Computer Standards & Interfaces 45 (2016) 90–99

Contents lists available at ScienceDirect

Computer Standards & Interfaces

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

Linking from Schema.org microdata to the Web of Linked Data: An empirical assessment

Alberto Nogales, Miguel-Angel Sicilia, Salvador Sánchez-Alonso, Elena Garcia-Barriocanal

Information Engineering Research Unit, Computer Science Department, University of Alcalá de Henares, Ctra. Barcelona km. 33.6, 28871 Alcalá de Henares, Spain

ARTICLE INFO

Article history:
Received 8 June 2015
Received in revised form 31 October 2015
Accepted 17 December 2015
Available online 29 December 2015

Keywords:
Ontologies
Microformats
Schema.org
Linked Data
LOV

ABSTRACT

The increase of Linked Open Data (LOD) usage has grown in the last few years, and the number of datasets available is considerably higher. Taking this into account, another way to make data available is microdata, whose aim is to make information more understandable for search engines to give better results. The Schema.org vocabulary was created for the enrichment of microdata as a way to give more accurate results for user searches. As Schema.org is a kind of ontology, it has the potential to become a bridge to the Web of Linked Data. In this paper we analyze the potential of mapping Schema.org and the Web of Linked Data. Concretely, we have obtained mappings between Schema.org terms and the terms provided by the Linked Open Vocabularies (LOV) collection. In order to measure the limitations of our mappings we have compared the results of our script with some matching tools. Finally, an analysis of the usability of interlinking Schema.org to vocabularies in LOV has been carried out. For this purpose, two studies in which we have been presented aggregated information. Results show that new information has been added a substantial number of times.

1. Introduction

On June 2, 2011, Bing, Google, and Yahoo! announced the joint effort Schema.org.[1] Schema.org ontologies are intended for the creation of microcontents targeted to improving indexing and search systems [20]. It consists of a set of tags introduced by HTML5[2] defining a vocabulary that lets webmasters to mark up Web sites with microdata. The purpose of microdata is to help search engines and other tools working with Web sites to better understand the information contained in them. This will eventually help the users to do more precise searches when they are looking for information on the Web. Mika and Potter [15] reported some statistics about the importance of using Schema.org. The increase of microdata used is also shown in Muhleisen and Bizer [16], demonstrating that it has increased among the different formats to embed structured data. The Schema.org initiative has supported the use of microdata, choosing it as its favorite syntax. Taking into account the information given by BuiltWith,[3] which is tool providing technology adoption, ecommerce data and usage analytics for the Internet, the usage of Microdata has increased from 750,000 Websites at the end of 2013 to 1,500,000 nowadays.

There are other initiatives aimed at making data and content more accessible for machine consumption, notably Linked Open Data (LOD). LOD has the objective of publishing open datasets using Resource Description Framework (RDF)[4] format and interlinking these datasets using RDF links. The way these datasets are published follows the well-known Linked Data principles [4]. Sometimes Linked Data is confused with microformats,[5] but the latter is another way to extend the Web completely differently that does not use the principles of Linked Data.

One of the characteristics of both Schema.org and LOD is that of bringing structure and vocabularies to the Web, so it appears to be promising to create links between them. One way to do this is to map the classes and properties from Schema.org with the principal vocabularies used in the Web of Linked Data. In LOD, there are no mandatory vocabularies; the communities using them are the ones that could make a vocabulary more popular at any time. Regardless, we need a way to measure the popularity of a vocabulary used in LOD. The Linked Open Vocabulary (LOV)[6] initiative quantifies the use of classes and properties. LOV consists of various vocabularies used in different fields. The objective of this initiative is to give access to the vocabularies, describe the relations between them and how they are linked with the Linked Data Cloud.

In previous work, Nogales et al. [17], we did a mapping between the classes and properties of Schema.org with LOV. Then using the statistics

E-mail addresses: alberto.nogales@uah.es (A. Nogales), msicilia@uah.es (M.-A. Sicilia), salvador.sanchez@uah.es (S. Sánchez-Alonso), elena.garciab@uah.es (E. Garcia-Barriocanal).

[1] http://www.schema.org
[2] http://www.w3org/TR/html5//
[3] http://builtwith.com/
[4] http://www.w3.org/RDF/
[5] http://www.microformats.org
[6] http://lov.okfn.org

http://dx.doi.org/10.1016/j.csi.2015.12.003
0920-5489/

provided by the project LODStats we measured the impact of Schema.org in the LOD. Finally we exposed a use case to retrieve information from datasets that could be aggregated to Webpages enriching its information.

In this paper, we report an assessment of the potential of linking microcontent and Linked Open Data through the mapping of Schema.org with Linked Open Vocabularies. We have developed a new mapping at a semantic level using synonyms of the Schema.org terms. Results show that only the third part of the vocabularies in LOV provides a class mapping between Schema.org and LOD. With regard to properties, just around the percentage can be mapped using our approach. Furthermore, in LOD it is easier to find particular values for Schema.org properties than for classes. Taking this into account, we can conclude that the reachability in properties is higher than in classes. Once we have demonstrated that there is an important amount of mappings between Schema.org and LOD, we are taking it into account in two studies. The first case will use the value of the classes and properties from Schema.org embedded in Web sites to aggregate new information retrieved from a dataset. The second case involves extending an ontology with properties that are used by a vocabulary from LOV, taking into account the mappings between classes that we have obtained previously. For both cases we have presented some real examples demonstrating its usage. We will get some conclusions about it, giving users a measure of LOD data in pages that are using Schema.org vocabularies. We will also present some example of software that could take advantages of our achievements.

The rest of this paper is structured as follows: in the second section, we present a background of papers using Schema.org and LOV. Then we have a section describing the materials and the methods used to obtain the results. The fourth section shows the results obtained and discusses them. The following section relates the potential use of the mappings. Finally, the last section offers some conclusions about the paper and the implications of future work.

2. Background

In 2011, Schema.org started to provide their official dump of their ontology in OWL[7] format. As a vocabulary, it addresses multiple areas and is not domain specific, but we can differentiate two parts. First, it provides a small set of elements to describe primitive data types like numbers or text in which we can find classes like Boolean, Date or Number. Second, the rest of the classes and properties are used to describe different fields like Organizations or entities related to Medicine, Media, etc. The schema can be extended by users themselves to add new vocabulary by marking up their own data. Nowadays the schema published on the Web can be found in three formats: one is represented in Microdata, the next is an experimental version in RDFa[8] and the third is in OWL, which is not yet fully-up-to-date.

One of the uses of Schema.org is improving the discoverability of data in order to obtain best results when searching. Rosati and Mayernik [21] compare the use of RDF and Schema.org to increase the discoverability and connectivity of resources on the Web to mark up HTML web pages. Researchers cited three cases, concluding that only one of them is more useful to make data more visible in public search engines. This paper does not give statistics about the use of Schema.org that could be use to decide which tags are more useful. Another approach to resource description, search optimisation and resource discovery can be found in Hawksey, Barker and Campbell [7] using Schema.org as embedding metadata. The limitation here is that it only works with open educational resources.

Schema.org has been used in previous research to enrich data. Ambiah and Lukose [1] used Schema.org in a case study to demonstrate the use of a tool that enriches Web sites automatically. The tool presented in the paper is Schema.org Microdata Creator (ScheMicCr), which is being tested in two cases. The first one builds a new Web site with microdata and the second enriches existing Web sites. In this paper the microdata is extracted from a patent knowledge base designed by the authors. In Li et al. [14] an application to publish media fragments and annotations is described using vocabularies defined in Schema.org. In this paper the authors only work with media fragments enriching them so they could be easier to find using search engines. Khalili and Auer [9] introduce the concept of WYSIWYM[9] (What-You-See-Is-What-You-Mean), which consists of manipulating structured content directly. For the implementation it uses a tool called RDFaCE,[10] which is an interface for semantic authoring of textual content, and Schema.org vocabularies to mark-up pages. The annotation in this case is made by the users who create their own subset from Schema.org. Also Schema.org is used as the vocabulary to classify a large collection of Web sites and categorize them ambiguously in Krutil, Kudelka and Snásel [11]. In this paper an algorithm is implemented using the microdata tag 'itemscope' to filter Web sites with recipes and 'itemprop' for getting extra information like rating or author. This paper only uses a few tags of the vocabulary. In Mynarz [27] a tool for validating and previewing structured data tagged with Schema.org in webpages is developed. In this case the tags are already part of the Web. Another approach to enrich Website with Schema.org is presented in Tort and Olivé [28]. This paper shows an approach that consists of using a human-computer task-oriented dialog to design the Web.

We also have some papers in which Schema.org is mapped with other vocabularies. Another paper where Schema.org is used is Atemezing and Troncy [2], where GeOnto[11] ontology is aligned with it, in order to represent and query geospatial data. This paper presents a mapping of Schema.org but only with one of the vocabularies of LOV. A Personalized Location Information System is presented in Viktoratos, Tsadiras and Bassiliades [25]; here a manual mapping is made between Google Places API[12] and Schema.org so the users can fetch extra information from the Web when they retrieve information about a location. In this case the mapping of Schema.org is not made with LOV. Finally, Veres and Elseth [24] present MaDaME, a tool for annotating Web sites with Schema.org, and add semantic metadata. This latter information is added when a concept that the user wants to add is not contained in Schema.org, importing it from WordNet[13] and using SUMO[14] to create a mapping if it is not available. Again a mapping is made between Schema.org and a vocabulary from LOV but only with one of them.

In this paper we analyze the potential of using Schema.org microdata in resources from the Web of Data using the work of Nogales et al. as a foundation [17]. As a link between them, we need to use LOV in order to provide statistics in the use of Schema.org in LOD. LOV provides users a collection of vocabularies from several fields like library science, e-commerce or biology. It also collects information about ontologies that represent vocabularies, detailed information about them, statistics related to LOD or graphical relations between vocabularies. LOV has been reported in several previous researches. Some of these vocabularies have been analyzed by Poveda, Suárez and Gómez [18] to display the reuse of ontologies in Linked Data. This paper gives statistics about how these vocabularies are used and related between them not about

Table 1
Example of a class mapping between Schema.org and LOV.

Class schema.org_iri	Class lov_iri
http://schema.org/Person	http://xmlns.com/foaf/0.1/Person

[7] http://www.w3.org/TR/owl-features/
[8] http://www.w3.org/TR/xhtml-rdfa-primer/
[9] http://en.wikipedia.org/wiki/WYSIWYM
[10] http://rdface.aksw.org/
[11] http://geonto.lri.fr/
[12] https://developers.google.com/places/
[13] http://wordnet.princeton.edu/
[14] http://www.ontologyportal.org/