# Schema.org Usage for Hotels

## An Analysis Based on the Web Data Commons Data Set

Elias Kärle[*]
elias.kaerle@sti2.at

Anna Fensel[*]
anna.fensel@sti2.at

Ioan Toma[*†]
ioan.toma@sti2.at

Dieter Fensel[*]
dieter.fensel@sti2.at

[*]Semantic Technology Institute (STI) Innsbruck
University of Innsbruck
Technikerstrasse 21a
6020 Innsbruck, Austria

[†]UMIT - University for Health Sciences
Medical Informatics and Technology
Eduard-Wallnöfer-Zentrum 1
6060 Hall in Tyrol, Austria

## ABSTRACT

It has been almost four years now since the world's leading search engine operators (Bing, Google, Yahoo! and Yandex), decided to start working on an initiative to enrich web pages with structured data; an initiative known as schema.org. Since then, many web masters and those responsible for designing web pages started adapting this technology to enrich websites with semantic information. This paper analyzes parts of the structured data in the largest web crawl available and open to the public, the *Common Crawl*, in order to find out how the tourism branch is using schema.org. On the use case of hotels, it studies the usage and distribution of *schema.org/Hotel*, examines who uses schema.org, how it is applied and whether or not the classes and properties of the vocabulary are used in a syntactically and semantically correct way.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms

Analysis

## Keywords

schema.org, semantic annotation, analysis, hotel, tourism

## 1. INTRODUCTION

Particularly in the tourism branch, the web has evolved to be the most important tool for representing businesses and distributing information about offers, events and other facts to potential customers. How search engines rank the popularity of certain pages changes frequently over time, and is probably the best kept secret of search engine providers, and hence a big challenge for web masters and search engine optimization experts. This makes it even more important to stick to certain recommendations or standards concerning content markup on web pages and to follow initiatives launched by search engine operators, such as *schema.org.*

On June 2nd 2011 the worlds biggest search engines, Google, Bing and Yahoo!, decided to "create and support a common set of schemas for structured data markup on web pages." [1], called *schema.org.* On November 1st of the same year, the operator of the largest Russian search engine, Yandex, joined the initiative and together they are constantly working on the refinement and the further development of this set of vocabulary. After these companies announced that the usage of *schema.org* will lead to significantly better search results and search engines presence and rankings, numerous websites started annotating their content with the vocabulary provided by *schema.org.*

The *Common Crawl*[2] is an organization which crawls the web several times a year and provides the collected archives and data sets to the public for free. *Web Data Commons*[3] is a project started in 2012 by *Freie Universität Berlin* and the *Karlsruhe Institute of Technology*, and it extracts different types of structured data from the *Common Crawl* and also provides them to the public for free.

The interest of this paper lies upon a data set within *Web Data Commons*, containing Microdata, RDFa and Microformat, used to annotate web page content with schema.org [3]. In this paper we present our work on getting a comprehensive overview of the distribution of tourism specific *schema.org* vocabulary over the web, using the example of the type schema.org/Hotel.

This paper is structured as follows: Section 2 describes related work, section 3 states the research questions and explains the methodology used to analyze the data. Section 4 presents the findings of the research, and section 5 concludes the paper.

[1] http://googlewebmastercentral.blogspot.co.at/2011/06/introducing-schemaorg-search-engines.html
[2] http://commoncrawl.org/
[3] http://webdatacommons.org/

## 2. RELATED WORK

During our work on this project, we came across work which is related to our research. First of all, in the paper by Stavrakantonakis et al. (2013), the authors survey the use of Web 2.0 technologies, the use of content management systems and social channels and Web 3.0 technologies, as well as the use of semantic web technologies and structured data on the websites of 2155 hotels in Austria. The outcome of this research is that only 5% of the websites employ semantic technologies and the vast majority of hotels ”completely ignore the existence of technologies that could enrich the website content with high level metadata and give machine readable meaning to the presented information” [4].

During the analysis, we came across several cases of wrong usage of schema.org. To detect, analyze and solve those problems the work of Meusel et al. (2015)[2] serves as a starting point for our further work, when we wish to give advice towards the semantically and syntactically correct usage of schema.org annotations.

When it comes to finding and choosing the most suitable vocabulary, a project worth mentioning is vocab.cc. It is an open source project which allows users to search for linked data vocabularies, based on the dataset of the Billion Triple Challange[4].

The available schema.org annotations have a commercial exploitation potential, which is currently pursued by several institutions. For example, current STI Innsbruck's start-up effort ONLIM[5] is applying annotations on online social media technologies in its product, social media marketing tool. The start-up already runs pilots, such as with touristic associations of Innsbruck[6], and implements semantic dissemination support by implementing schema.org support on their website and publishing the touristic data of the regions as linked open data.

Another direction towards widespread real life application of schema.org is in the development of tools assisting web developers to easily and correctly introduce schema.org annotations. One example here is the WYSIWYM project described in Khalili et al. (2013) [1].

## 3. RESEARCH QUESTION AND METHODOLOGY

As a starting point for our analysis, we define key research questions we want to answer. These are:

1. **How many hotels use schema.org?** This question triggers an analysis on whether or not it is possible to indicate a number of hotels that are annotated with schema.org, either on their own website or on third party websites?
2. **Is schema.org used syntactically and semantically correctly or are there many mistakes?** The answer to this question surveys the mistakes made when it comes to annotating hotels.
3. **Who is using schema.org in the touristic field?** The last question is looking for answers w.r.t. whether or not hotels use schema.org on their own web sites and which other platforms annotate hotels with schema.org.

As mentioned in the introduction, the primary source of our data was the result of the *Common Crawl.* Since our analysis should be based only on structured data and, to be more precise, on schema.org, we took advantage of a project called *Web Data Commons*. This project uses the data from *Common Crawl* and extracts all sorts of structured data which then are divided into three main data sets. The *Hyperlink Graph*, the *Web Tables*, and the RDFa, Microdata and Microformat dataset, upon which our interest lies. From this dataset we are using the ”Schema.org Class Specific Data-Subsets” and from those subsets the one containing all triples related to schema.org/Hotel.

The schema.org/Hotel specific subset of the 2013 crawl was 2.2GB in compressed and 35GB in uncompressed size.

Over all, we used 37 different queries. The measurement and analysis of the collected data, which was present in CSV tables, was mostly done by hand or by arithmetic functions in Microsoft Excel, as well as through generating charts and diagram.

## 4. RESULTS

In the following Section we will present the results of our analysis of the schema.org/Hotel related structured data on the 2013 corpus of the *Web Data Commons* project. In Section 2 of the paper, we have defined three main questions which will be answered below.

### 4.1 How many hotels use schema.org?

When trying to find out how many hotels are present in the triple store, one can first query for all triples with predicate *rdf:type* and object *schema.org/Hotel* and count them. The output would be about 4.841.000 hotels in the whole data set. But after a little manual inspection it is clearly visible that many hotels are annotated more than once because, for example, they have schema.org annotations on their own website and are annotated in listings of one or several booking platforms. Trying to do the same query with the restriction of only counting hotels with unique names results in a reduced number, about 740.000, which is also not expressive, because details about the hotels with same names, like for example *Hotel Post* or *Hotel Adler* - which are very common hotel names in Austria, are still not distinct.

A solution to that problem would have been to perform a search on unique hotel names and locations or addresses, but we observe that less than 75% of the hotels in the dataset have proper annotation for an address. To be more specific, only about 3 million hotels added used schema.org/Address, 2.2 million Hotels used schema.org/street, 2 million hotels used schema.org/zip, 1.9 million hotels used schema.org/land and schema.org/Region and only 1.1 million hotels used schema.org/name as a country name. See Table 1 for more details.

If we count all appearances of annotations of hotels per coun-

[4] http://km.aifb.kit.edu/projects/btc-2012/

[5] http://www.onlim.com

[6] http://www.innsbruck.info

**Table 1: Classes and properties used in the data set**

| Class or Property | Usage sum | Percentage |
|---|---|---|
| schema.org/Hotel | 4.841.000 | 100% |
| schema.org/PostalAddress | 3.035.000 | 62,7% |
| schema.org/addressCountry | 1.904.000 | 39,3% |
| schema.org/Country/name | 1.125.000 | 23,2% |
| schema.org/addressRegion | 1.902.000 | 39,3% |
| schema.org/postalCode | 2.011.000 | 41,5% |
| schema.org/streetAddress | 2.284.000 | 47,2% |
| schema.org/Rating | 2.377.000 | 49,10% |
| schema.org/ratingValue | 2.375.000 | 49,06% |

**Table 2: Distribution of hotel triples per country**

| Rank | Country | Sum | Percentage |
|---|---|---|---|
| 1 | US | 1.021.513 | 90,8% |
| 2 | CA | 52.360 | 4,7% |
| 3 | CN | 20.648 | 1,8% |
| 4 | GB | 11.580 | 1,0% |
| 5 | DE | 3.163 | 0,28% |
| 6 | MX | 1.921 | 0,17% |
| 7 | PR | 1.250 | 0,1% |
| 8 | AR | 1016 | 0,09% |
| 9 | PH | 765 | 0,07% |
| 10 | IN | 699 | 0,06% |
|  | other | 10.085 | 0,9% |

**Table 3: Distribution of ratings among annotated hotels**

| Rating | Usage sum | Percentage |
|---|---|---|
| 5 | 866.932 | 36,5% |
| 4.5 | 35.079 | 1,5% |
| 4 | 651.606 | 27,4% |
| 3.5 | 66.208 | 2,8% |
| 3 | 426.925 | 18% |
| 2.5 | 15.476 | 0,6% |
| 2 | 176.800 | 7,4% |
| 1.5 | 941 | 0,03% |
| 1 | 135.958 | 5,7% |

**Table 4: Usage of properties in hotel triples. (For space reasons http://schema.org/ is shortened to sc:)**

| Property name | Usage sum | Percentage |
|---|---|---|
| sc:Hotel/name | 5.666.474 | 117% |
| sc:Hotel/review | 5.226.132 | 108% |
| rdf:type | 4.841.353 | 100% |
| sc:Hotel/image | 3.439.579 | 71,0% |
| sc:Hotel/address | 3.035.301 | 62,7% |
| sc:Hotel/aggregateRating | 2.723.587 | 56,3% |
| sc:Hotel/rating | 2.377.406 | 49,1% |
| sc:Hotel/description | 1.934.486 | 40% |
| sc:Hotel/url | 1.749.830 | 36,1% |
| sc:Hotel/geo | 1.323.333 | 27,3% |

try, of course, only for those 23.2 % of hotels which have an annotation for schema.org/postalAddress and schema.org/name within postalAddress, we come to the conclusion that the large majority of triples is found within the United States, followed by Canada, China, Great Britain, Germany and others. See a detailed listing in Table 2.

Another interesting aspect of this data set was to find out which categories of hotels are either using schema.org on their own, or are annotated by others. For this purpose we inspected the appearance of schema.org/Rating, which aims, due to the documentation[7], to show the rating on a numerical scale from one to five, as it is done in hotels with the stars rating (*, ..., *****). In our understanding, values with .5, like in 3.5 stars, indicate a higher level hotel, such as for example the ***Superior rating. But again, the observation is that only about 2.3 million hotel triples even make use of the schema.org/Rating class (see also Table 1 for details). Analysis of the mentioned 2.3 million triples showed a clear tendency of higher rated hotels to be annotated more accurately and more frequently, see Table 3 for datails.

## 4.2 How is schema.org used in the hotel domain?

This question will be answered by taking a detailed look at which classes are used when it comes to annotating hotels and which attributes are in use.

To find out which classes and properties are used and how often they appear, we iterated over all hotel triples and all related properties. We grouped those properties by name and counted the appearance. With this method we found 37.192.502 triples, directly related to hotel triples. The most frequently used property was schema.org/Hotel/name, 5.666.474 times, which is interesting, because there are only about 4.8 million hotel triples. Obviously some hotels were annotated with two or more names. The second most frequently used property was schema.org/Hotel/review, 5.226.132 times, which is not very surprising, because as we will see in Section 4.3, a large number of hotels are annotated with schema.org on rating websites. Place three in this ranking is rdf:type, with 4.841.353 appearances, which is the attribute that tells a triple that it is a hotel - this number of course equals the number of total hotel triples. Overall there are 119 different properties in use which either refer to literals or to classes. To find more details about the top ten used properties, see Table 4.

In the documentation for schema.org/Hotel there are 62 properties mentioned from either the Hotel class itself or inherited from *LocalBusiness*, *Organization*, *Thing* and *Place*, while our analysis came up with 111 different properties. This again is an indicator that large inaccuracies take place when it comes to annotations. Attributes are written syntactically wrong, for example *makeOffer* instead of *makesOffer* and some properties even get invented out of thin air, like *Hotel/wedding*, *Hotel/telefax* or *Hotel/?description?*. Even tough almost all properties of schema.org are generally in use, only 8 properties appear in more than 30% of hotel triples and only 20 of the 62 described properties are used in more than 1% of the hotel triples.

[7] https://schema.org/Rating

To sum up this question, there is a movement observable towards semantic annotation of hotels but there still a lot to be done to match a sufficient annotation.

## 4.3 Who is using schema.org in the hotel domain?

With this question we wanted to find out if it is the individual hotel that uses schema.org most or best to describe its properties, or if it is a third party page which displays and annotates hotels for whatever reason (e.g. for providing the hotel information in order to collect the hotel bookings). After manually browsing through some of the hotels in the data set during the process of the analysis, it appeared, by looking at the mentioned fourth column of the NQuad (the data provenance column), that only a very small number of hotels showed their own url as a provenance. The vast majority of the hotels appeared to be annotated by third party websites. So we came up with a hypothesis which says: "In the tourism domain, schema.org is predominantly used by booking- or rating webs sites, barely by hotel web sites themselves".

The approach we took to prove the derived hypothesis was the following: iterating over all hotel triples found on booking- and rating websites which offer a hotel-URL (as hotel-URL schema.org/Hotel/url is used) and checking if the hotel web site is schema.org annotated. Further, we use the hotels pay-level domain as a unique identifier and note if a schema.org annotation was found on the hotel web site or not. And finally, check if the specific hotel appears multiple times in the data set, and if so, note on which other web sites and count the appearance. With this method we get a detailed overview of how many hotels use schema.org themselves and which other websites, rating- or booking sites use schema.org to annotate hotels.

# 5. CONCLUSIONS AND FUTURE WORK

To conclude this paper we would like to highlight that schema.org is used in the touristic domain. Hotels start annotating web sites for more visibility in search engines and to power rich snippets. Also third party web sites such as rating- or booking platforms are using schema.org more often-sometimes even excessively- to increase search engine visibility as well as to make their data more visible and useful for other developments, like the usage in mobile apps. Nevertheless, especially for the hotels' own web sites, there is much more that could and should be done when it comes to annotation. Very often schema.org classes and properties are used incorrectly. Some properties are invented by the website developers and often, very important classes and properties- such as the URL, telephone number, description or geographic location- are totally omitted. It appears that the hotel owners' only concern is to be visible and highly ranked in the web search engines, but they completely ignore what could be created from of their hotel's data if properly annotated by third party apps such as event platforms or other services- or information orientated web sites.

We also wish to highlight that since May 2015 (when schema.org version 2[8] was released), a newly introduced schema.org extensions mechanism has been enabling extessions for various domains. One other idea for future work we are currently addressing, is to create an extension (similar to that of schema.org) for tourism. As we discovered in the hotel domain (and this is true for other touristic fields as well), a lot of important information can not yet be annotated by schema.org: for example, number of beds per hotel room, availabilty of a TV or a whirlpool, etc. Extending schema.org with terminology for describing hotels, hotel rooms, amenities and in general any other aspect of accommodations and their features could really enrich schema.org and make it even more valuable for tourism.

As mentioned in Section 4.3, we would also like to survey the whole data set to find out who is using schema.org most and to get specific numbers about the distribution of schema.org on hotels' own websites. As a very interesting part of our future work, we would like to compare all the findings we described in this paper with the newly published 2014 data set of the Web Data Commons project, and perhaps even newer data sets as soon as they are published.

## Acknowledgments

The authors would like to thank the Online Communications working group[9] for their active discussions and input during the OC meetings, Christian Bizer and Rober Meusel from the University of Mannheim for their input on how to work with the WDC data set, Ontotext[10] for a free license of the GraphDB repository software, used as our triple store and third-party funded projects in which STI is involved such as *Byte*, *FITMAN*, *LDCT*, *TourPack*, *OntoHealth*, *OpenFridge* and *EuTravel* for their support.

# 6. REFERENCES

[1] A. Khalili and S. Auer. Wysiwym authoring of structured content based on schema. org. In *Web Information Systems Engineering–WISE 2013*, pages 425–438. Springer, 2013.

[2] R. Meusel and H. Paulheim. Heuristics for fixing common errors in deployed schema. org microdata. In *The Semantic Web. Latest Advances and New Domains*, pages 152–168. Springer, 2015.

[3] R. Meusel, P. Petrovski, and C. Bizer. The webdatacommons microdata, rdfa and microformat dataset series. In *The Semantic Web–ISWC 2014*, pages 277–292. Springer, 2014.

[4] I. Stavrakantonakis, I. Toma, A. Fensel, and D. Fensel. Hotel websites, web 2.0, web 3.0 and online direct marketing: The case of austria. In *Information and Communication Technologies in Tourism 2014*, pages 665–677. Springer, 2013.

[8] http://schema.org/version/2.0/

[9] http://oc.sti2.at/

[10] http://ontotext.com/products/ontotext-graphdb/