality check is applied, mostly integrity checks (13), followed by inconsistency checks (10) and data cleansing (8) (Figure 9). Troublesome remains the fact that in quite a few cases (6) there isn't any assessment.

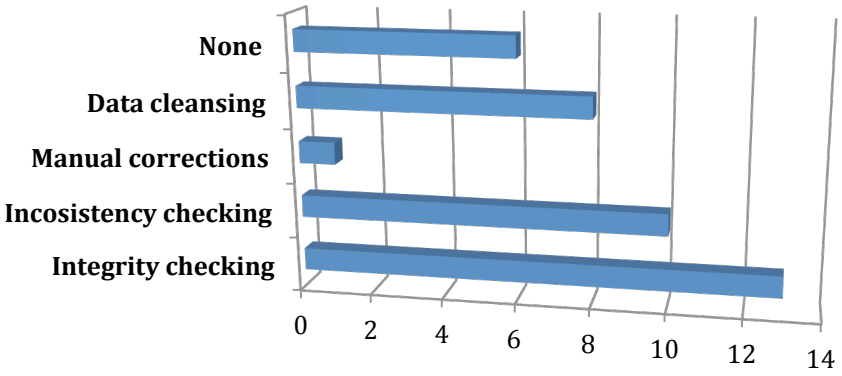


Figure 9. Which quality assessment techniques are applied over the data?

In terms of Quality Assessment techniques, fully automatic checks are applied in 10 cases, semi-automatic checks in 11 cases and manual checks in 4 cases only (Figure 10).

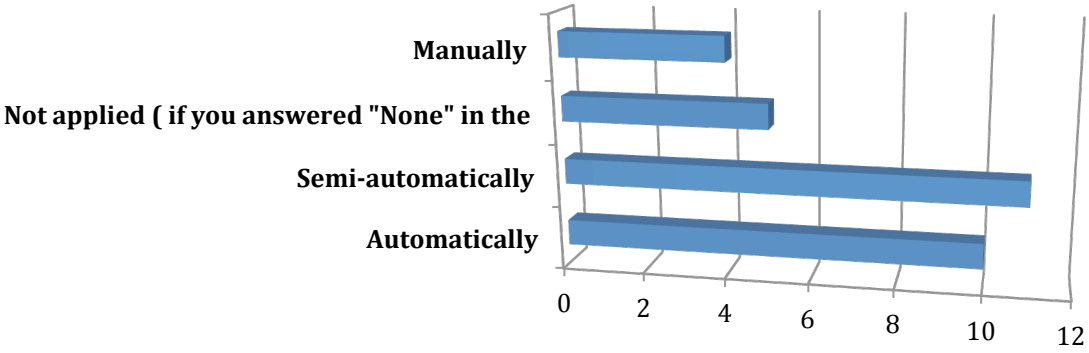


Figure 10. How are these quality assessment techniques applied?

Regarding data content, answers provide a broad mix of supported domains. Although no organisation belongs to the AEC industry most data are somehow related to buildings (Figure 14). Energy (Figure 12) and urban (Figure 15) data are nearly on the same level, followed by environmental data (Figure 13). Interestingly, only 7 of 22 participants claim that their data are related to other data sources, either from inside or outside the organization.

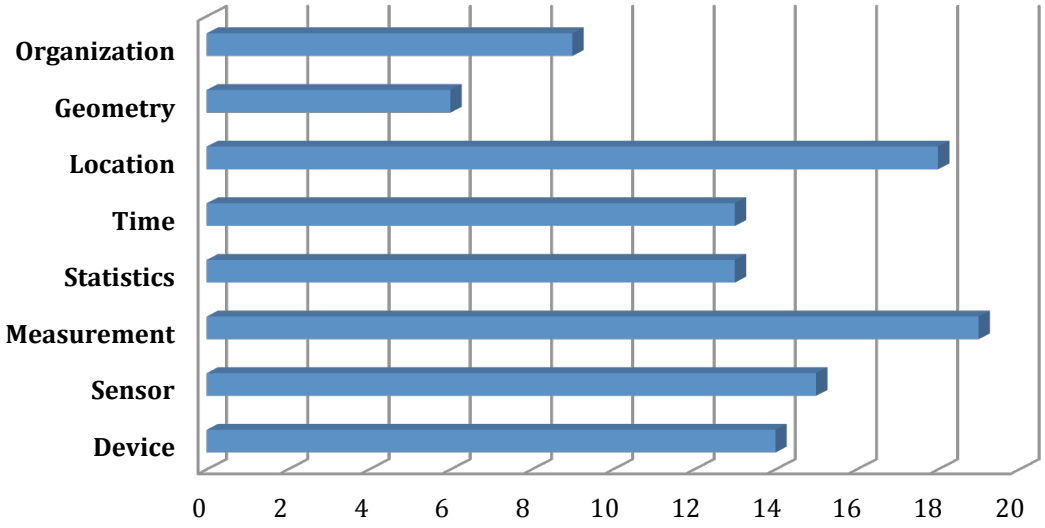


Figure 11. Which domains are covered by the data (General)?

Regarding energy-related domains, most of the data cover energy consumption (19), followed by energy production (15). Domains that do not affect directly the organisations and the individuals seem to have less interest over the participants.

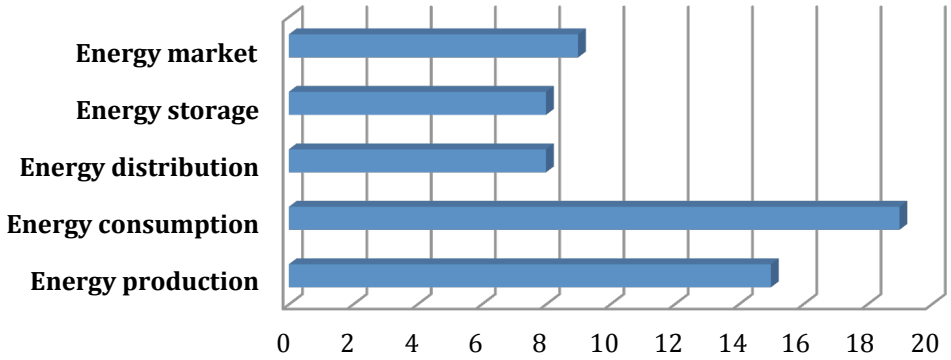


Figure 12. Which domains are covered by the data (Energy)?

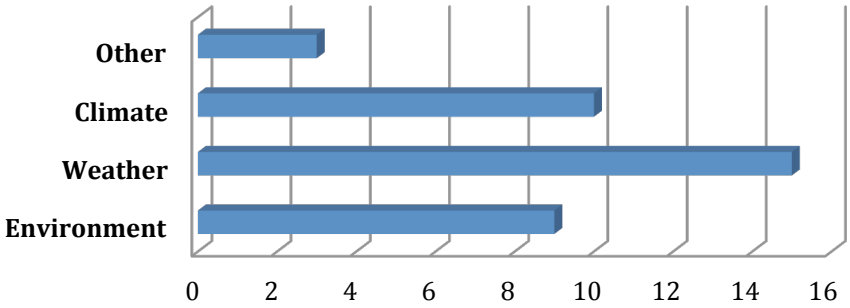


Figure 13. Which domains are covered by the data (Environmental)?

On the contrary, participants show high interest in building-related domains such as appliances (11), HVAC (15), electrical system (12), building automation (11), building elements (9), and the building in general (15). However, more detailed domains such as building's furniture and materials where not taken into high consideration.

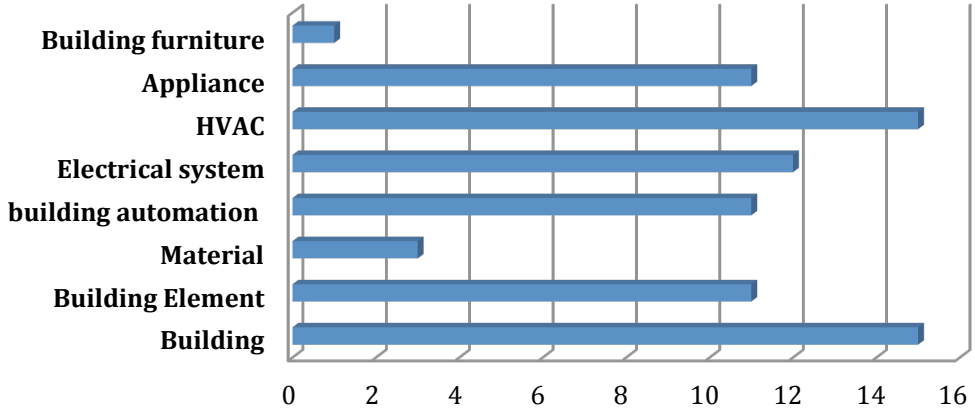


Figure 14. Which domains are covered by the data (Building)?

Furthermore, an elevated awareness in urban domains was observed with multiple answers covering domains such as neighbourhood/district (11), city in general (10), city lightning (6), and user behaviour (8).

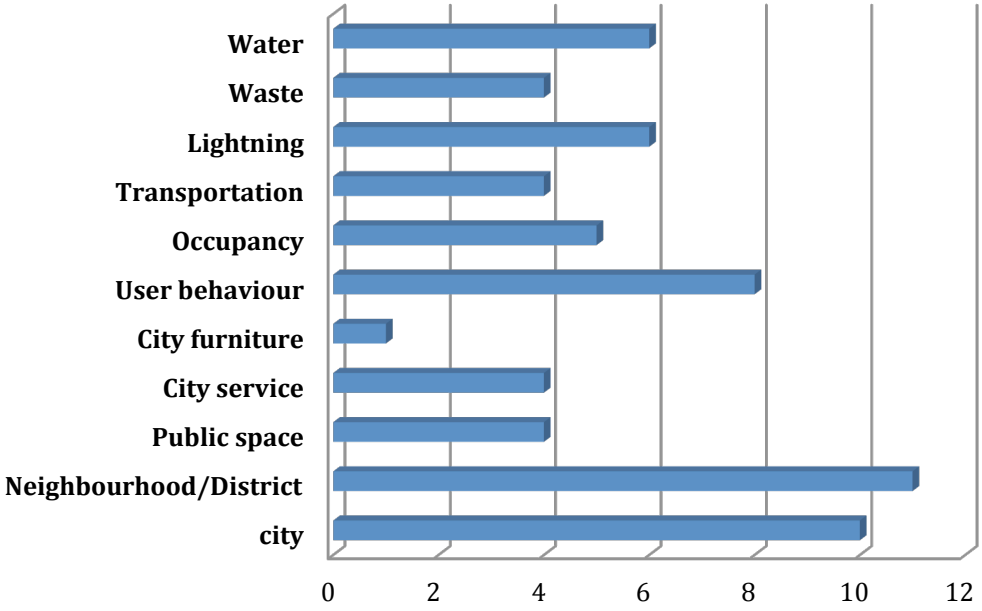


Figure 15. Which domains are covered by the data (Urban)?

2.4.5 Organisational issues and motivation of participants

Some questions try to identify the motivation to fill out this survey and to ask for barriers and open questions. A third of the organisations don't seem to be interested in publishing their data as Linked Data. Another third is not sure about the business case or is not fully clear about legal consequences. However, 86% are interested to learn more about Linked Data and would be interested in attending a workshop or tutorial on these topics. This shows that the topic as such is of high interest but that there are a lot of basic questions.

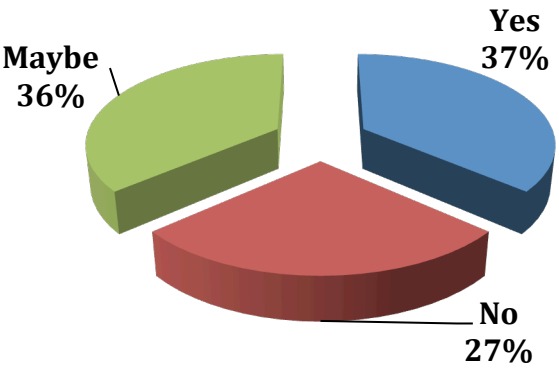


Figure 16. Would you be interested in publishing your data on the Web and linking them to other data?

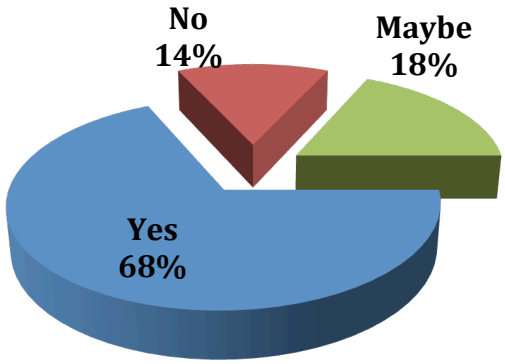


Figure 17. Would you be interested in attending a workshop or tutorial on these topics?

2.4.6 Conclusion and recommendation

In general, the survey shows a high diversity of data sources with a good mix of topics that have to be discussed when publishing information as Linked Data. Also, the survey is not fully representative of the target population due to having only 25 responses. Accordingly, it is difficult to develop clear Linked Data guidelines. However, the following requirements seem to be of high relevance in a more general sense:

- R1. Table structures stored in SQL, XLS or CSV file format (other data structures like XML seem to be not that important in practice or are unstructured and thus would require much more work to transfer to structured data),
- R2. Legal aspects like licensing and data ownership - clear recommendations could help to lower the barrier to publish data,
- R3. Access rights management or mechanisms for extracting public data - there is a lot of personal, confidential or otherwise protected data,
- R4. Changing data (dynamic or streaming data) and related questions like versioning, (automatic) data quality assurance and reliability, and
- R5. Data access through web services, proprietary APIs and data files seem to be equally important.

The graphs for the remaining questions of the survey that were not presented above are included in ANNEX II.

3 Guidelines

A large number of both private and public companies and institutions from various domains have transformed their data into Linked Data, or have done so with data not coming from their institutions. The process of Linked Data generation is an intensive engineering process and requires high attention in order to ensure the high quality of the produced Linked Data.

Tim Berners-Lee defined the following four principles of Linked Data [Berners-Lee, 2009], and the best practice is to follow these principles whenever producing new Linked Data:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information using the standards (RDF, SPARQL).
4. Include links to other URIs so that they can discover more things.

This chapter presents the guidelines for Linked Data generation in the energy domain (Figure 18); consecutive tasks and its inputs are represented with full lines, while the inputs from non-consecutive tasks are represented with discontinuous lines. Outputs that are final resources are represented with double lines. These guidelines define a process that consists of eight tasks; this document describes each task in detail through several important aspects:

- **Inputs** - inputs necessary to perform the task.
- **Outputs** - outputs to be obtained after the task is finished.
- **Description** - detailed description and steps to be performed within the task.
- **Tools** - a list of tools that help in performing the task or some parts of the task.
- **Alternatives** - different possibilities to perform the task or some parts of the task.
- **Tips** - best practices and recommendations that help in achieving better quality of the outputs of the task, and therefore of the generated Linked Data.

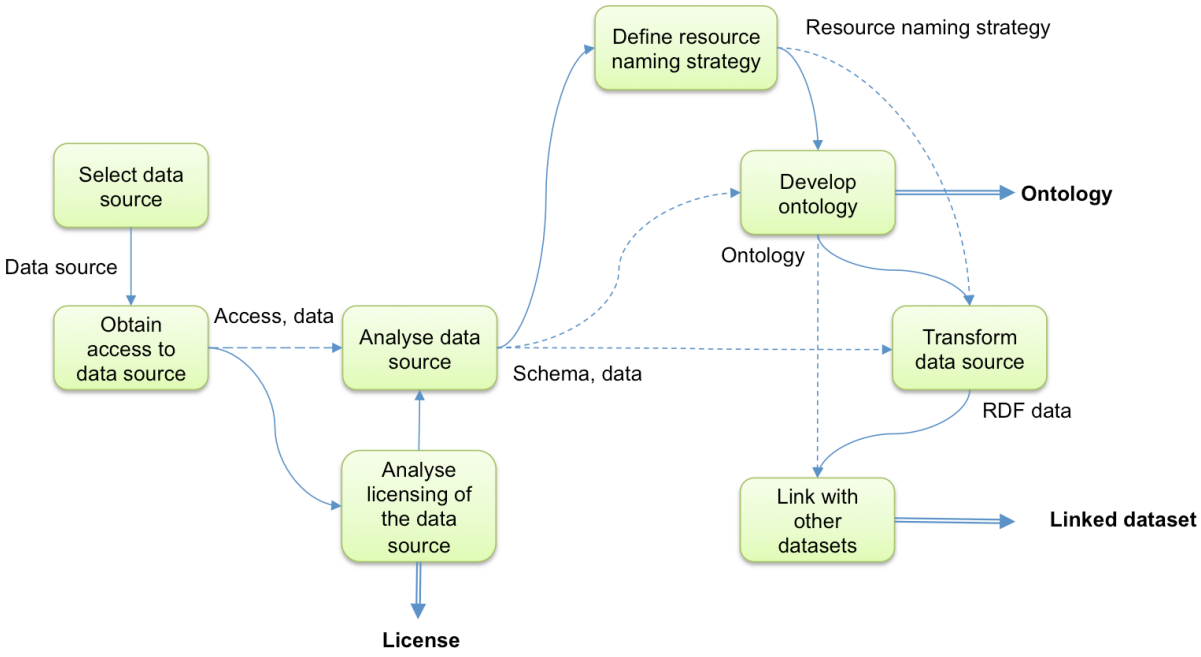


Figure 18. Steps of the guidelines for Linked Data generation.

In this deliverable, two examples will be provided. The first example is related to energy consumption in social housing scenarios, while the second example is related to data coming from building information models.

The process of Linked Data generation is performed mostly by developers; however, some of the tasks can involve other actors, such as managers or ontology engineers.

The Linked Data guidelines described in this chapter address the requirements derived in Section 2.4.6. However, since one particular requirement is not directly related to data generation (R3), this requirement is not addressed in this deliverable. The rest of the requirements are addressed as follows:

R1. The guidelines address the generation of Linked Data from tabular (SQL, XLS, or CSV) file formats, among others, which are the formats that are currently the most used in the energy domain. This is mainly covered in the "Analyse data source" and "Transform data source" tasks. Furthermore, one of the running examples deals with data in the CSV format.
R2. The guidelines address the issue of legal aspects, licenses, and data ownership, which is regarded as an important topic that could help lowering the barrier to publish data. To this end, there is a task that deals with this topic ("Analyse licensing of the data source").
R4. The guidelines cover the generation of static data, as well as dynamic data. One of the running examples provides details about how static data (or dynamic data that does not change frequently) can be managed.
R5. The guidelines describe various means of obtaining and accessing the data, including data stored in files, which is in line with the specified requirements. There is a task that deals with this topic ("Obtain access to data source").

In the following sections, we describe each task of the Linked Data generation process in detail.

3.1 Select data source

Inputs: -

Outputs: Data source

Description: The first step of the Linked Data generation process is the selection of the data source that will be transformed into Linked Data. In this step, a data source owned by the organization is selected depending on the specific needs of the organization or the expected value to be obtained. This task is achieved in two consecutive steps:

1. To define the requirements for the selection of the data source. These requirements might be related to the intended use of the generated Linked Data or to individuals or organizations that will use the generated Linked Data, among others.
2. To select one or several data sources to be transformed into Linked Data.

Alternatives:

- In most of the cases, the selected data sources will be owned by the organization.
- Alternatively, an organization may be interested in extending its data with data from other data sources not owned by the organization. In this case, these external data sources should also be selected.

3.1.1 Energy consumption example

In order to present an example of how to implement these steps in a real scenario, since in the project we do not have any energy data in our possession, we were limited to data sources owned by third parties. Therefore, it was necessary to first search for an appropriate data source for this example. In this search, we have looked for data in both private and public organizations as well as in the academic community, which address the energy domain in research projects. For the purpose of the example, we specified several requirements for such data:

- Data contain different types of energy-related information; preferably in some domain that is relevant to WP2 or WP3.
- Data are available.
- Data are represented in some machine-processable format (e.g., Excel, CSV, XML); the more structured the data are, the better.
- The data model used to represent the data is available (with documentation if possible).
- Data can come from one data source or from multiple data sources.
- Data can be easily linked with generic real-world entities (e.g., locations).
- Data have clear licensing and access policies.
- Data come from some real scenario.

After performing the search, data from the BECA (Balanced European Conservation Approach) project[6] appeared as an interesting example. BECA is a European ICT PSP project that aims to reduce energy consumption in European social housing. In order to achieve this goal, BECA has developed a set of innovative services for resource use awareness and resource management. Balance is achieved by addressing not only energy but also water, by including all key energy forms (electricity, gas and heating) and by including strong activities in Eastern Europe as well as in the North, South and West of the EU.

The services developed in the project are being used and tested in several pilot sites: Örebro, Darmstadt, Havirov, Manresa, Torino, Ruse, and Belgrade; and the project has collected data about energy consumption in households from such pilots, which is stored in Excel format. The collected data is fed into the eeMeasure[7] tool, which enables ICT PSP projects to calculate and record energy saving results using a consistent methodology.

This particular data set was selected because it was the one that complied most with the previously specified requirements. Furthermore, two READY4SmartCities partners are members of the BECA consortium (empirica, who are coordinating the BECA project, and Politecnico di Torino) and this would facilitate getting detailed information about the data.

3.1.2 Building example

Buildings have a high potential for energy savings, not only in the operation phase. The whole lifecycle is of interest, which starts with a smart and sustainable design, going further with the construction and operation of the building, and ends with demolition and recycling. A lot of data are generated and collected in these phases that can be used to improve energy efficiency. For instance, better building design will reduce overall energy consumption and building monitoring enables to identify unexpected states and to react accordingly. Due to the multi-disciplinary character of the AEC/FM industry there are many different domains that are involved in these phases and need to collaborate, which means that they have to share and merge their data in order to simulate and optimize all kinds of building behaviours.

A new methodology called Building Information Modelling (BIM) was introduced in the AEC/FM industry to tackle challenges related to data sharing and coordination activities. The BIM approach also tries to collect all relevant data in a consistent data model that can be used as an information hub for all project partners. A public standard, namely the Industry Foundation Classes[8] developed by buildingSMART and published as an ISO standard provides the basis for our building example. As there are a lot of general questions, in particular related to privacy, no specific building project was chosen to be discussed as a showcase. If possible, links to research projects are given for further reading.

[6] http://www.beca-project.eu/home/

[7] http://eemeasure.smartspaces.eu/eemeasure/generalUser/

[8] http://www.buildingsmart-tech.org/

For our building example we expect BIM data to be available in the IFC format and to describe a built facility. No particular requirements about the data content are defined for this single source of information.

3.2 Obtain access to data source

Inputs: Data source

Outputs: Access to data source, data

Description: A data source that is already owned by the organization is easier to access and in most cases such data can be accessed without obstacles. In the case of external data sources, the selected data source can be either in a public domain or not accessible. In the case when a data source is in the public domain, it can be accessed in a straightforward way. However, not all data sources are in the public domain and some of them are not accessible. In this case, it is necessary to first obtain access to the data source.

Data access consists of having the technical means to retrieve the data (addressed in this step) and of having the legal rights to use that data (addressed in the next task).

In the case of a data source that is not accessible, the steps to perform are:

1. To identify the person to contact in order to request access to the data source.
2. To request the access to the data source.
3. To retrieve the data from the data source.

Alternatives: There are several possibilities for retrieving the data (provided the user has the required credentials):

- Accessing the file or files containing the data.
- Accessing the data via a programming interface (e.g., a library API or a web service).
- Accessing the data stored in a database.
- Accessing the data from a stream of data (e.g., a sensor network, a social network feed).

3.2.1 Energy consumption example

The data source from the BECA project selected for this example is not accessible. Therefore, it was necessary to obtain access to it, which was done as follows:

- The person in charge of the data from the BECA project was identified and contacted by the partners from READY4SmartCities that are also participating in BECA.
- A request for accessing the BECA data source was sent to the identified person.
- The data from BECA data source was provided in an Excel spreadsheet and can be analysed by accessing a local copy of the file.

3.2.2 Building example

From a technical point of view, getting access to standardized BIM data is straightforward. Besides using IFC files there are different ways to get data access through APIs, which depend on the chosen hosting solution and could be either a proprietary API or a standardized data access interface (ISO 10303-22). In most cases an IFC file will be available that is stored according to the ISO standard 10303-21, the so-called STEP physical file format. This file format is supported by all major AEC-CAD tools and thus is seen as the solution with highest practical relevance.

Therefore, getting access to BIM data means to have an IFC file that represents a specific building state like as-designed or as-built. Depending on the state (lifecycle phase of building) there are different types of “data managers” that are responsible for having consistent and an up-to date building information model. For instance, in the design phase it is likely to be the project manager whereas in the operation phase it will be the facilities

manager or building owner. The amount of stored BIM data depends on the kind of information that is contained in an IFC file, in particular the state and domains, the way how geometry is represented, and the level of detail. It could easily reach several hundreds of megabytes even for small buildings.

3.3 Analyse licensing of the data source

Inputs: Data source, data

Outputs: License

Description: Licenses declared for a dataset specify the legal terms under which a dataset can be used and exploited. In the absence of specific contracts between the data provider and the data user, the generic licenses specify which actions can be executed on a dataset, provided that certain conditions are satisfied.

Therefore, in order to prevent legal conflicts, it is necessary to determine who is the rightsholder and which licenses have been declared for the data. This may be a non-trivial task as, in practice, the same dataset can be found in different Internet locations having different descriptions and formats and, what is worse, having different licenses specified.

An additional problem is that there are neither legal prescriptions on how to declare the license nor common standard practices to do so. For the own interest of the data publisher, the licensing terms are usually exhibited in the most explicit manner.

A list of steps for the task of license analysis could be formalized as follows:

1. To identify the authoritative dataset publisher. Prior knowledge of who is the rightsholder is essential to assess if that data has been published by (or in behalf of) the rightsholder or an authorized distributor.
2. To find the applicable license. Once having identified the authoritative dataset publisher, and assuming an access to data is obtained, the following places can be checked to determine the license:
    a. Browse the web page hosting the data. Typically, licensing information is provided as a text in the HTML footer (possible in separated page), as a well-known icon (e.g., Creative Commons), or as a combination of both.
    b. Browse the dataset metadata. For example, for RDF data, looking in the VoiD/DCAT description for structured information. DublinCore `license`, DublinCore `rights`, or XHTML `license` are the most common licensing elements.
    c. Inspect the dataset**,** as licensing information is sometimes present within the data. Comments in XML format, explicit DublinCore labels within the data, and simple textual notes are not uncommon.
    d. Contact the dataset publisher if the above steps have not proven sufficient, or if doubts exist about the applicable license.
3. To read the license and determine if the terms are satisfactory**.** Licenses are legal texts written with a particular style, but they are usually not long in extent and are easily understandable. If the intended uses of the dataset match the waived rights in the license, and the required conditions are satisfied, then the resource can be used.

The data publishers should always publish a license along with the published dataset. Therefore, in the case that a dataset does not have an associated license, an appropriate license should be defined. The following steps are proposed in order for the publisher to define the license of some data to be published:

1. To find out if the publisher is allowed to publish the dataset.
    a. If the publisher is the data creator, and this data are not based on data from other parties, then he/she is the rightsholder and can freely publish the data.
    b. If the publisher is not the data creator (or if parts of the dataset to be published are based on data from other parties) then permission from the other parties is needed. These parties may

have released their data under a license permitting derivative works; in that case, no specific permission is needed.

2. To choose the right license.
    a. If the publisher is the data rightsholder, any license can be chosen.
    b. If the data to be published includes (or is based on) data from other parties, these parties may have released their data imposing restrictions that must be satisfied. These restrictions may impose which kind of license is allowed.
    c. Accommodating the possible restrictions from the former point, the data publisher should choose a license suitable to his/her purposes. This choice, if possible, should be a well-known license, as it eases its understanding by data consumers.
3. To choose an appropriate method to publish the license.
    a. Make the license visible to humans. An easily recognizable placeholder is recommended in order to publish the license.
    b. Make the license visible to machines. If metadata is published along with the date, common elements to introduce the license should be used. In particular, a DublinCore license element is the most recommended choice. If using HTML, introducing RDFa is a good practice. Although not the best practice, the license can be referenced within the dataset.

Finally, access policies govern the availability of resources in a computer system. Although generic policy languages exist (e.g., XACML), these access policies are completely dependent on the access control system, and given that such systems are in no manner standard or homogeneous, no guidelines can be given.

Alternatives:

- In general, license discovery is not an automatable task and machines cannot be relied for this sensitive task. Yet, some resources may be available for machines under an access control schema, where queries can be made within a pay-per-access system. Policies governing these access control systems then will need to be understood. Yet, this is an unlikely scenario and out of the scope of the READY4SmartCities project.
- Legal rights to use the data can be achieved by either asking for the permission or paying the amount requested by data owners.

Tips:

- The analysis of the licenses of a data source should be performed upon all the available copies and formats of the data. Furthermore, all analyses should be performed by the same person or group of persons.
- If data are to be integrated within a larger dataset, make sure that licenses are compatible and their terms are not mutually exclusive.

3.3.1 Energy consumption example

We have analysed the license of the data from the BECA example following the previously specified steps:

1. We have contacted the partners from empirica, and we have concluded that the rightsholder of the data from the BECA example is the ATC Torino (Agenzia Territoriale per la Cassa della Provincia di Torino).
2. There is no license associated to the BECA dataset and, furthermore, to this date the dataset has not been published. Furthermore, after asking the rightsholders we can conclude that the data are not public.

Having in mind the previous findings, the license for the Linked Data that will be generated is specified:

1. Since the dataset is not public, no use of the data is permitted and, therefore, this is also the case for the Linked Data to be generated.
2. All rights have to be reserved for the RDF data produced in this deliverable.

3. The appropriate license will be specified in the RDF data using the Dublin Core *license* element.

3.3.2 Building example

Due to the nature of the AEC industry, for instance having one of a kind products with creative architectural design and complex engineering solutions, or the multiple authorship of created BIM data, there are substantial questions related to legal aspects. Also, there is a lot of sensitive, confidential or otherwise restricted data that are currently shared within a project consortium only.

It might be easy to answer the question of who is the owner of the data, but it is not so easy to decide about the legal status (authorship and copyright) of developed solutions and the privacy of user profiles. Depending on the data that should be published there are two main parties who must be asked for permission:

1. Building owner
2. User(s) of the building

At the time of writing this report it seems that readiness to publish building data is very low. Main reasons could be: (1) the uncertainty about all the consequences of making such data publicly available and (2) there is no clear business case or return of investment that makes it attractive to spend additional resources to ask for permissions and to publish the data. For instance, the showcases of the HESMOS project[9] that are used to validate developed eeBIM developments are not publicly available due to those legal reasons.

3.4 Analyse data source

Inputs: Access to data source, data

Outputs: Data, schema

Description: In the case when the license of the selected data source permits its further use, the next step is the analysis of the data source. The analysis allows getting insight into the data in it and into how such data are structured and organized. Data source analysis is a task achieved through two steps:

1. To analyse the data. In this step, it is observed which data are present in a data source and the characteristics of such data, such as quantities, value ranges, etc.
2. To obtain the schema of the data, describing the domain concepts that are described in the data source, together with all the relevant relationships between them. In some cases, the schema already exists; in those cases when the schema exists but does not describe all the data in the data source, this initial schema can be used with the results of data analysis in order to complete it if necessary. If the schema is not available, it has to be extracted directly from the data.

Alternatives: Data can be available in several different forms:

- Structured data: structured data implies that all the available data are stored in a structured format like CSV, XML, etc. Structured data are easier to analyse and reuse and, therefore, are less resource demanding.
- Unstructured data: unstructured data implies that all the available data are represented in natural language (e.g., in plain text files, PDFs, etc.). In order to use unstructured data it is necessary to make a deeper analysis, and such data are often more difficult to reuse.

[9] http://hesmos.eu/

Tips:

- If the data source schema does not exist, clear and unambiguous information about such schema should be provided; in this case, it is recommendable to use a standard modeling language (e.g., UML) when representing the extracted schema.

3.4.1 Energy consumption example

As mentioned before, the BECA project provided us with Excel files containing energy consumption data (heating, cold water and hot water) from several social housing pilot sites in Torino. Figure 19 shows an excerpt of those data related to hot water consumption.

HOT WATER — Monthly measurements

Eval ID (combined) matching var for other maps	Building	Kind of Setup	Suchkriterium	Internal Dwelling Number (Dwelling_ID)	Tenant ID or Name	Date of first meter reading	COMMENTS	Consumption YEAR 2011 (cbm, m³)											
	"Name"	"Name"		Unique Dwelling ID within each Site	changes if new tenant moves in (insert new row for new ID)	Date	remarks on data problems	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
712821111	[redacted]	n/a	BC89_0010005_1	-77	BC89_001 0005_1			7	2	8	3	4	2	3	3	3	2	4	3
712821121	[redacted]	n/a	BC89_0010006_1	-77	BC89_001 0006_1			5	3	8	2	4	3	4	4		3	3	5
712821131	[redacted]	n/a	BC89_0010007_1	-77	BC89_001 0007_1			19	2	16	9	10	8	9	9	9	11	9	7
712821141	[redacted]	n/a	BC89_0010008_1	-77	BC89_001 0008_1			5	2	5	11	3	10	8	8		0	0	4
712821151	[redacted]	n/a	BC89_0010009_1	-77	BC89_001 0009_1			24	3	8	1	1	0	0	0	0	2	0	0
712821161	[redacted]	n/a	BC89_0010010_1	-77	BC89_001 0010_1			5	3	8	3	5	4	5	5	2	3	7	5
712821171	[redacted]	n/a	BC89_0010011_1	-77	BC89_001 0011_1			6	4	10	5	7	5	6	6	6	8	6	4
712821181	[redacted]	n/a	BC89_0010012_1	-77	BC89_001 0012_1			1	3	4	1	1	1	6	6	1	3	1	1
712921191	[redacted]	n/a	BC93_0020013_1	-77	BC93_002 0013_1			36	5	21	6	9	6	8	8	8	4	10	6
712921201	[redacted]	n/a	BC93_0020014_1	-77	BC93_002 0014_1			2	2	5	3	3	3	2	2	2	3	4	4
712921211	[redacted]	n/a	BC93_0020015_1	-77	BC93_002 0015_1			7	2	7	0	0	0	0	0	0	0	0	0
712921221	[redacted]	n/a	BC93_0020016_1	-77	BC93_002 0016_1			11	2	10	6	10	3	8	8	3	6	8	7
712921231	[redacted]	n/a	BC93_0020017_1	-77	BC93_002 0017_1			18	4	13	2	11	9	12	12		4	9	9
712921241	[redacted]	n/a	BC93_0020018_1	-77	BC93_002 0018_1			12	3	13	5	8	6				16	0	1
712921251	[redacted]	n/a	BC93_0020019_1	-77	BC93_002 0019_1			7	3	10	7	9	6	5	5	5	8	6	4
712921261	[redacted]	n/a	BC93_0020020_1	-77	BC93_002 0020_1			1	3	6	1	4	0	0	0	0	8	6	4
712921271	[redacted]	n/a	BC93_0020021_1	-77	BC93_002 0021_1			7	2	8	3	5	3	3	3	3	4	5	4
712921281	[redacted]	n/a	BC93_0020022_1	-77	BC93_002 0022_1			9	2	11	4	2	5	16	16		6	8	11
712921291	[redacted]	n/a	BC93_0020023_1	-77	BC93_002 0023_1			19	1	20	8	14	6	9	9	9	5	9	6
712921301	[redacted]	n/a	BC93_0020024_1	-77	BC93_002 0024_1			15	2	15	7	9	6	7	7	7	10	9	6

Figure 19 - Hot water consumption in the BECA dataset.

Data about dwelling characteristics in a pilot site consists of several identifiers to identify sites, buildings, dwellings, tenancies, and evaluation groups, among others. All these fields have integers for their values. Data about pilot name, building name, and kind of setup are also present as strings, together with tenant identifiers, and with comments that are mostly related to problems in the data related to a certain tenancy.

For each dwelling, it is indicated with a value of zero or a value of one whether there have been some changes in tenancy, as well as the dates of changes (from which date until to which date). Furthermore, the number of persons and the size of dwellings in square meters are indicated, together with some technical details such as whether a dwelling is equipped with a mechanical ventilation system.

Each table related to energy consumption (heating, hot water and cold water) contains data about tenancies (these data are the same as in the dwelling table and serve the purpose of identification). For each tenancy, consumption data are stored from January 2011 to October 2013; consumption data are represented as decimal values, and every entry denotes the consumption in a tenancy for the indicated month. Also, data about heating degree days (HDD) for the pilot site are present for the same time period and in the same form as consumption data.

Part of the data about heating actually relates to both heating and hot water; i.e., the data about a tenancy in this case denotes the sum of heating and hot water consumption.

It is important to notice that the Excel spreadsheet contains empty sheets and empty columns, since not all entries are filled. These data have not been taken into account in this example.

Figure 20 shows the schema of the data from the BECA example, which is based on the tables in the Excel spreadsheet.

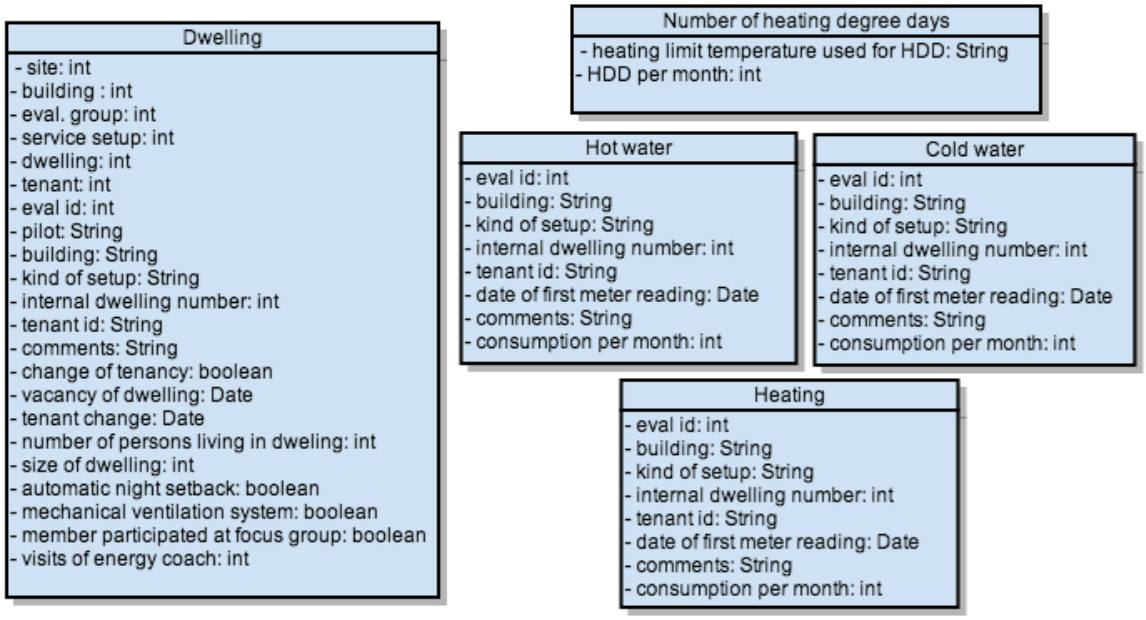


Figure 20 - Schema of the BECA dataset.

3.4.2 Building example

IFC data are stored according to a well-defined, publicly available data structure that is defined in EXPRESS (ISO 10303-11). EXPRESS is a powerful modelling language that is based on the extended entity relationship model and supports well-known data modelling concepts like inheritance, different types of constraints (rules, typing, cardinality of relationships, etc.) and even functions. The meaning of IFC data can be derived from the standard and is based on a common understanding of main concepts within the AEC industry. But the IFC data structure also enables to add own content where the meaning may not be known and thus has to be demanded from the data owner. However, such proprietary data may not be in scope to be published as Linked Data as they typically define some tool-specific settings. In general, data filtering is the main task within this step in order to limit the amount of data and to focus on use case specific information. Besides removing unnecessary or protected data it could also be used to anonymize information in order to deal with privacy issues.

3.5 Define resource naming strategy

Inputs: Schema, data

Outputs: Resource naming strategy

Description: Resources are one of the key concepts in Linked Data, and choosing the right strategy for naming them is of high importance. The global identification system on the Web is the URI, and it is considered a good practice to identify resources using URIs. Furthermore, the principles of Linked Data defined by Tim Berbers-Lee (see the beginning of Chapter 3) already state that URIs must be used for naming resources.

URIs can be used both for identifying a thing that exists outside of the Web and a web document containing the description of that thing. Since any entity (e.g., a person) can also have a Web document with its description, a clear distinction has to be made between the identifier for a web document describing the entity and the identifier for the entity itself. For example, a web page with the description of an energy company could be *http://www.energycompany.com/about*, which implies that the same URI cannot be used to identify the company itself.

Because of the previous, and having in mind the third Linked Data principle, it is important to contemplate the mechanism that allows web clients and servers to communicate, i.e., that allows a client to request a specific representation of a resource using its URI, and a server to return this representation (HTML, RDF, etc.). This mechanism is called content negotiation, and there exist different alternatives for content negotiation based on different forms of the URI.

An HTTP URI (e.g., *http://www.energycompany.com/about/energyCompany*) has different parts: protocol, domain and path structure. The protocol usually refers to the hypertext transfer protocol (*http://*); the domain is the part between the protocol and the first slash (e.g., *www.energycompany.com*) and is processed by Domain Name Servers; the path structure is the part of a URI after the domain (e.g., *about/energyCompany*) and is processed by the web server.

There are two basic forms of URIs:

- Hash URIs (#). In this case, a URI contains a fragment that is separated form the rest of the URI by a hash character ("#"). An example of this type of URI for an energy company could be *http://www.energycompany.com/about#energyCompany.* In the case of hash URIs, the fragment part has to be stripped off when the URI is requested from the server. This is because a hash URI does not necessarily identify a web document and cannot be retrieved directly; however, hash URIs can be used to identify non-document resources.
- Slash URIs (303 URIs). These URIs imply a 303 redirection to the location of a document that represents the resource. An example of this type of URI for an energy company could be *http://www.energycompany.com/about/energyCompany.* In this case, there are two possibilities for content negotiation:
    - Forwarding to one generic document. In this case, after requesting a description of a resource, the 303 redirection points to a single document that returns the appropriate description using content negotiation, as demanded by the client.
    - Forwarding to different documents. In this case, after requesting a description of a resource, the 303 redirection with content negotiation points to a document with the desired representation, which is one of several documents, all containing different representations of the resource.

When designing URIs, it is advisable to consult well-established guidelines, such as Cool URIs [Sauermann and Cyganiak, 2008], design guidelines for the UK public sector [UK Cabinet Office, 2010], ten rules for persistent URIs [Semantic Interoperability Community, 2012], or Linked Data patterns [Dodds and Davis, 2012].

Developing a resource naming strategy for a dataset is achieved in several steps:

1. To choose a URI form and an appropriate content negotiation alternative[10]. This step has to define whether hash or slash URIs are going to be used. In the case of choosing slash URIs, it is also needed to choose one of the two specified content negotiation alternatives.
2. To choose a domain for the URIs.
3. To choose a path for the URIs. The strategy for the path of a URI should include choosing the folder on the server if necessary (e.g., *www.energycompany.com/**about***), and together with this folder the domain forms the base URI.
4. To choose a pattern for ontology classes and properties in the ontology, as well as for individuals (e.g., a name of the class or property can be contained in the path of the URI of that class or property).

[10] http://www.w3.org/TR/cooluris/#choosing

Tips:

- Be unambiguous. One URI should identify only one item, regardless whether it is about a web page or a real-world object.
- URIs should be persistent and should not contain anything that can be changed (e.g., state information). This is important because changes in resources can affect the applications that depend upon them. One possible way to achieve this is to use persistent uniform resource locator (PURL[11]), which is a service for resource management and redirection settings.
- Use a domain that is under your control. Alternatively, PURL is a service that allows third party control.
- Separate the ontology model from its instances:
    - Append the string *ontology* and the ontology name to the base URI in the case of an ontology model.
    - Append the string *resource* and the ontology class name to the base URI in the case of instances.
- Define URIs in a readable manner so that people are able to understand them.

3.5.1 Energy consumption example

The resource naming strategy defined for the data coming from the BECA project follows the best practices and previously specified tips:

- According to the tips provided by [Sauermann and Cyganiak, 2008], since our dataset is rather small hash URIs will be used.
- The URI domain to be used is http://smartcity.linkeddata.es/. This domain is chosen because it is under UPM's direct control.
- A subfolder "BECA" will be added to the domain. The purpose of this is to be clear that the ontology model and data are related to the BECA example.
- Ontology model classes: following the tips, the classes in the ontology will have the form *http://smartcity.linkeddata.es/BECA/ontology/<ontologyName>#<className>*. An example for the class of dwellings in the residence ontology could be *http://smartcity.linkeddata.es/BECA/ontology/ResidenceOntology#Dwelling.*
- Ontology model properties: following the tips, the properties in the ontology will have the form *http://smartcity.linkeddata.es/BECA/ontology/<ontologyName>#<propertyName>*. An example for the *hasId* property of the *Dwelling* class in the residence ontology could be *http://smartcity.linkeddata.es/BECA/ontology/ResidenceOntology#hasId.*
- Instances: following the tips, the instances in the ontology will have the form *http://smartcity.linkeddata.es/BECA/resource/<className>#<identifier>.* An example of an instance of a dwelling (e.g., a dwelling in Belgrade) could be *http://smartcity.linkeddata.es/BECA/resource/Dwelling#347fns7ebcuwdb.*

3.5.2 Building example

As mentioned at the beginning of this section, IFC data are already based on a sound data modelling approach including a formal representation of the schema and the data. Several mapping approaches are available to transfer (1) the schema to an OWL ontology and (2) the data to ontology instances. It's a fully automatic conversion process developed on the meta-model level of EXPRESS and OWL. Depending on the use case of such auto-generated ontology, different mapping strategies are applied. If one of those strategies fits the expected use case, it is suggested to follow the proposed resource naming strategy. It would typically mean that

[11] http://www.purlz.org/

the IFC ontology is using hash URIs, ideally accessible under this address in the Web, whereas IFC instances are using slash URIs due to the amount of data. For the latter, any domain can be used to publish the data. It is worth to mention that IFC differentiates between uniquely identifiable elements and so called resource elements. In the mapping approaches published at http://linkedbuildingdata.net/ identifiable elements are using the global unique identifier in the URI, whereas resource elements are represented as anonymous individuals (blank nodes). Thus, if two IFC data sets shall be published under the same domain the node IDs of anonymous individuals may have to be changed to be unique and thus to avoid naming conflicts.

3.6 Develop ontology

Inputs: Schema, data, resource naming strategy

Outputs: Ontology

Description: By describing the concepts in a domain and the relationships between them, ontologies represent formal representations of knowledge about a certain domain and are the cornerstone of the Linked Data initiative since they are the formal models for representing data on the Web. Ontologies can be implemented in various languages; the most widely used and accepted implementation is that standardized by the W3C, the Web Ontology Language (OWL) [W3C, 2012].

Ontologies contain different components (e.g., classes, properties, instances and axioms). Ontologies denoted as lightweight contain only classes, properties, and instances. On the other hand, heavyweight ontologies are developed having in mind all the components.

Since the ontology in the energy domain is developed to represent the data that are already available in a data source, data-driven development (i.e., development which relies on already available data) has been considered; furthermore, the guidelines focus on lightweight ontologies.

With respect to these requirements, ontology development can be achieved in several consecutive steps [Poveda-Villalón, 2012]:

1. *To define requirements*. The first step of the ontology development process is to define the requirements that have to be fulfilled. These requirements can be related to the purpose of usage of the ontology, to the domain that the ontology is covering, or to technical details of the ontology, among others.
2. *To extract terms*. The second step is to extract the terms from the data source that are related to basic concepts in the data and to relationships between those concepts. In the case when the schema of the data source already exists or has been previously extracted, it can be used together with the data as a reference for terms in the data source. Furthermore, the extracted terms should consist of not only the terms from the data source, but also of the synonyms of those terms. In order to find the synonyms, some online services can be used[12].

   Once the terms are extracted, they are divided into terms for classes, terms for properties, and terms for instances. A property can be an object property (the range of a property is another class) or a data type property (the range of the property is a literal).
3. *To search existing ontologies*. Reusability is one of the main principles to follow when developing ontologies. The best practice is to reuse existing ontologies whenever possible and, therefore, it is necessary to first perform a search to find which existing ontologies best fit the previously-extracted terms.

[12] Examples of online services to search synonyms include *synonyms.com* or *thesaurus.com*

Existing ontologies are searched based on keywords in such a way that previously extracted terms (including synonyms) are searched using one or more tools in order to find ontologies in which classes and properties related to those terms are already defined.

In this step, the search results often need to be filtered because for some general terms (e.g., *person*) they can consist of several hundred ontologies and it is not possible to inspect all of them. Also, in the case when the search gives satisfying results for a specific term, it is not needed to search for its synonyms.

In those cases when widely-used ontologies are already known and can be reused with certain classes or properties, these ontologies can be selected for reuse and there is no need to perform the ontology search for the terms related to these classes or properties.

4. *To select existing ontologies*. After the search for ontologies is performed, based on the search results and on the extracted terms, the appropriate ontologies that are going to be reused are selected.

   For every extracted term, an ontology is selected for reuse in such a way that

   - The class or property in the ontology relates to the context of the searched term, i.e., the semantics of the class or property in the ontology is related to the term.
   - If the term relates to a class, the class in the ontology has as much properties that correlate to the term as possible.
   - The ontology that describes the class or property related to the search term is widely accepted and used.

5. *To draft an initial conceptualization*. Once the ontologies to be reused were selected, the next step is to integrate the concepts from the selected ontologies into an initial conceptualization. This is done by
   - Deciding upon which classes can be reused. The classes selected for reuse are those that describe the concepts that are related to a search term. Furthermore, a class can also be reused as a superclass.
   - Deciding upon which properties can be reused. Properties that are selected have to be related to search terms, i.e., to the terms extracted form the data or schema.
   - Making an initial conceptualization of the ontology that includes the selected classes and properties.

   If all the needed classes and properties are available in existing ontologies, the next step to be performed is step 8. Usually, this is not the case and steps 6 and 7 have to be performed.

6. *To define the complete conceptualization*. If existing ontologies do not provide all the information needed to represent the data, it is necessary to complete the ontology by introducing
   - New classes. New classes are introduced only in the case when existing ontologies do not describe the desired classes; these new classes have to be related to the terms extracted in the first step.
   - New properties. New properties can be introduced to newly introduced classes as well as to classes from other ontologies that are selected for reuse; these properties have to be related to terms extracted in the first step.
7. *To implement the ontology*. In order to be used in software systems, the ontology has to be implemented. When implementing the ontology, its domain and the URIs of all the classes and properties have to conform to the resource naming strategy. Ontology implementation can be done using one of the well-known ontology engineering tools.
8. *To evaluate the ontology*. Once the ontology is implemented, it has to be evaluated. Several dimensions for ontology evaluation exist (Figure 21) [Poveda-Villalón et al. 2012]:

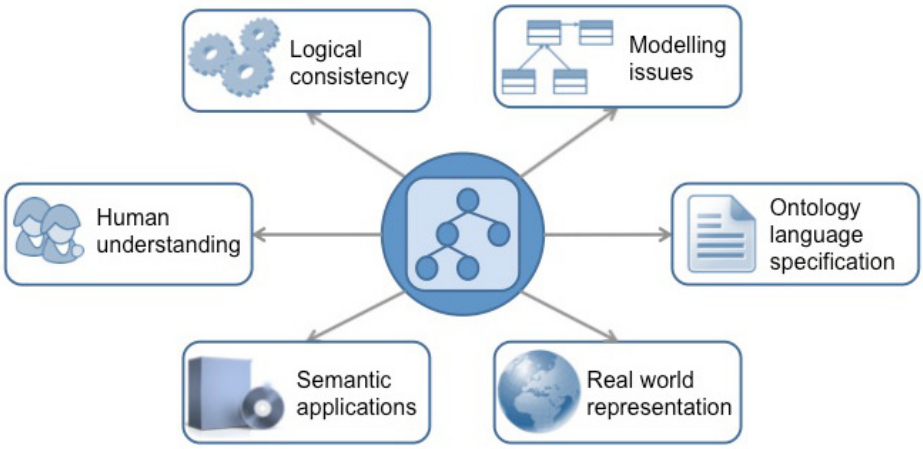


Figure 21. Ontology evaluation dimensions.

- The human understanding dimension refers to whether the ontology provides enough information so that it can be understood from a human point of view. This aspect is highly related to ontology documentation and clarity of code.
- The logical consistency dimension refers to whether (a) there are logical inconsistencies or (b) there are parts of the ontology that could potentially lead to an inconsistency but they cannot be detected by a reasoner unless the ontology is populated.
- The modeling issues dimension refers to whether the ontology is defined using the primitives provided by ontology implementation languages in a correct way, or whether there are modeling decisions that could be improved.
- The ontology language specification dimension refers to whether the ontology is compliant (e.g., syntax correctness) with the specifications of the ontology language used to implement the ontology.
- The real world representation dimension refers to how accurately the ontology represents the domain intended for modeling. This dimension should be checked by humans (e.g., ontology engineers and domain experts).
- The semantic applications dimension refers to whether the ontology is fit for the software that uses it, for example checking availability, format compatibility, etc.

Tools: Well-known and widely-used tools for searching ontologies in general are Swoogle[13], Watson[14], and LOV[15] (Linked Open Vocabularies). Furthermore, in the READY4SmartCities project we have developed a domain-specific ontology catalogue for the domain of smart cities[16]. One additional tool to search for ontologies is also Google.

Among the tools to be used for developing the ontology, well-known tools are Protégé, WebProtégé, the NeOn Toolkit, and TopBraid Composer, among others.

Several tools for ontology evaluation also exist:

[13] http://swoogle.umbc.edu/

[14] http://watson.kmi.open.ac.uk/

[15] http://lov.okfn.org/dataset/lov/

[16] http://smartcity.linkeddata.es/

- The ontology syntax can be validated using the OWL validator[17].
- A tool that can be used for automatically evaluating the mentioned dimensions is the OOPS! pitfall scanner[18], which provides mechanisms to automatically detect a number of pitfalls and thus helps developers in the diagnosis activity.
- One possibility for ontology evaluation is to use one of the available reasoners (e.g., HermiT, Fact++, etc.) in order to evaluate the consistency of the ontology, or to use the Protégé tool with a reasoner plugin.

Alternatives:

- Data-driven development, in which ontologies are developed starting from the available data (the methodology described above is a data-driven methodology).
- Competency question-driven development, in which ontologies are developed starting from competency questions. Competency questions are questions for which the answers can be found in the data described according to developed ontology; an example of a methodology based on competency questions is the NeOn methodology [Suárez-Figueroa, 2010].

Tips:

- In the case when the search does not provide satisfactory results, additional keyword-based search can be performed. To improve search results with respect to the context of the search, terms such as "ontology" can be added, or only OWL files can be searched. For example, when searching for the concept of "residence", the query can be "residence ontology" or "residence filetype:owl" (if using the Google search engine).
- Classes and properties should not be created with the goal of just allowing the use of a specific term for a concept that already exists.
- The common lexical conventions should be followed (for example, those presented by the Semantic Interoperability Community [Semantic Interoperability Community, 2013]).
- Abbreviations and acronyms should be expanded.
- The ontology should be implemented in the RDF-S or OWL languages, since these are the W3C standard languages to represent ontologies in the Web.

3.6.1 Energy consumption example

Since in the BECA example the data are already available in the Excel format, the ontology to develop for this example relies on the data-driven methodology presented above. The purpose of this ontology is to capture the knowledge about energy resources related to the BECA example and to provide a model for the representation of data from such example. Next, we describe each ontology development step carried out in the BECA example.

1. For the ontology to be developed for the BECA example, several requirements were specified:
    a. The ontology will try to adopt concepts and design patterns in other ontologies where possible (for example, a range of a property can be changed, additional classes could be introduced, etc.).
    b. The ontology should be implemented in OWL 2 DL.
2. As the schema of the BECA example is already available within the Excel spreadsheet (Section 3.3), it was used, together with available data, as the reference for the terms and their synonyms and for the

[17] http://mowl-power.cs.man.ac.uk:8080/validator/

[18] http://www.oeg-upm.net/oops

identification of classes, properties and instances. Table 1 shows the list of terms and their synonyms in brackets.

Classes
Dwelling (residence, habitat), city, building, evaluation group, tenancy (occupancy), pilot, heating degree days, hot water, cold water, heating (heat), energy, consumption (utilization), unit of measurement, month (time)
Instances
Kilowatt hour (kWh), cubic meter (cbm), square meter (sqm), thermal unit
Properties
Service setup, evaluation identifier, building identifier, building name (name), pilot name, service setup name, tenant identifier, comment, dwelling identifier, change of tenancy, vacancy of dwelling, number of persons living (number of persons), size of dwelling (size), night setback, ventilation system, participated in focus group, value

Table 1. List of terms and their synonyms for the BECA example.

3. To search for existing ontologies that describe the extracted terms, we used Swoogle, Watson, LOV, Google, and the smart cities ontology catalogue.

   Some widely-known ontologies that contain classes and properties that can be reused in the BECA example exist, and these ontologies were selected without performing the search for related terms. Example of this is the *TimeInterval* class from the DUL ontology that can be used to represent the time period to which the energy consumption relates to, as well as the *comment* property from the *RDFS schema,* the *identifier* property (to represent evaluation identifier, tenant identifier, building identifier and dwelling identifier) and the *title* property (to represent building name and pilot name) from the *Dublin Core* ontology, and the *hasValue* property from the DUL ontology (to relate the energy consumption value to energy consumption data).

   Table 2, Table 3, and Table 4 show the part of the results of the search related to classes, instances, and properties, respectively; only those terms that gave satisfying search results are showed. For every term and its synonyms shown in brackets, URIs of one or several ontology concepts that can be reused are listed.

   As an example, for the search of ontologies including the concept of dwelling, we have performed the following steps:

   - We have first used the term *dwelling* with the previously mentioned tools to search for ontologies. The search results contained more than three hundred ontologies, and we have included only a number of those that are available (excluding links with errors and no content) and that can be used to represent the concept of dwelling.
   - We have also performed the search using the synonyms of the term dwelling and using the same search tools as when searching for the term dwelling. In this case, the search results contained more than six hundred results.

Term	Ontology	Class
Dwelling (residence, habitat)	**http://schema.org/**	Residence
	http://www.semanticweb.org/ontologies/2012/2/OpenStreetMapFeatures.owl#	Isolated_Dwelling
	http://www.semanticweb.org/ontologies/2011/5/Ontology1307456124031.owl#	RESDW
	http://www.cyc.com/2003/04/01/cyc#	ModernHumanResidence
Building	**https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/gbBuildingOntology.owl#**	Building
	http://contextus.net/ontology/ontomedia/misc/location/	Building
Heating degree days	https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/gbBuildingOntology.owl#	HDD
Energy	**https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/EnergyResourceOntology.owl#**	UsefulEnergy
	http://schema.org/	Energy
Time	http://www.loa-cnr.it/ontologies/DUL.owl#	TimeInterval
City	**http://schema.org/**	City
	http://purl.org/ontology/places#	City
	http://sw-portal.deri.org/ontologies/swportal#	City
	http://www.loc.gov/standards/mads/rdf/v1.html#	City
Unit of measurement	**http://www.wurvoc.org/vocabularies/om-1.8/**	Unit_of_measure
	http://idi.fundacionctic.org/muo/muo-vocab.html#	UnitOfMeasurement
	http://purl.oclc.org/NET/muo/muo#	MetricUnit
	http://www.loa-cnr.it/ontologies/DUL.owl#	UnitOfMeasure

Table 2. Terms from the BECA data and the related classes in existing ontologies.

Term	Ontology	Instance
Kilowatt-hour	**http://www.wurvoc.org/vocabularies/om-1.8/**	kilowatt_hour
	http://qudt.org/1.1/vocab/unit#	Kilowatthour
	http://purl.obolibrary.org/obo/	UO_0000224
Cubic meter	**http://www.wurvoc.org/vocabularies/om-1.8/**	cubic_metre
	http://qudt.org/1.1/vocab/unit#	CubicMeter
	http://purl.obolibrary.org/obo/	UO_0000089
Square meter	**http://www.wurvoc.org/vocabularies/om-1.8/**	square_metre
	http://qudt.org/1.1/vocab/unit#	SquareMeter
	http://purl.obolibrary.org/obo/	UO_0000080
Heating	https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/EnergyResourceOntology.owl#	Heat

Table 3. Terms from the BECA data and the related instances in existing ontologies.

Term	Ontology	Property
evaluation id	http://purl.org/dc/terms/	identifier
building id	http://purl.org/dc/terms/	identifier
dwelling id	http://purl.org/dc/terms/	identifier
tenant id	http://purl.org/dc/terms/	identifier
building name (name)	http://purl.org/dc/terms/	title
comment	http://www.w3.org/2000/01/rdf-schema#	comment
value	http://purl.oclc.org/NET/ssnx/ssn#	hasValue

Table 4. Terms from the BECA data and the related properties in existing ontologies.

4. Several ontologies were found in the previous step. The *schema.org* one provides a class for describing residences, which can be used for dwelling description, and includes a number of properties to describe it (e.g., address and geographical coordinates, among others). In this case, *schema.org* was selected to be used because it is widely-known and an accepted vocabulary. An ontology that was reused for describing buildings is the *gbBuilding Information Ontology (BIO)*. BIO provides some additional classes and properties that can be used for building description.

   Several ontologies were also found to represent the concept of city. In this case, *schema.org* was selected because it is a well-known vocabulary and was also already selected to represent dwellings (residences).

Only one ontology was found in the case of the heating degree days concept and, therefore, this ontology was selected for reuse. The general concept of energy was found in two ontologies, *Energy Resource Ontology (ERO)* and *schema.org*; however, since the *UsefulEnergy* class from ERO is semantically closer to the context of the BECA example, and since ERO also describes some instances of the mentioned class that are of interest for the BECA example (e.g., *Heat*), it was selected for reuse.

The concept of heating is found in the ERO ontology as an instance of the *UsefulEnergy* class. Because of this, although this concept is initially searched as a class in the ontology, it will be represented using the *Heat* instance from the ERO ontology.

To represent energy consumption data (measured values), the widely-known *Semantic Sensor Network Ontology* (SSN) from the W3C was reused since it provides a set of classes for representing measurements (*Observation, SensorOutput, ObservationValue, Property*). The key concept is the *Observation* class, which allows capturing the BECA consumption values and provides relationships for capturing the time period related to these values. To capture the time period to which the captured values relate to, the widely-accepted DOLCE ontology is reused.

In the case of units of measurement and their instances, several ontologies were found. The *Ontology of units of Measure (OM)* was selected in all the cases because, unlike other ontologies, it describes both the classes and instances needed to represent the units in the BECA example.

In those cases where more than one ontology was found for a specific term, the ontology that was selected for reuse is marked in bold in Tables 2-4.

5. Figure 22 shows the initial conceptualization of the ontology related to buildings and observations (energy consumptions), together with the classes, properties and instances that were selected for reuse. Figure 23 shows the initial conceptualization of the ontology related to units of measurement. All the reused elements contain the reused ontology namespace before their names, and Table 5 shows the mappings between the ontology URIs and their namespaces. Instances and the *rdf:type* property are represented with a discontinuous line. Primitive data types are represented with rectangles.

Ontology	Namesace	URI
gbBuilding Information Ontology	bio	https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/gbBuildingOntology.owl#
DOLCE+DnS Ultralite	dul	http://www.loa-cnr.it/ontologies/DUL.owl#
Dublin Core	dc	http://purl.org/dc/terms/
Energy Resource Ontology	ero	https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/EnergyResourceOntology.owl#
Ontology of units of Measure	om	http://www.wurvoc.org/vocabularies/om-1.8/
RDFS Schema	rdfs	http://www.w3.org/2000/01/rdf-schema#
Schema	schema	http://schema.org/
Semantic Sensor Network	ssn	http://purl.oclc.org/NET/ssnx/ssn#
XML Schema	xsd	http://www.w3.org/2001/XMLSchema#

Table 5. Mapping between the ontology URIs and their namespaces.

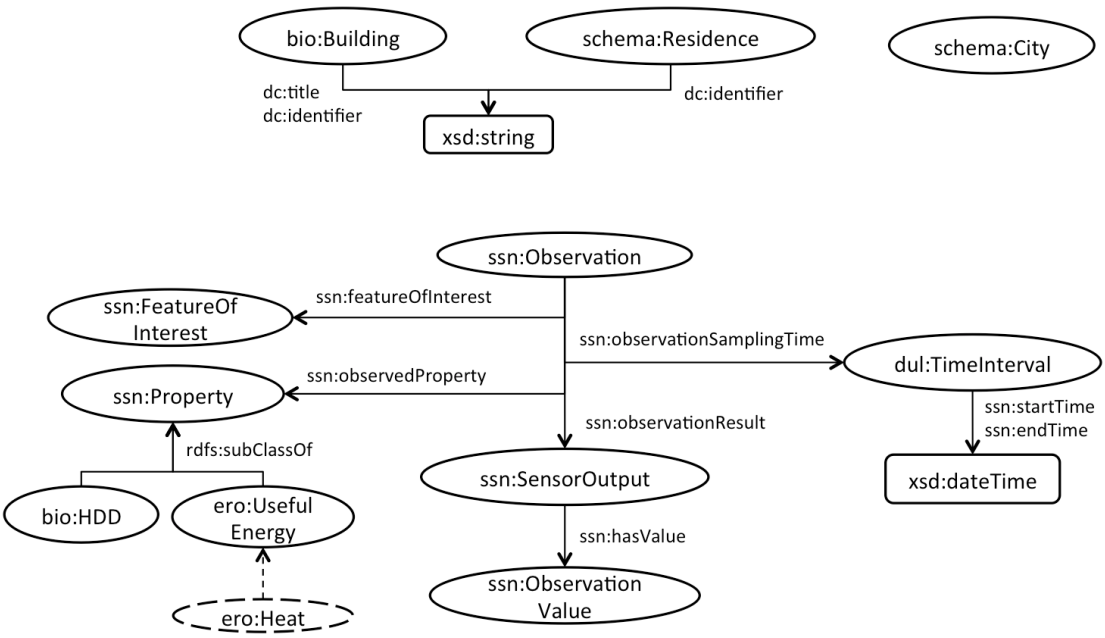


Figure 22. Initial conceptualization of the ontology related to buildings and observations.

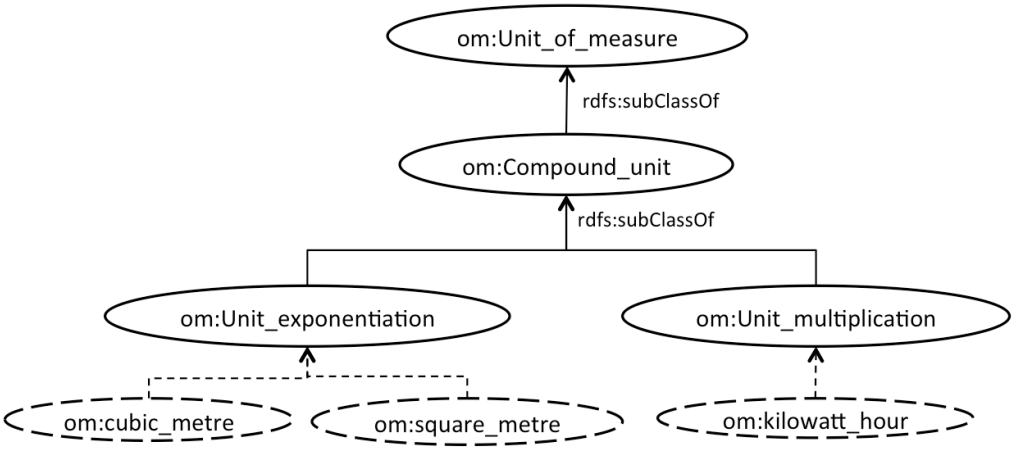


Figure 23. Initial conceptualization of the ontology related to the units of measurement.

6. Since the search for existing ontologies did not provide results for all extracted terms and their synonyms, it was necessary to complete the ontology. Therefore, several classes, properties, and instances were introduced. Due to space reasons, the complete ontology conceptualization is shown in several figures. Figure 24 shows the part of the ontology related to tenancies; Figure 25 shows the part related to residences; Figure 26 shows the part related to observations and features of interest; Figure 27 shows the part related to observation values and units of measurement; finally, Figure 28 shows the part related to properties. Similarly as in the previous case, classes, properties and instances that are reused contain the reused ontology namespace before their names; instances and the *rdf:type* property are represented with a discontinuous line; and primitive data types are represented with rectangles.

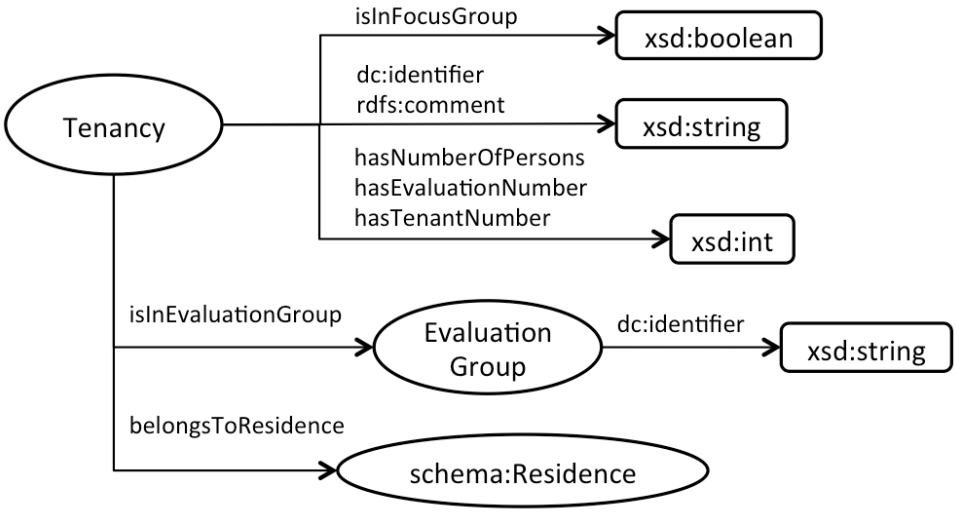


Figure 24. Part of the complete ontology related to tenancies.

The *evaluation identifier* term denotes an identifier of an evaluation run over a specific tenancy. Although the *identifier* property from the *Dublin Core* ontology is identified for reuse in this case, since tenancies have their own identifiers represented with the same property, a new property *hasEvaluationNumber* was introduced.

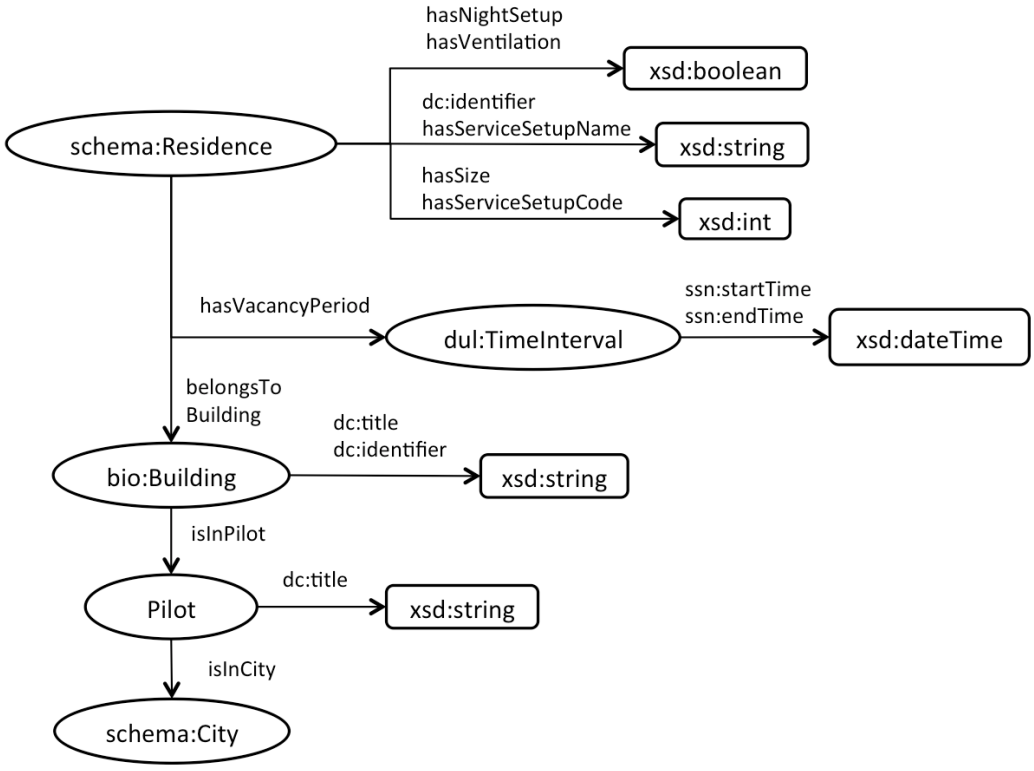


Figure 25. Part of the complete ontology related to residences.

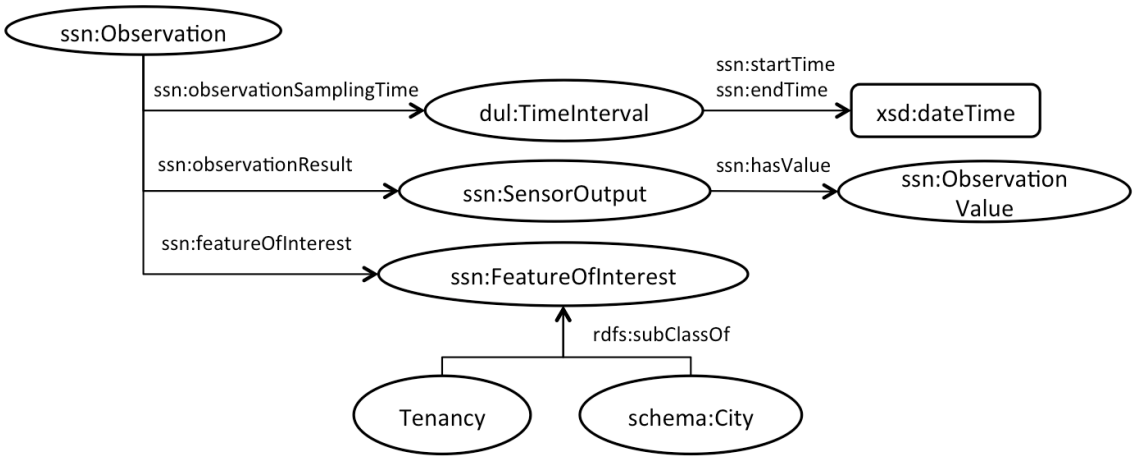


Figure 26. Part of the complete ontology related to observations and features of interest.

One of the measurement units used in the BECA example is the thermal unit. Although a few instances related to thermal units exist in the ontologies listed in Table 4, they are related to different specifications and since the data from the BECA project is not sufficient to determine to which specification the thermal units measured in the project refer to, a new instance was introduced.

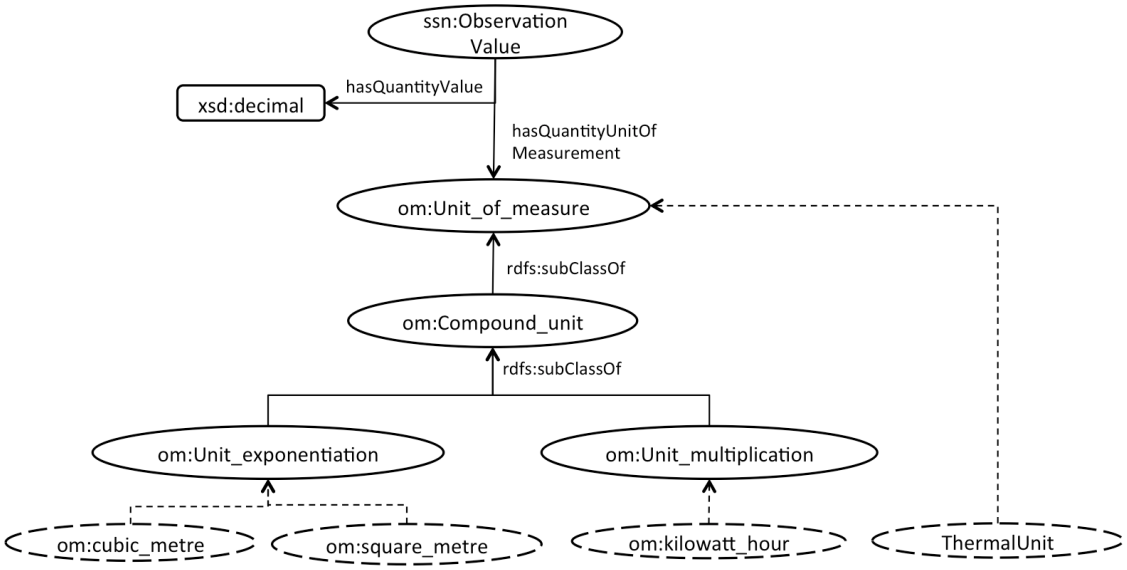


Figure 27. Part of the complete ontology related to observation values and units of measurement.

In the case of energy data (heating, cold water and hot water), as heating will be represented with an instance from an already existing ontology, the remaining energy categories were also modeled as instances in the ontology.

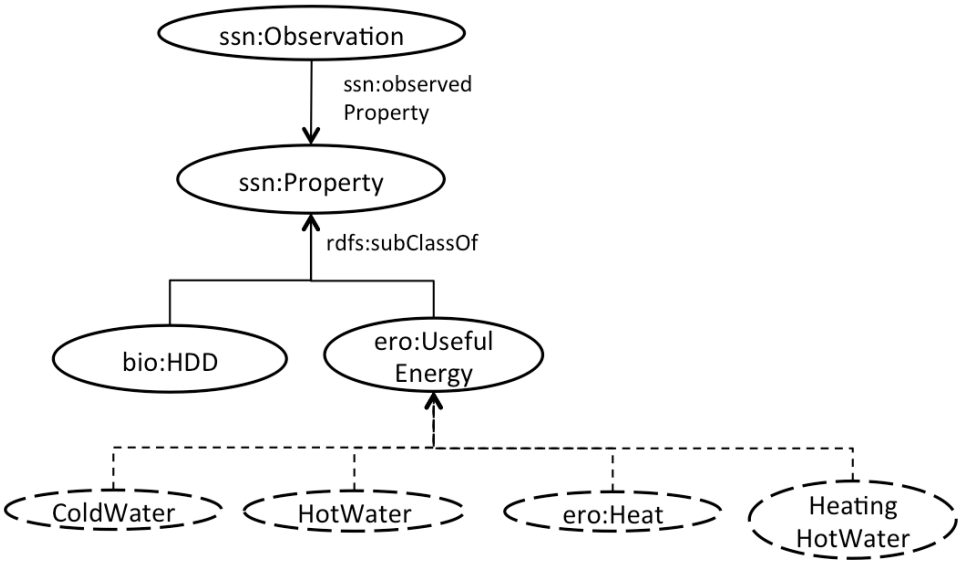


Figure 28. Part of the complete ontology related to properties.

7. The ontology developed for the BECA example was implemented in OWL using the Protégé tool. The implemented ontology is available online at the following link: http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption.owl.
8. The ontology developed for this example was evaluated using the OOPS! pitfall scanner. Several errors were found, both minor and important ones. Through several evaluation iterations, the important errors were corrected and only one minor warning remains in the current version of the ontology. This warning is related to classes and properties that miss annotations. In this case, these are the classes and properties that are reused and, therefore, these annotations were purposely omitted.

   During the evaluation process, in order to correct some important pitfalls, a set of axioms was added to the ontology. Because of this, the resulting ontology is heavyweight, which is not in line with the initial guideline requirements; however, this step of defining ontology axioms was performed in order to provide an ontology of higher quality.

   Furthermore, the syntax of the ontology was also validated, since OOPS! only works with ontologies with valid syntax, and we have used the Falcon reasoner in order to evaluate the logical consistency of the developed ontology.

3.6.2 Building example

The IFC standard is a result of many years of development, in particular in reaching consensus about how to represent building data that is to be integrated and shared throughout its lifecycle. Lots of resources went not only into specification work but also into implementation and certification of tools as well as all kinds of training activities. What is missing is an agreed W3C-ifcOWL representation, which needs to be derived from the IFC-EXPRESS reference schema. It may need to be configured and extended by additional knowledge to meet the requirements of a specific use case. Accordingly, besides choosing a proper mapping approach it has to be decided if additional knowledge on the schema level needs to be added. For instance, the following mapping approaches are already available to convert IFC EXPRESS to different flavours of OWL:

- NIST OntoSTEP Converter[19] (plug-in for Protégé):
  creates an OWL 1 DL representation
- IFC2X3 - University of Ghent[20]
  creates an OWL 1 Full representation
- Linked Building Data[21] initiative (in collaboration with University of Ghent)
  The following configurations for OWL 1 and OWL 2 are available:
  - OWL 1 - DL (two configurations)
  - OWL 2 - EL (two configurations)
  - OWL 2 - Full (one configuration)
  - OWL 2 - RL (two configurations)

The list of mapping approaches is even longer. Thus, a lot of work has already been done in that area that can be used without further effort. It is worth mentioning that those mapping approaches will lose some knowledge encoded in the IFC schema as not all modelling concepts from EXPRESS have a direct equivalent in OWL. Also, as those solutions are based on a fully automatic mapping approach there is no additional knowledge encoded in the generated ifcOWL representations.

There might be the need to simplify the IFC schema in order to support a specific use case. This can be done by reducing definitions, e.g., to exclude specific domains that are out of scope, or by re-organizing class definitions, its attributes and relationships. The former would be interesting if BIM data should be as close as possible to standardized IFC whereas the latter would be interesting if BIM data can be mapped to a simplified target ontology (that might be already available) as used for instance in the example discussed in [Curry et al. 2013]. However, any kind of optimization (or simplification) requires having understanding of the use case and thus is not a universal approach for BIM/IFC datasets.

3.7 Transform data source

Inputs: Data, resource naming strategy, ontology

Outputs: RDF data

Description: After the ontology has been developed, the data from the selected data source can be transformed into the RDF format. This task can be achieved in several consecutive steps:

1. To select the RDF serialization. While several serializations of RDF exist, no serialization is better than other and the benefits of using a specific serialization may include simplicity, speed of processing, and readability by humans:
   - RDF/XML[22] is a serialization based on the XML specification. This serialization is considered to be less readable by humans.
   - Turtle[23] is a compact format that is considered to be more readable by humans, and it can be processed quickly.
   - N-Triples[24] is a line-based format that is very easy to read by humans, and it can also be processed quickly.

[19] http://www.nist.gov/el/msid/ontostep.cfm

[20] http://multimedialab.elis.ugent.be/organon/ontologies/IFC2X3#

[21] http://linkedbuildingdata.net.previewdns.com/tools/tool-ifc-to-rdf-conversion-tool/

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

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

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

- JSON-LD[25] is a serialization that is based on the widely-accepted JSON standard.

All previously mentioned serializations are W3C recommendations[26] (i.e., standards).

2. To select a tool for data transformation. Several tools exist for the purpose of transforming the data from various formats into RDF. The tool to be selected depends on the format of the data (database, spreadsheets, etc.), and on concrete needs of the transformation process (e.g., dynamicity).
3. To use the selected tool in order to obtain the RDF. The tool to be used usually requires a mapping between the data and the ontology. Furthermore, such mapping specifies the naming of all instances in a dataset according to the resource naming strategy defined.

   In this deliverable, concrete guidelines are not given for this step since every tool is different. However, one example of using a particular tool can be found in the example for this task.

4. To evaluate the obtained RDF dataset. Several approaches for the evaluation of Linked Data exist. In the context of Linked Data generation in the energy domain, the evaluation could consist of:
   - Validation of the syntax of the RDF produced.
   - Completeness evaluation by examining whether the dataset contains missing values for a specific property.
   - Licensing evaluation, which includes
     - Checking whether the dataset contains machine-processable indication of a license.
     - Checking whether the dataset contains human-readable indication of a license.
   - Accuracy evaluation, which consists of checking for literals incompatible with the data type ranges.
   - Conciseness evaluation, which includes
     - Checking whether the dataset contains redundant objects, i.e., if it contains any pair of two equivalent objects with different identifiers.
     - Checking whether the dataset contains duplicate entries.
   - Evaluation of representational consistency, which consists of checking whether the dataset uses existing established ontologies to represent its entities.
   - Understandability evaluation, which includes checking whether the dataset contains human-readable labels of instances.
   - Versatility evaluation, which includes checking whether the data are represented in various languages.
   - Usage evaluation, which allows determining whether a dataset provides possibilities for answering certain questions and obtaining the necessary information (e.g., in terms of SPARQL queries).

Tools: The tools that can be used for transforming data into RDF, depending on the format of the data source, are the following:

- Transforming databases into RDF
  - morph-RDB[27] is an RDB2RDF engine that follows the W3C R2RML specification. Apart from data generation, this tool also supports query translation. morph-RDB can be used with MySQL, PostgreSQL and MonetDB.

[25] http://www.w3.org/TR/rdf-json/

[26] More details about the RDF and serialization formats can be found on http://www.w3.org/TR/rdf11-primer/

[27] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/315-morph-rdb

- D2R Server[28] is a tool that creates instances of an ontology on demand. It contains a declarative language for describing relationships between an ontology and a relational model; resulting instances are expressed in RDF.
- TopBraid Composer[29] is another tool that can be used for RDF data generation from databases, in which the rows in a table of a database become instances.

- Transforming data streams into RDF
  - morph-streams[30] is a tool for generating in real time RDF data from streams. It also provide functionalities for querying such data.
  - D2R server is often used for this task. In this case, data from data streams are first stored in a database, and D2R Server is then used to generate RDF data from this database.
- Transforming XML files into RDF
  - XML2RDF[31] is a tool that supports the transformation of XML files into ontology instances.
  - TopBraid Composer also supports the conversion of XML files into ontology instances.
  - LODRefine[32] is a tool based on OpenRefine[33] that adds several extensions. It can be used to transform data from one format into another, XML and RDF among them.
- Transforming spreadsheets into RDF
  - TopBraid Composer supports the conversion of data from Excel spreadsheets.
  - Excel2rdf[34] is a java-based command-line utility for converting Excel files to RDF.
  - RDF123[35] is an open source tool for transforming spreadsheet data into RDF that works with CSV files and Google spreadsheets also. This tool provides a public web service for performing the translation task.
  - XLWrap[36] is a tool that supports the transformation of Excel and OpenDocument spreadsheets and provides the possibility of either loading local files or using remote files via HTTP.
  - LODRefine and OpenRefine are tools that, besides XML, also support Excel spreadsheets and RDF.

The tools that can be used for the evaluation of Linked Data are:

- W3C RDF validator[37], which validates the syntax of RDF data.
- Databugger[38] is a tool that uses a SPARQL endpoint for verification of data against a schema.

[28] http://d2rq.org/d2r-server

[29] www.topbraidcomposer.com

[30] https://github.com/jpcik/morph-streams

[31] http://rhizomik.net/html/redefer/

[32] http://code.zemanta.com/sparkica/index.html

[33] http://openrefine.org/index.html

[34] https://github.com/waqarini/Excel2rdf

[35] http://rdf123.umbc.edu/

[36] http://xlwrap.sourceforge.net/

[37] http://www.w3.org/RDF/Validator/

[38] http://databugger.aksw.org:8080/rdfunit/

3.7.1 Energy consumption example

We have used the data from the BECA Excel spreadsheet for the generation of RDF data. According to the guidelines, we have performed the following steps:

1. We have selected the Turtle serialization for this example. Although the dataset is small and the speed of processing is not an issue, Turtle was selected because it is easy to read by humans.
2. Since the data are available in an Excel spreadsheet, we have selected OpenRefine for transforming the data into RDF. This tool was selected because it is easy to use and it is widely-known in the community.
3. We have used the selected tool to generate RDF Data from the BECA Excel spreadsheet. Using OpenRefine, this can be achieved in several steps:
    a. First, the Excel spreadsheet has to be imported into the working environment (Figure 29). In this step, various options exist for making initial transformations of the spreadsheet (e.g., choosing the worksheet to import, or excluding a number of rows from the beginning, among others).
    b. The RDF extension for OpenRefine allows creating the mappings between the columns and rows in the spreadsheet and the ontology (Figure 30). For each cell in a column that has to be transformed into an RDF instance, it is possible to specify the URI of the instance and its type, as well as to specify the properties for that instance and the values of those properties.

Column7	Column8	Column9	Column10	Column11	Column12	Column13
Eval ID (combined) matching var for other maps	Pilot	Building	Kind of Setup		Internal Dwelling Number (Dwelling_ID)	Tenant ID or Name
	"Name"	"Name"	"Name"	Suchkriterium (corresponding to tenant ID but without blanks)	Unique Dwelling ID within each Site	changes if new tenant moves in (insert new row for new ID)
710816171	[illegible]	[illegible]	TOP SET UP_HW+HE	O_002 0001_1	-77	O_002 0001_1
710816181	[illegible]	[illegible]	TOP SET UP_HW+HE	O_0020002_1	-77	O_002 0002_1
710816191	[illegible]	[illegible]	TOP SET UP_HW+HE	O_0020003_1	-77	O_002 0003_1

Figure 29. OpenRefine working environment.

In this step, manipulation of the table can be performed in order to make the mapping process easier to complete. For example, new columns can be added with values based on existing columns, values of existing cells can be changed based on some pattern, etc.

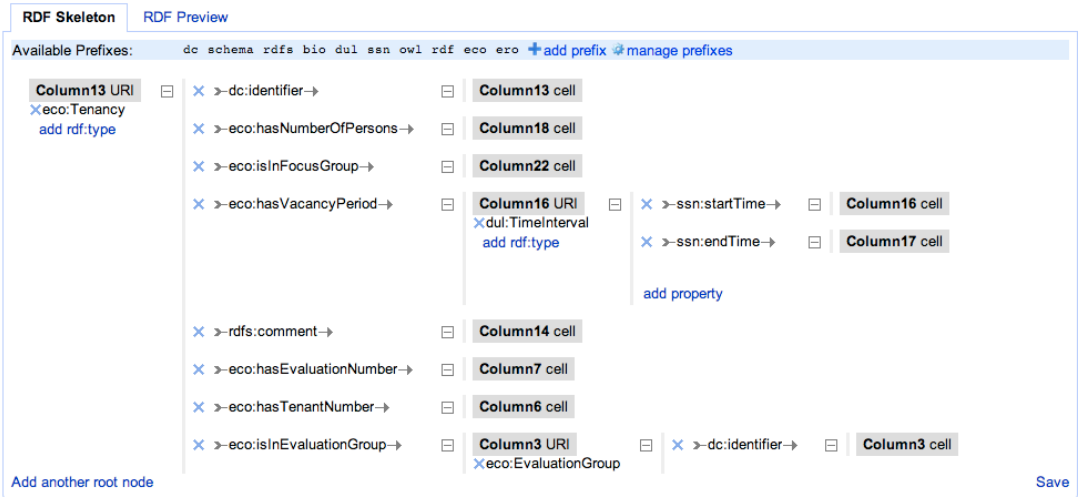


Figure 30. RDF extension mapping in OpenRefine.

c. Once the mapping is finished, the Linked Data can be generated. The RDF extension gives the possibility to choose the syntax of the RDF data, and the data are then stored in a file.
d. If needed, the generated data has to be cleaned. This step is performed in those cases when the imported spreadsheet contains rows with values that are not intended to be generated as RDF data (e.g., table headers).

4. Since the primary goal of the BECA dataset is to provide data about energy consumption from BECA pilot sites, we have evaluated the generated RDF dataset by performing SPARQL queries over it and examining the retrieved information. For this purpose, all the generated data were imported into the Virtuoso[39] tool, and several queries were performed:
   a. Select all residences in building "108" that have three persons living in it. Figure 31 shows the SPARQL query executed, while Figure 32 shows the results obtained for this query.

```
PREFIX schema: <http://schema.org/>
PREFIX eco: <http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#>
PREFIX bio:<https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/gbBuildingOntology.owl#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

Select ?residenceId ?tenantId ?number WHERE {
    ?tenancy a eco:Tenancy;
    eco:hasNumberOfPersons ?number;
    dc:identifier ?tenantId;
    eco:belongsToResidence ?residence.
    FILTER (?number = 3)

    ?residence a <http://schema.org/Residence> ;
    dc:identifier ?residenceId;
    eco:belongsToBuilding ?building.

    ?building a bio:Building;
    dc:identifier "108".
}
```

Figure 31. SPARQL query for residences with three persons in building "108".

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

residenceId	tenantId	number
"1617"	"O00200011"	3
"1622"	"O00200061"	3
"1626"	"O00200101"	3
"1637"	"O00200211"	3
"1638"	"O00200221"	3
"1642"	"O00200261"	3
"1647"	"O00200311"	3
"1654"	"O00200381"	3
"1655"	"O00200391"	3
"1656"	"O00200401"	3

Figure 32. Query results - residences with three persons in building "108".

b. Select all cold water consumptions of tenancy "M3600100011". Figure 33 shows the SPARQL query executed, while Figure 34 shows an excerpt with 17 records of the results obtained for this query.

```
PREFIX schema: <http://schema.org/>
PREFIX eco: <http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#>
PREFIX bio: <https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/gbBuildingOntology.owl#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ssn: <http://purl.oclc.org/NET/ssnx/ssn#>
PREFIX dul: <http://www.loa-cnr.it/ontologies/DUL.owl#>

Select ?startTime ?endTime ?consumption WHERE {
    ?tenancy a eco:Tenancy;
    dc:identifier "M3600100011".

    ?observation a ssn:Observation;
    ssn:featureOfInterest ?tenancy;
    ssn:observedProperty eco:ColdWater;
    ssn:observationSamplingTime ?time;
    ssn:observationResult ?sensor.

    ?sensor a ssn:SensorOutput;
    ssn:hasValue ?value.

    ?value a ssn:ObservationValue;
    eco:hasQuantityValue ?consumption.

    ?time a dul:TimeInterval;
    ssn:startTime ?startTime;
    ssn:endTime ?endTime.
}
```

Figure 33. SPARQL query for cold water consumption of tenancy "M3600100011".

startTime	endTime	consumption
2011-03-31T22:00:00+02:00	2011-04-30T21:59:59+02:00	9
2012-03-31T22:00:00+02:00	2012-04-30T21:59:59+02:00	5.8
2013-03-31T22:00:00+02:00	2013-04-30T21:59:59+02:00	2.9
2011-07-31T22:00:00+02:00	2011-08-31T21:59:59+02:00	0.9
2012-07-31T22:00:00+02:00	2012-08-31T21:59:59+02:00	0.3
2013-07-31T22:00:00+02:00	2013-08-31T21:59:59+02:00	0
2011-11-30T22:00:00+02:00	2011-12-31T21:59:59+02:00	13.6
2012-11-30T22:00:00+02:00	2012-12-31T21:59:59+02:00	10.5
2011-01-31T22:00:00+02:00	2011-02-28T21:59:59+02:00	12.2
2012-01-31T22:00:00+02:00	2012-02-29T21:59:59+02:00	10.4
2013-01-31T22:00:00+02:00	2013-02-28T21:59:59+02:00	8.3
2010-12-31T22:00:00+02:00	2011-01-31T21:59:59+02:00	11.7
2011-12-31T22:00:00+02:00	2012-01-31T21:59:59+02:00	8.2
2012-12-31T22:00:00+02:00	2013-01-31T21:59:59+02:00	9.8
2011-06-30T22:00:00+02:00	2011-07-31T21:59:59+02:00	0.8
2012-06-30T22:00:00+02:00	2012-07-31T21:59:59+02:00	0.5
2013-06-30T22:00:00+02:00	2013-07-31T21:59:59+02:00	4.5

Figure 34. Excerpt of query results - cold water consumption of tenancy "M3600100011".

c. Select all consumptions and related time periods of tenancy "M3600100011". Figure 35 shows the SPARQL query executed, while Figure 36 shows an excerpt with 15 records of the results obtained for this query.

```
PREFIX schema: <http://schema.org/>
PREFIX eco: <http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#>
PREFIX bio: <https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/gbBuildingOntology.owl#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ssn: <http://purl.oclc.org/NET/ssnx/ssn#>
PREFIX dul: <http://www.loa-cnr.it/ontologies/DUL.owl#>

Select ?property ?startTime ?endTime ?consumption WHERE {
    ?tenancy a eco:Tenancy;
    dc:identifier "M3600100011".

    ?observation a ssn:Observation;
    ssn:featureOfInterest ?tenancy;
    ssn:observedProperty ?property;
    ssn:observationSamplingTime ?time;
    ssn:observationResult ?sensor.

    ?sensor a ssn:SensorOutput;
    ssn:hasValue ?value.

    ?value a ssn:ObservationValue;
    eco:hasQuantityValue ?consumption.

    ?time a dul:TimeInterval;
    ssn:startTime ?startTime;
    ssn:endTime ?endTime.
}
```

Figure 35. SPARQL query for all energy consumptions of tenancy "M3600100011".

http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2012-04-30T22:00:00+02:00	2012-05-31T21:59:59+02:00	6.5
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2013-04-30T22:00:00+02:00	2013-05-31T21:59:59+02:00	6.69
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2011-10-31T22:00:00+02:00	2011-11-30T21:59:59+02:00	10.9
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2012-10-31T22:00:00+02:00	2012-11-30T21:59:59+02:00	9.7
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2011-09-30T22:00:00+02:00	2011-10-31T21:59:59+02:00	8.5
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2012-09-30T22:00:00+02:00	2012-10-31T21:59:59+02:00	6.7
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2013-09-30T22:00:00+02:00	2013-10-31T21:59:59+02:00	7.29
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2011-08-31T22:00:00+02:00	2011-09-30T21:59:59+02:00	4.5
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2012-08-31T22:00:00+02:00	2012-09-30T21:59:59+02:00	3.4
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#ColdWater	2013-08-31T22:00:00+02:00	2013-09-30T21:59:59+02:00	3.6
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#HeatingHotWater	2011-03-31T22:00:00+02:00	2011-04-30T21:59:59+02:00	158
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#HeatingHotWater	2012-03-31T22:00:00+02:00	2012-04-30T21:59:59+02:00	488
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#HeatingHotWater	2013-03-31T22:00:00+02:00	2013-04-30T21:59:59+02:00	423
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#HeatingHotWater	2011-07-31T22:00:00+02:00	2011-08-31T21:59:59+02:00	7
http://smartcity.linkeddata.es/BECA/ontology/EnergyConsumption#HeatingHotWater	2012-07-31T22:00:00+02:00	2012-08-31T21:59:59+02:00	2

Figure 36. Excerpt of query results - all energy consumption of tenancy "M3600100011".

The performed queries gave satisfying results, which were checked against the BECA excel spreadsheet.

For the validation of syntax, the W3C syntax validator could not be used because the size of the dataset is too big. However, as the data were successfully imported into Virtuoso, the conclusion is that the syntax of the produced RDF data is valid.

3.7.2 Building example

In general, conversion of IFC data to RDF is similar to mapping IFC EXPRESS to OWL. Based on the chosen configuration (or OWL profile) the data are imported from an IFC file, mapped and finally re-exported as RDF data. If not all IFC data should be mapped to RDF, for instance to exclude confidential or private data, then a subset of the IFC dataset has to be generated before starting the mapping process. This could be done by using predefined filters or by editing the IFC data in a BIM tool. Depending on the quality and completeness of the IFC data source as well as the chosen mapping approach further data evaluation might be skipped or may be done before the data mapping, in particular if all constraints and rules of the original IFC schema shall be checked and if no further knowledge is added to the ifcOWL representation.

3.8 Link with other datasets

Inputs: Ontology, RDF data

Outputs: Linked dataset

Description: The key concept in Linked Data are links to data from other datasets; these links ensure that datasets are not just isolated data islands.

The term data linking denotes connecting different data from various heterogeneous data sources. Several categories are defined for the problem of linking [Nikolov et al. 2011]:

- *Value matching* consists of determining if two values of two properties expressed in different ways are equivalent. This can be because of different formatting or because of the use of synonyms. However, a value match of two properties of two instances does not always denote an equivalence of those instances, but it is a starting point for deducing this equivalence.

  Value matching is based on various string similarity measures and various tools implement these measures. Furthermore, some tools also rely on keyword-based search (e.g., using Google) as an indicator of the similarity.

Another technique for value matching involves the discovery of semantic relations between the two values and deducing the similarity between them. Tools that use this approach rely on external knowledge resources (e.g., WordNet or high level vocabularies such as OpenCyc).

- *Individual matching* consists of taking two instances from different datasets and deciding whether these instances can be linked or not. This decision is taken based on the descriptions of these instances which helps in determining whether the instances refer to the same object.

  Individual matching can be performed by aggregating the results obtained with a value matching technique and various approaches for aggregation tasks have been developed. Another approach consists of using the transitivity of the *owl:sameAs* property and deducing the equivalence between the instances by examining the chains of this property.

  Deducing the links between individuals can be based also on ontology mappings, where datasets described according to different ontologies can be interlinked using such mappings.

- *Dataset matching* consists of utilizing a set of potential individual mappings between the two datasets. This way, the problem where information about individual descriptions is insufficient or incomplete is solved; datasets are analysed as a whole, often using a whole network of repositories and publicly available data.

The task of data linking (value matching, individual matching and dataset matching) can be achieved in several consecutive steps:

1. To identify classes whose instances can be the subject of linking.
2. To identify datasets that may contain instances for the previously-identified classes.
3. To select the tools for performing the task. Different tools for data linking exist, and each tool has its advantages and provides different functionalities for certain matching tasks. However, in some cases, the linking can be performed manually (e.g., when the generated dataset is small, or when the number of instances to link is low), and the next step is not necessarily performed.
4. To use the tool in order to obtain links. Different tools are used differently, and each tool requires configuration from the user in a specific form.

Tools: Tools that can be used for data linking include:

- LN2R[40] is an instance matching system that consists of logical and numerical matching engines. The logical matching engine exploits ontology axioms, after which the numerical matching engine analyses the similarity of different entities. LN2R supports all three categories of matching.
- LD mapper[41] is a dataset linking tool based on Prolog. It uses a similarity aggregation algorithm and requires user configuration in order to work properly. LD mapper supports value matching and individual matching.
- Silk[42] is a tool for interlinking datasets and maintaining the links obtained. Besides string matching techniques, Silk uses a variety of similarity measures that are configured by the user. It takes two datasets as inputs using their SPARQL endpoints and supports value matching and individual matching.
- LIMES[43] is a semi-automatic interlinking tool which is configured using XML files. The novelty of this tool is the usage of optimization techniques in order to minimize the number of comparisons that are

[40] https://sourcesup.renater.fr/projects/ln2r-lt/

[41] http://sourceforge.net/p/motools/code/HEAD/tree/

[42] http://sourceforge.net/projects/silk2/

[43] http://lod2.eu/Project/LIMES.html

needed in order to match the two datasets. It is available both as a web service and through a user interface. LIMES only supports value matching.

- RDF-AI[44] is a tool that produces a set of links containing *owl:sameAs* relationships. The tool uses an XML file with the specification of matching parameters, such as the dataset ontology structure or the matching technique. RDF-AI supports all three categories of matching.
- Serimi[45] is a tool that uses a two-phase approach. In the first phase, the candidate resources to be linked are selected according to information retrieval strategies. In the second phase, the descriptions of candidate resources are examined using an algorithm in order to determine which candidates are the most suitable for linking. Serimi supports value matching and dataset matching.

3.8.1 Energy consumption example

The linking of the BECA RDF dataset with other datasets was performed through the previously described steps:

1. The classes whose instances can be subjects of linking were identified first. In the case of the ontology developed for the BECA example (Section 3.6), there is only one identified class, which is the *City* class from *schema.org*.
2. Dbpedia[46] is a database containing the structured content from the Wikipedia pages in Linked Data format, it was identified as the dataset that might contain relevant instances of the *City* class.
3. Since we have identified only one class whose instances can be linked to Dbpedia dataset, and since the data in the BECA example contain only one instance of the mentioned class (that of the city of Torino, to which all pilot sites in the data belong to), we have performed the linking manually. The instance of this class in the BECA RDF data (http://smartcity.linkeddata.es/BECA/resource/City#Torino) is linked to the instance in Dbpedia (http://dbpedia.org/page/Turin) using the *owl:sameAs* property.

3.8.2 Building example

Linking BIM data with other data sources typically requires a manual setup of linking rules. Those linking rules then enable to link two datasets without further user interaction. Ideally, they are not limited to one specific BIM dataset and thus can be used to link other BIM datasets too. However, this is not always the case. If for instance a specific naming convention is used for a specific building only, then all linking rules based on that naming convention are limited to that BIM dataset. Furthermore, linking rules may require to make sophisticated data conversions or to combine more than one value in order to identify links. For instance it could be necessary to convert a postal address to geo coordinates or vice-versa. Also, it might be possible that linking-rules are not always decidable or cannot be defined due to missing or imprecise information. In those cases the linking of two datasets requires manual assignment of individuals, which could be time consuming and error prone. Thus, manual assignment is doable if only few individuals must be linked or if some tool support with filter options is provided. Finally, linking may require changing or extending BIM data in order to meet requirements for linking data. For instance, if there is a wall that is partially outside it may have to be split into an outside and an inside wall in order to be linked with weather data or occupant behaviour.

Linking BIM data to other data sources typically requires noteworthy engineering knowledge and is already done for instance in the design phase (e.g., linking BIM data with libraries or catalogues to assign default settings – see the HESMOS[47] or ISES[48] projects) or in the monitoring phase (e.g., linking measured values about temperature,

[44] https://code.google.com/p/rdfai/

[45] https://github.com/samuraraujo/SERIMI-RDF-Interlinking/

[46] http://dbpedia.org/

[47] http://www.hesmos.eu/

humidity, energy consumption, air flow rate, etc. to a building in order to combine static and dynamic elements of the building – see [Curry et al. 2013]).

[48] http://ises.eu-project.info/

4 Conclusions

This deliverable has presented a set of guidelines for Linked Data generation in the energy domain, together with two examples. These guidelines help organizations in the energy domain from both public and private sectors in generating Linked Data from already-existing data, by providing detailed descriptions of each task in the generation process. Furthermore, the examples provided within the guidelines help developers from different organizations to gain better insight into the process of Linked Data generation, thus ensuring the highest quality of the outputs of the process.

The requirements for Linked Data generation in the energy domain, presented in this deliverable, helped us in developing the guidelines by pointing out some important information, such as the formats of the data currently available in the energy domain and information on how often the data are updated.

The guidelines presented in this deliverable satisfy those requirements for the Linked Data generation derived from the survey with relevant stakeholders (Chapter 2). The guidelines address each requirement as follows:

R1. The guidelines address the generation of Linked Data from SQL, XLS, or CSV file formats, among others, which are formats that are currently used the most in the energy domain.
R2. The guidelines address the issue of legal aspects, licenses, and data ownership, which is regarded as an important topic that could help lowering the barrier to publish data.
R4. The guidelines cover the generation of static data, as well as dynamic data.
R5. The guidelines describe various means of obtaining and accessing the data, including data stored in files, which is in line with the specified requirements.

However, since we did not gather requirements from a large number of stakeholders, we plan to further validate the guidelines with relevant stakeholders in future events (e.g., workshops or tutorials).

One requirement (R3) is not directly related to the generation of Linked Data. This requirement is related to Linked Data publication, and will be addressed in the future. Also, we plan to further examine all the requirements specified in this deliverable.

The data from the BECA dataset used as an example in the guidelines are not public. Because of this, we plan to generate synthetic data that follow the same template as the data in BECA example, so it can help interested parties in even better understanding the guidelines and its examples and also allows us to publish the materials of the example along with the guidelines.

In order for the data to be available on the Web and to enable their exploitation, the generated Linked Data have to be published according to well-established criteria and best practices. Therefore, future work includes the development of the guidelines for publishing Linked Data on the Web, which will also include details about how to properly publish an ontology developed to describe those data. Together with the guidelines defined in this deliverable, the guidelines for Linked Data publishing will form a comprehensive manual for developers and organizations to follow in order to make their data available as Linked Data, and therefore contributing to the vision of the Semantic Web as a new generation of the World Wide Web.

For each of the tasks in the guidelines that can be supported by some software tool (either manually or automatically), we provide a list of potential tools to be used. This way, the document also provides a catalogue of the different technologies that can be used in the Linked Data generation process.

5 References

[Curry et al. 2013] E. Curry, J. O'Donnell, E. Corry, S. Hasan, M. Keane and S. Riain. *Linking building data in the cloud: Integrating cross-domain building data using linked data.* Advanced Engineering Informatics 27 (2013) 206–219

[Dodds and Davis, 2012] L. Dodds, I Davvis. *Linked Data Patterns*. URL: http://patterns.dataincubator.org/book/. Last retrieved on 26.03.2014.

[Poveda-Villalón, 2012] M. Poveda-Villalón. *A Reus e-based Lightweight Method for Developing Linked Data Ontologies and Vocabularies*. PhD symposium at the 9th Extended Semantic Web Conference (ESWC2012). 27th – 31st May 2012. Heraklion, Greece

[Poveda-Villalón et al. 2012] M. Poveda-Villalón, M. C. Suárez-Figueroa, A. Gómez-Pérez. *Validating Ontologies Using OOPS!*. Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management. Galway City, Ireland. 2012.

[Berners-Lee, 2009] T. Berners-Lee. *Linked Data*. July 2009. URL: http://www.w3.org/DesignIssues/LinkedData.html. Last retrieved on 14.03.2014.

[Nikolov et al. 2011] A. Nikolov, A. Ferrara, F Scharffe. *Data Linking for the Semantic Web*. International Journal on Semantic Web & Information Systems. 7(3) pp. 46-76. 2011.

[Sauermann and Cyganiak, 2008] L. Sauermann, R. Cyganiak. *Cool URIs for the Semantic Web*. December 2008. URL: http://www.w3.org/TR/cooluris. Last retrieved on 17.03.2014.

[Semantic Interoperability Community, 2012] SEMIC - Semantic Interoperability Community, European Commission. *10 Rules for Persistent URIs*. Technical report. 2012.

[Semantic Interoperability Community, 2013] SEMIC - Semantic Interoperability Community, European Commission. *Cookbook for translating Data Models to RDF Schemas*. Technical report. 2013

[Suárez-Figueroa, 2010] M.C. Suárez-Figueroa. Doctoral Thesis. *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse. Spain*. Universidad Politécnica de Madrid. June 2010

[UK Cabinet Office, 2010] UK Government Cabinet Office . *Designing URI Sets for the UK Public Sector*. October 2010

[W3C, 2012] W3C OWL Working Group. *OWL 2 Web Ontology Language*. 2012. URL: http://www.w3.org/TR/owl2-overview/. Last retrieved on 18.03.2014.

ANNEX I - Questionnaire

This annex contains the screenshots of the READY4SmartCities online questionnaire used to extract requirements from relevant stakeholders.

Goal

The goal of this survey is to discover which requirements or restrictions exist for generating energy-related data as Linked Data.

Motivation

The Linked Data paradigm, which is built upon standard Web technologies, allows publishing structured data on the Web and linking such data to other data so they can be more useful.

Nowadays there are plenty of energy-related data coming from different sources; however, the exploitation of such huge amounts of data is hindered by their low interoperability.

In the READY4SmartCities FP7 CSA we are developing a set of guidelines to facilitate data providers generating their energy-related data as Linked Data. To this end, we want to know which requirements and restrictions do current energy-related data providers have in order to support those necessities through those guidelines.

And this is where your individual participation is important; with your feedback we will be able to take your needs into account when elaborating the guidelines.

Survey Process

This survey is oriented to people with technical profile from organizations that provide energy-related data. Even if the initial target population are the stakeholders of the READY4SmartCities project, the survey is open to any interested party.

The survey consists of filling a questionnaire online. The estimated time required to complete the questionnaire is of 5 minutes.

The questionnaire will be available for a month (3 February to 28 March 2014). After that date, the results obtained will be analysed.

The questionnaire does not include any personal question and the confidentiality of the answers will be preserved. We only ask for an email address just in case you want to obtain information about the results of the survey and the guidelines that we will produce.

Contact

This questionnaire is being performed in the context of the READY4SmartCities FP7 project (http://www.ready4smartcities.eu).
If you have any question or comment about the questionnaire contact Dr. Raúl García-Castro at rgarcia@fi.upm.es.

Continue »

Figure 1. The first page of the questionnaire.

Related Domains

Which domains are covered by the data? General:

- ☐ device
- ☐ sensor
- ☐ measurement
- ☐ statistics
- ☐ time
- ☐ geometry
- ☐ location
- ☐ organization

Energy:

- ☐ energy production
- ☐ energy consumption
- ☐ energy distribution
- ☐ energy storage
- ☐ energy market

Building:

- ☐ building
- ☐ building element
- ☐ material
- ☐ building automation
- ☐ electrical system
- ☐ HVAC
- ☐ appliance
- ☐ building furniture

Figure 2. Domains covered by the data (part 1).

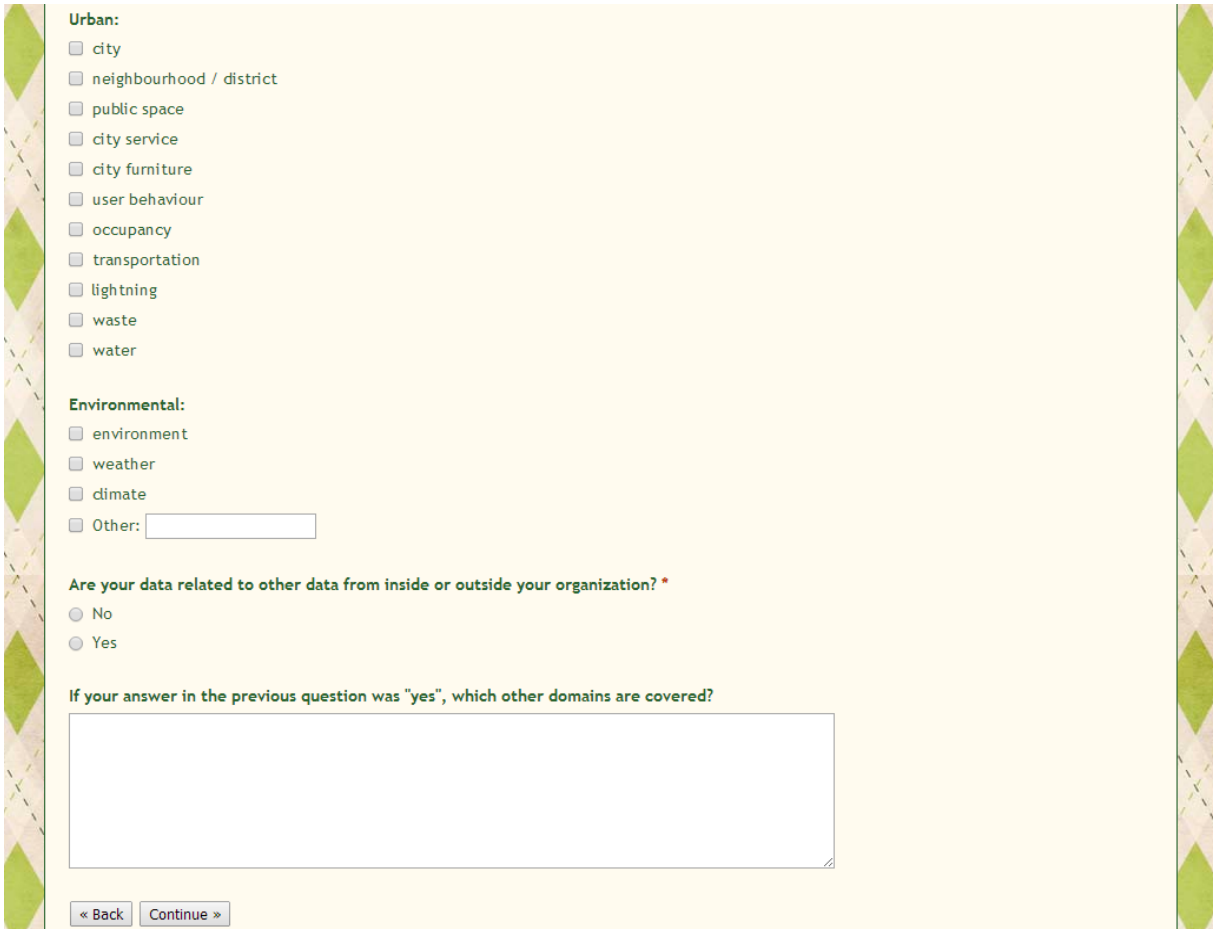


Figure 3. Domains covered by the data (part 2).

Data Characteristics

In what format are the data stored? *

- [ ] Plain Text
- [ ] RTF/DOC
- [ ] PDF
- [ ] SQL
- [ ] XLS
- [ ] CSV
- [ ] XML
- [ ] HTML
- [ ] RDF
- [ ] JSON
- [ ] Other:

In what format is the data model used with the data *

- [ ] Data are unstructured
- [ ] The data model is not formalized
- [ ] SQL
- [ ] XSD
- [ ] RDF-Schema
- [ ] OWL
- [ ] EXPRESS
- [ ] Other:

Are there any metadata that describe your data or parts of it (e.g., creator, modification date)? *

- ( ) No
- ( ) Yes

« Back | Continue »

Figure 4. Data characteristics.

Figure 5. Data generation process.

Figure 6. Data access.

Ethical and legal issues

Is your organization the rightsholder of the data? *

- No
- Do not know
- Yes

Are the terms of use of the data specified? *

- No
- Do not know
- Yes, in terms of use document
- Yes, using an existing license

If you selected "Yes, using an existing licence" in the previous question, provide some information about the existing licence.

Do the data contain personal data? *

- No
- Do not know
- Yes

Do the data contain confidential data (e.g. trade secrets, etc)? *

- No
- Do not know
- Yes

« Back | Continue »

Figure 7. Ethical and legal issues.

Demographic

In which country does your organization reside? *

What sector does your company belong to? *
- Private sector
- Public sector
- Non-profit

Which of the following categories best describes the organization you primarily work in (regardless of your actual position)? *
- Public authority (European, national, regional, local)
- Energy
- Utility
- ESCO
- AEC
- ICT
- Academia
- Other:

Are you familiar with the Linked Data paradigm and its related technologies? *
- No, never heard of them
- Yes, I know something about them
- Yes, Iknow about them and have used them

Would you be interested in publishing your data on the Web and linking them to other data? *
- No
- Maybe
- Yes

If your answer in the previous question was "Maybe", please explain with a few words why.

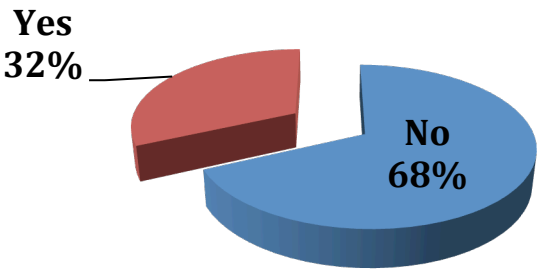
Figure 8. Demographics (part 1).

Figure 9. Demographics (part 2).

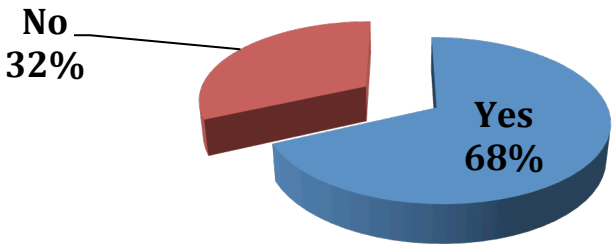
ANNEX II - Survey replies

This annex includes the replies to those questions in the READY4SmartCities survey not included in Chapter 2.

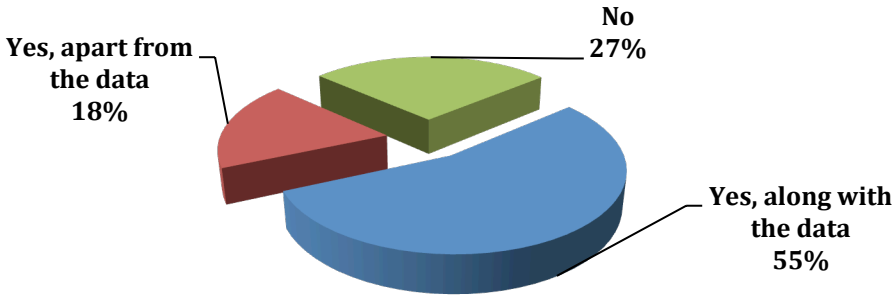
Are your data related to other data from inside or outside your organization?



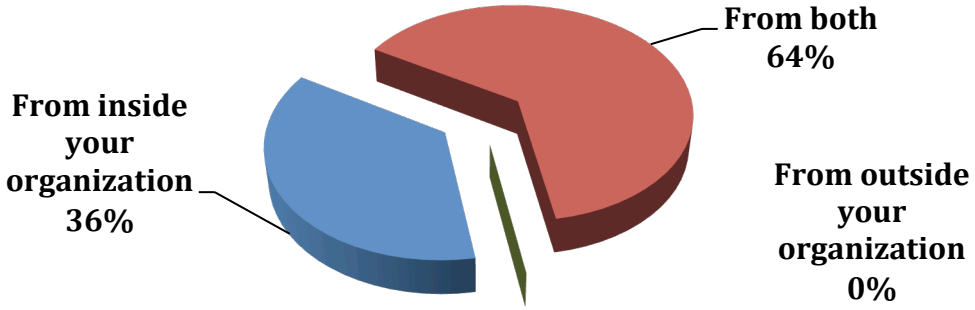
Are there any metadata that describe your data or parts of it (e.g., creator, modification date)?



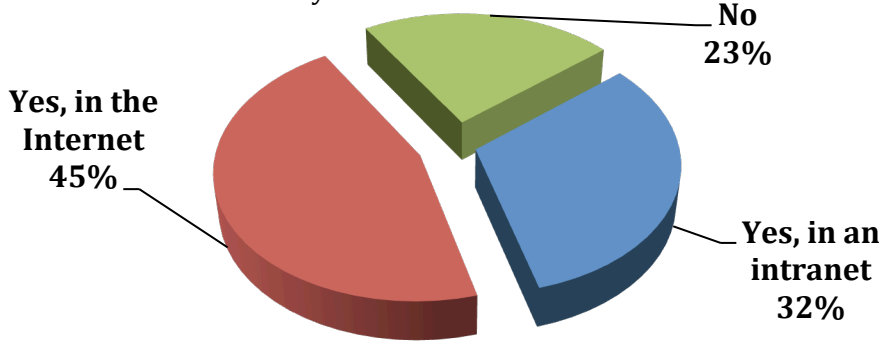
Is information about the provenance of the data stored?



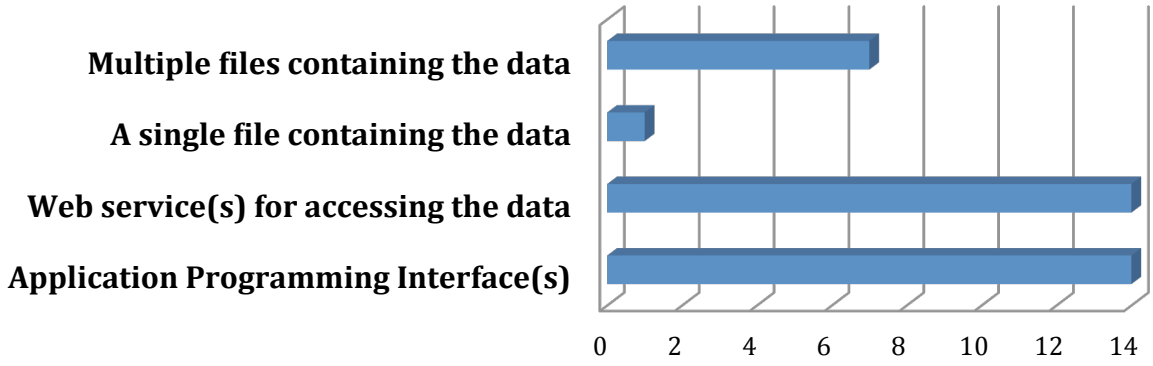
From where are the data currently accessed?



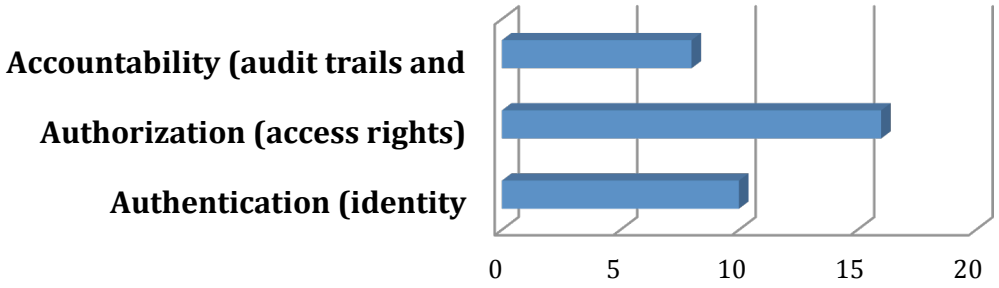
Are your data available online?



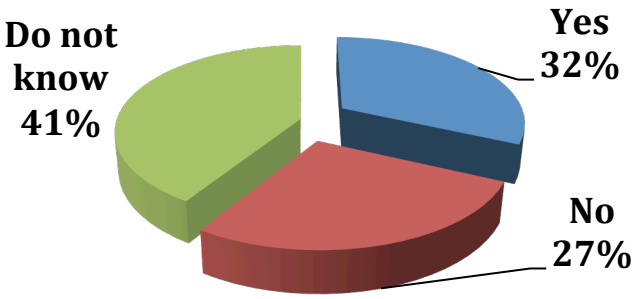
How is access provided to the data?



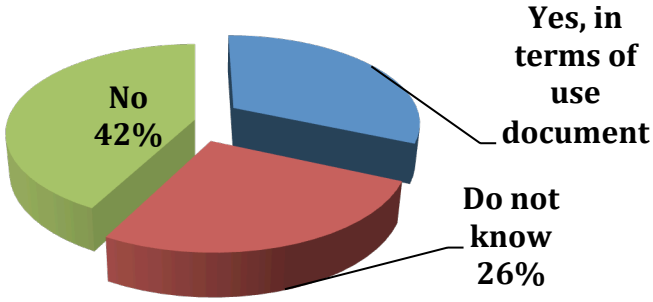
Which of the different services (if any) are provided for controlling the access to your data?



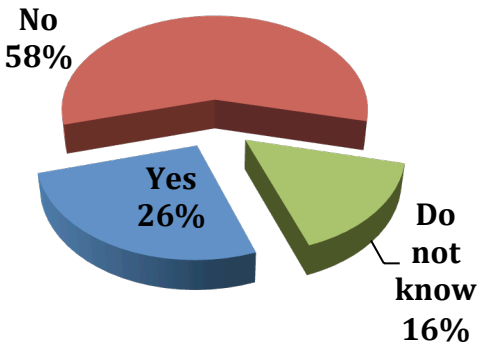
Is your organization the rightsholder of the data?



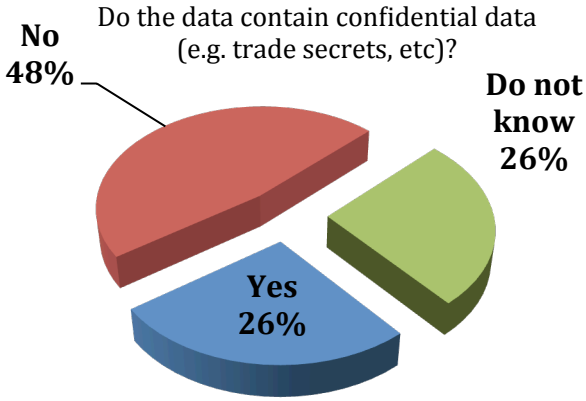
Are the terms of use of the data specified?



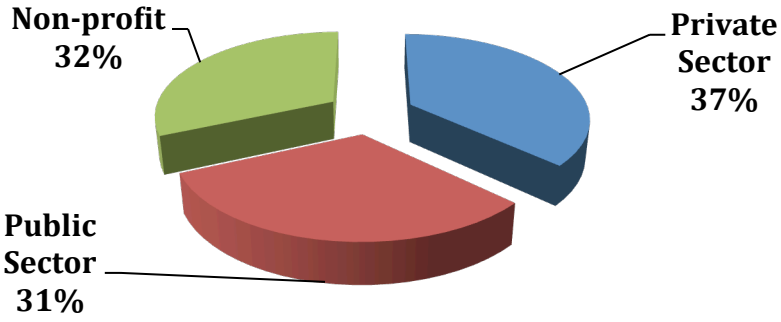
Do the data contain personal data?



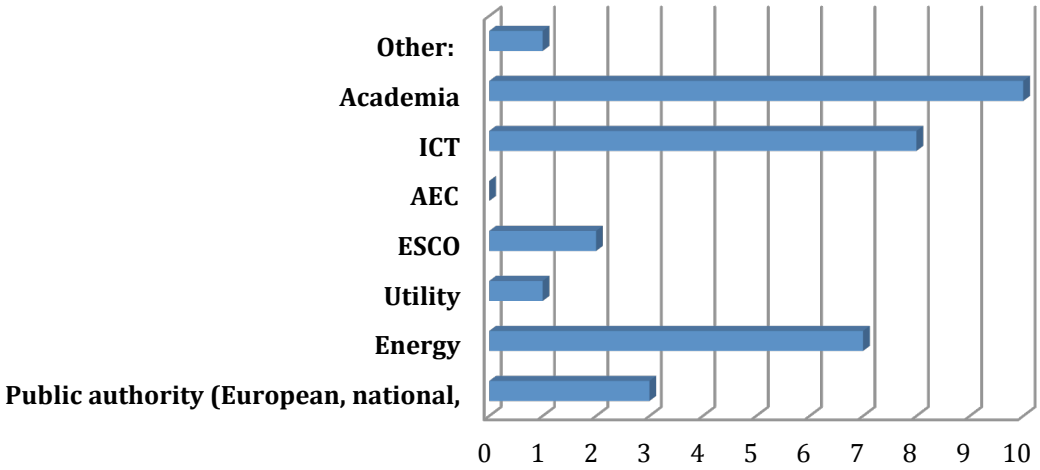
Do the data contain confidential data (e.g. trade secrets, etc)?



What sector does your company belong to?

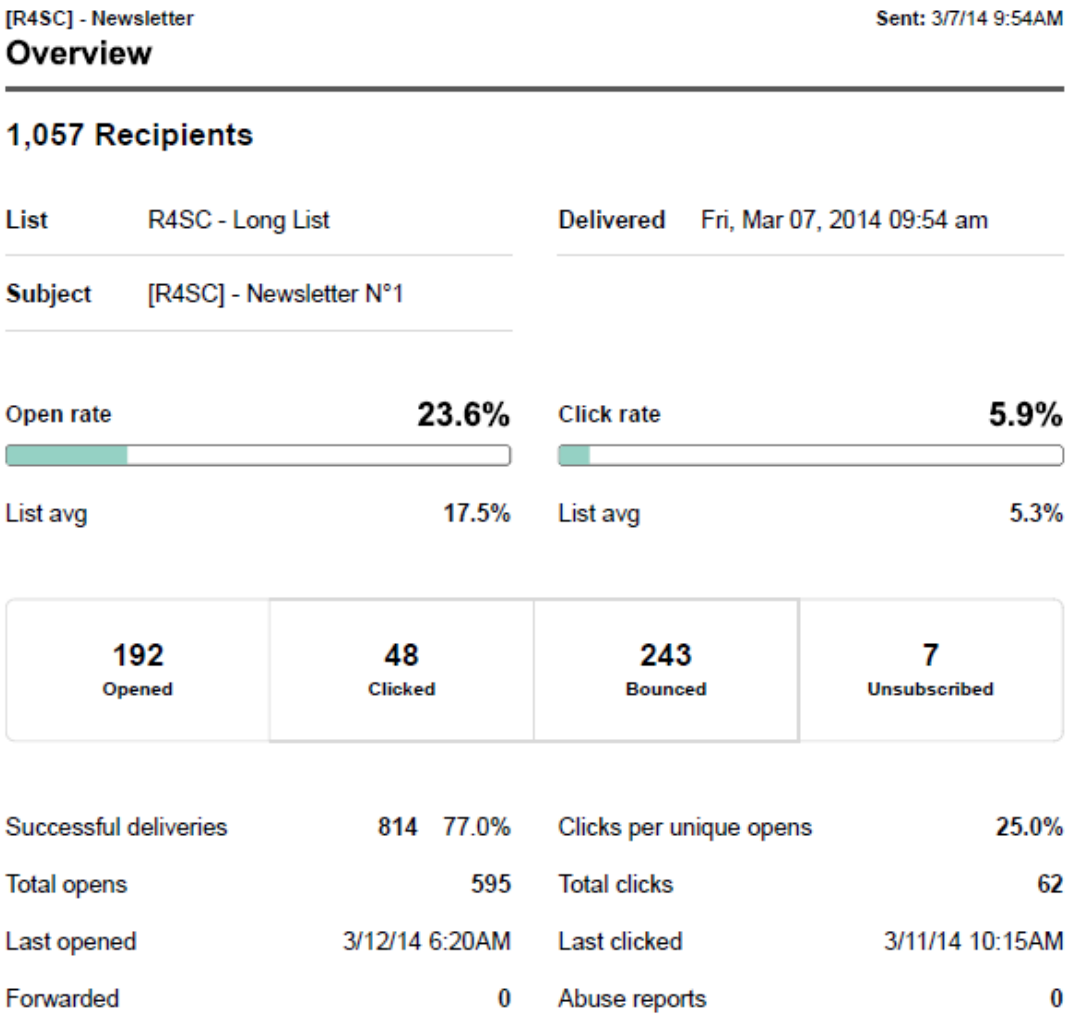


Which of the following categories best describes the organisation you primarily work in (regardless of your actual position)?



ANNEX III - Report for the invitation to the survey

This annex contains the report generated by the MailChimp online mailing list platform, which was used to send the invitation to the survey to more than 1000 recipients.



[R4SC] - Newsletter

Sent: 3/7/14 9:54AM

Overview

1,057 Recipients

List	R4SC - Long List	**Delivered**	Fri, Mar 07, 2014 09:54 am
Subject	[R4SC] - Newsletter N°1		

Open rate	**23.6%**	Click rate	**5.9%**
List avg	17.5%	List avg	5.3%

192 Opened	**48** Clicked	**243** Bounced	**7** Unsubscribed

Successful deliveries	814 77.0%	Clicks per unique opens	25.0%
Total opens	595	Total clicks	62
Last opened	3/12/14 6:20AM	Last clicked	3/11/14 10:15AM
Forwarded	0	Abuse reports	0

Figure 1. Overwiev of the report.

[R4SC] - Newsletter

Sent: 3/7/14 9:54AM

Opens by location

Country	Opens	Percent
Spain	65	13.1%
Switzerland	59	11.9%
Poland	58	11.7%
France	44	8.9%
Germany	43	8.7%
Austria	34	6.9%
USA	30	6.1%
Ireland	24	4.8%
Netherlands	24	4.8%
	19	3.8%

Figure 2. Locations where survey was opened.

[R4SC] - Newsletter

Sent: 3/7/14 9:54AM

Subscriber activity

24-hour performance

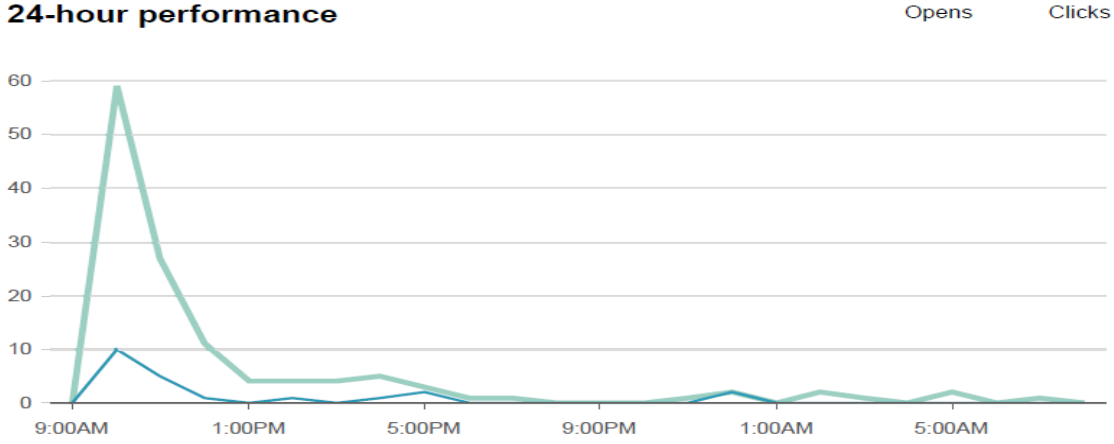


Figure 3. Subscribers activity.