Automation in Construction 57 (2015) 178–187

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Guidelines for Linked Data generation and publication: An example in building energy consumption

Filip Radulovic *, María Poveda-Villalón, Daniel Vila-Suero, Víctor Rodríguez-Doncel, Raúl García-Castro, Asunción Gómez-Pérez

Ontology Engineering Group, ETSI Informáticos, Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form 3 April 2015
Accepted 8 April 2015
Available online 29 April 2015

Keywords:
Linked Data
Guidelines
Buildings
Energy consumption

ABSTRACT

Linked Data is the key paradigm of the Semantic Web, a new generation of the World Wide Web that promises to bring meaning (semantics) to data. A large number of both public and private organizations have published their data following the Linked Data principles, or have done so with data from other organizations. To this extent, since the generation and publication of Linked Data are intensive engineering processes that require high attention in order to achieve high quality, and since experience has shown that existing general guidelines are not always sufficient to be applied to every domain, this paper presents a set of guidelines for generating and publishing Linked Data in the context of energy consumption in buildings (one aspect of Building Information Models). These guidelines offer a comprehensive description of the tasks to perform, including a list of steps, tools that help in achieving the task, various alternatives for performing the task, and best practices and recommendations. Furthermore, this paper presents a complete example on the generation and publication of Linked Data about energy consumption in buildings, following the presented guidelines, in which the energy consumption data of council sites (e.g., buildings and lights) belonging to the Leeds City Council jurisdiction have been generated and published as Linked Data.

1. Introduction

Last years have witnessed the growing interest of many practitioners in publishing semantic data on the Web, mainly powered by the Linked Data[1]1http://www.w3.org/standards/semanticweb/data. initiative, the key paradigm in the next generation of the World Wide Web called the Semantic Web [1]. The concept of Linked Data comes from the idea of using the Web to connect data and aims at transforming the Web into a global knowledge base. The key concepts in Linked Data are links between data from different data sets, which ensure that data sets are not just isolated data islands and support data integration.

By describing the concepts in a domain and the relationships between them, ontologies are 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 contain different components (e.g., classes, properties, instances and axioms), and can be implemented in various languages, being the most widely used and accepted language the one standardized by the W3C, the Web Ontology Language (OWL) [2].

The basic principles for developing Linked Data are the following[2]: i) to provide URIs for each entity to be represented; ii) to provide HTTP URIs for those entities; iii) to use web standards such as RDF [3] for describing data; and iv) to include links to resources already available in the Web. In addition to these principles, in order to realize the notion of Linked Data, not only data must be available in a standard format (i.e., RDF), but also concepts and relationships among data sets must be defined by means of ontologies.

A significant number of energy-related companies posses data about energy consumption, which is one aspect of Building Information Modeling (BIM), that are represented in different formats (e.g., SQL, CSV or XLS), have different update frequencies, and are accessed through different means (e.g., web services or files). Furthermore, having in mind that these data belong to private companies, legal aspects such as licensing are of high importance [4].

The technologies and principles underlying Linked Data are successfully applied in various domains in order to enhance interoperability among systems, and are starting to be applied to the architecture, engineering and construction (AEC) and BIM domains. These opportunities are being analyzed, discussed and promoted along different initiatives

* Corresponding author at: Tel.: +34 913363670; fax: +34 913524819.
E-mail addresses: fradulovic@fi.upm.es (F. Radulovic), mpoveda@fi.upm.es (M. Poveda-Villalón), dvila@fi.upm.es (D. Vila-Suero), vrodriguez@fi.upm.es (V. Rodríguez-Doncel), rgarcia@fi.upm.es (R. García-Castro), asun@fi.upm.es (A. Gómez-Pérez).

[1] http://www.w3.org/standards/semanticweb/data.

[2] Adapted from http://www.w3.org/DesignIssues/LinkedData.html.

http://dx.doi.org/10.1016/j.autcon.2015.04.002
0926-5805/

such as, for example, the LDAC workshop series.[3] More precisely, we can mention the work carried out by Pauwels and Van Deursen [5] to transform BIM data based on the IFC standard [6] into RDF. In this case, the authors reported following no particular methodology and only the development of ad-hoc wrappers is mentioned.

Other approaches focus on consuming and integrating existing Linked Data data sets with AEC data in order to overcome interoperability issues, such as the work described by Curry et al. [7,8] and by O'Donnel et al. [9], or also acting as consumers and publishers of data as the reegle data portal[4] [10] in the energy field.

Linked Data generation and publication are intensive engineering processes that demand high attention in order to achieve high quality and, because of this, some general guidelines and best practices have been developed to this date. However, Villazón-Terazas et al. [11] argue that, although it is possible to have a general guidelines, practitioners should rely on different techniques, technologies and tools for a particular domain. Besides, Villazón-Terazas et al. [12] argue that existing guidelines do not cover all the required steps with enough detail and including the related technologies.

Furthermore, as experience has shown, generic guidelines are often not sufficient to be applied in every domain. In order to overcome this problem, more domain-specific guidelines have been developed with the aim of addressing particular characteristics and of providing concrete domain-specific examples that help practitioners and Linked Data adopters to better understand and use these guidelines. Examples of such domain-oriented guidelines include guidelines for government [12,13] or heritage data [14].

This paper aims at guiding through the process of developing Linked Data related to energy consumption in buildings, including the process of transforming the data available in any format into Linked Data and its publication according to the Linked Data principles. To this end, it provides a methodology for generating and publishing Linked Data with advice on design decisions. In this paper, we describe each task of such methodology in detail through several important aspects, which include a detailed description and the steps to be performed within the task. Furthermore, where possible, we give a list of tools that help in performing the task or some parts of the task, different alternatives to perform the task, or we outline the best practices and recommendations that help in achieving a better quality in the task outputs.

The guidelines presented in this paper are general, in the sense that they can applied to the broad spectrum of different scenarios. However, these guidelines have been developed having in mind characteristics that are specific and useful to the energy consumption in buildings scenario [4,15]. These include data licensing, legal compliance, Smart Cities and Open Data requirements [16], and concrete tools to be used.

Furthermore, the paper also presents an instantiation of the Linked Data generation and publication methodology through the transformation into Linked Data of a data set about building energy consumption. The selected data set comes from the Leeds City Council Open Data office and includes data about electricity, gas and oil consumption from various council sites (e.g., buildings and lights) belonging to the Leeds City Council jurisdiction.

This paper is organized as follows. Section 2 expounds related research efforts. Sections 3 and 4 present the Linked Data generation and publication processes respectively, together with the Leeds City Council example. Finally, Section 5 gives some concluding remarks and discusses lessons learnt and future lines of work.

2. Related work

Different works have explored the advantages and potential of using the Linked Data approach for integrating and enriching AEC data as exposed by Pauwels and colleagues [17], Abanda and colleagues [18], Madrazo and Costa [19] or, more recently, by Törmä [20]. Other related work, such as the work by Törmä and colleagues exposed in [21], already points out specific research problems in this area (e.g., link-type modeling and link generation). In this section we review existing works regarding the generation of Linked Data in the AEC field and also existing general and cross-domain literature about Linked Data generation and publication.

On the Linked Data side, we should mention key publications such as Heath and Bizer's book for driving the Linked Data generation and publication process [22] and consequent works such as the outcomes of the LOD2 project [23]. These represent the starting point for following the process of contributing to the Linked Data initiative; however, as argued in the introduction, existing general guidelines do not provide specific level of details and do not take into account specific characteristics of a particular domain and related tools and techniques. To that extent, some resources might be specialized depending on the field at hand, as has happened in other areas (e.g., biology or cultural heritage) where domain experts together with Linked Data developers have accommodated tools, techniques, and guidelines to their specific requirements.

In general we could state that Linked Data publication in the area of building and construction data is in its infancy. Since experience unveils that practices are too general and not enough to be directly applied to every single domain, it still needs methodological guidelines supporting its evolution towards a mature and repeatable process and providing clear examples in the domain at hand.

3. Linked Data generation process

This section presents the guidelines for the generation of Linked Data for some existing data by describing the different tasks to be performed in the process. Fig. 1 shows an overview of the generation process; 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.

The Linked Data generation process consists of eight tasks. After a data source that will be transformed to Linked Data is selected and access to that data source is obtained, the license has to be analyzed in order to determine the terms of use. Next, the data source is analyzed in detail and a resource naming strategy is specified. Afterwards, an ontology for describing the data is developed and the data is transformed into the RDF format. Finally, generated data are linked to data from other Linked Data data sets.

Next, we describe each task of the Linked Data generation process in detail.

3.1. Select data source

The first step of the Linked Data generation process is the selection of the data source that will be transformed into Linked Data. Such data source is usually owned by the organization and is selected depending on the specific needs of the organization or the expected value to be obtained. Alternatively, an organization may be interested in extending its data with data from other sources not owned by the organization.

This task is achieved by first defining the requirements for the selection of the data source and, then, by selecting one or several data sources that satisfy those requirements.

3.1.1. LCC example

We specified several requirements for the example data: i) to include data about energy consumption in buildings; ii) to have a clear license stated and to be available for use; iii) to be represented in some machine-processable format (e.g., Excel, CSV, XML); iv) to be easily

[3] 2014 edition http://linkedbuildingdata.net/events/ldac2014/.
[4] http://data.reegle.info/.