Schema.org: How is it used?

Minh-Hoang Dang[1], Alban Gaignard[2], Hala Skaf-Molli[1] and Pascal Molli[1]

[1]Nantes Université, LS2N, Nantes, France
[2]Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France

Abstract

Schema.org defines a shared vocabulary for semantically annotating web pages. Due to the vast and diverse nature of the contributed annotations, it is not easy to understand the widespread use of Schema.org. In this poster, we rely on the characteristic sets computed from the web data commons datasets to provide insights into property combinations on various websites. Thanks to in-depth experiments, this poster establishes a comprehensive observatory for schema.org annotations, visually presenting the most frequently used classes, commonly used combinations of properties per class, the average number of filled properties per class, and the classes with the greatest property coverage. These findings are valuable for both the communities involved in defining Schema.org vocabularies and the users of these vocabularies.

1. Introduction

Schema.org[1] defines a standardized, general-purpose vocabulary for semantically annotating web pages [1]. It covers entities such as people, places, events, products, and relationships between entities. A total of 803 types and 1,464 properties are defined by June 2023. Since 2011, major search engines (Google, Bing, Yahoo, Yandex) have encouraged webmasters to use Schema.org for annotating their web pages [2]. The Web Data Commons [3, 4] project extracts semantic annotations from the Common Crawl annually since 2010[2]. It provides a reference dataset to study the evolution and adoption of semantic annotations in web pages. The extracted data is represented with RDF quads, which consist of RDF statements along with the URL of the corresponding web page. The abundance of annotations on the web and the diversity of contributors raise challenges in understanding how Schema.org is used at the web-scale.

Previous studies have provided statistics regarding the adoption of Schema.org [5, 2], i.e., the evolution of the standard, trends in adopting Schema.org annotations, the number of websites implementing it, and the growth of popular classes. Nevertheless, none of the existing studies provide us the means to ascertain how webmasters use and combine properties for specific types (classes), e.g., an e-commerce webmaster may use the *schema:Product* class and the combination of *name* and *description* properties to annotate a product, while others use *name* and *review*.

ISWC'23 Posters and Demos: 22nd International Semantic Web Conference, November 06–10, 2023, Athens, Greece
minh-hoang.dang@etu.univ-nantes.fr (M. Dang); alban.gaignard@univ-nantes.fr (A. Gaignard); hala.skaf@univ-nantes.fr (H. Skaf-Molli); pascal.molli@univ-nantes.fr (P. Molli)
0000-0003-1062-6659 (H. Skaf-Molli); 0000-0001-8048-273X (P. Molli)

[1]https://schema.org
[2]http://webdatacommons.org/structureddata/

(a) Cakes entities at four different websites

Entity	name	image	description	totalTime
Cake1	X	X		X
Cake2			X	X
Cake3	X	X		
Cake4	X	X		X

(b) Number of entities of characteristic sets

Combinaison of properties	Count
name + image + totalTime	2
description + totalTime	1
name + image	1

Figure 1: Example of characteristic sets of the class *Recipe*

By relying on the *Product* class specification defined by Schema.org, it is impossible to know which combination of properties is the most used on the web. As a webmaster, am I following good practices? Finding the most used combinations of properties can help discover latent soft schemas [6] in semantic annotations on the web.

This poster establishes an observatory for schema.org annotations, providing comprehensive insights into the commonly used combinations of class-specific properties and the quality of class descriptions. This information is crucial for both communities specifying Schema.org vocabularies and Schema.org profiles (*e.g.* Bioschema.org) and for the users of these specifications. The remainder of the paper is organized as follows: Section 2 presents our approach. Section 3 details our experimental study. The last section concludes the paper and points out perspective works.

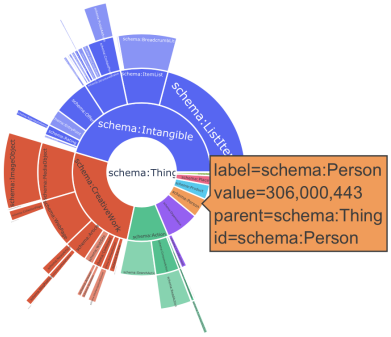
2. Approach

We rely on characteristic sets [6] to build our observatory. Characteristic sets describe semantically similar entities by grouping them according to the set of properties the entities share. For an entity s of in an RDF dataset R, the *characteristic set* is defined as: $S_C(s) = \{p | \exists o : (s, p, o) \in R\}$. To illustrate, in Table 1a, each row represents a different website, and the X indicates the presence of a property in that particular website. Table 1b presents the cardinality of each characteristic set. We consider a class well described if many entities of that class share a combination of properties with many properties. We computed characteristic sets (CSets) for the JSON-LD dataset (most used format) of WebDataCommons (October 2021) [3]. The CSets are available at (https://doi.org/10.5281/zenodo.8167689) and used as a basis to answer different questions detailed in the next section. All results are available at (https://schema-obs-demo.onrender.com).

3. Experimental study

The experimental study answers the following questions: i) Which classes are the most commonly used? ii) What are the common combinations of properties per class? iii) Which classes are accurately described?

[3] http://webdatacommons.org/structureddata/2021-12/stats/how_to_get_the_data.html (623GB)

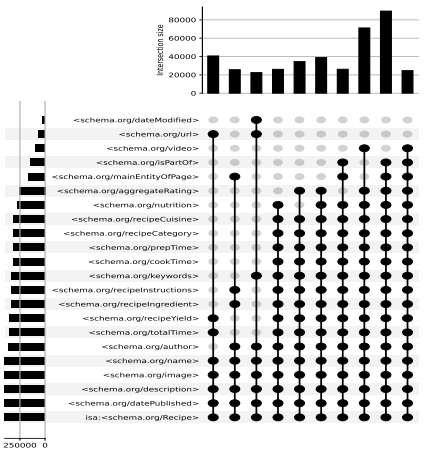


(a) Hierarchy of used classes

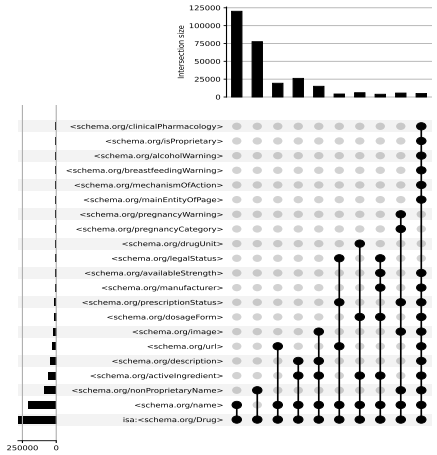
Rank	Type	Count
1	schema:ListItem	1.39 B
2	schema:ImageObject	591 M
3	schema:BreadcrumbList	460 M
4	schema:Organization	435 M
5	schema:WebPage	405 M
6	schema:SearchAction	372 M
7	schema:Offer	351 M
8	schema:Person	306 M
9	schema:ReadAction	245 M
10	schema:Product	219 M

(b) Top-10 most used Schema.org classes

Figure 2: Schema.org Classes and instances per class.



(a) Top-10 properties combination of *Recipe*



(b) Top-10 properties combination of *Drug*

Figure 3: Upset plots displaying the top-10 Schema.org characteristic sets.

Data corpus. We used the JSON-LD (most common formats) dataset from the WebDataCommons [3] released in October 2021. This dataset is derived from crawling 35 million websites, of which 42% utilized Web Entities. It comprises 82 billion RDF quads (16 terabytes uncompressed) and 6.7 billion Schema.org entities.

Distributed computing infrastructure. To analyze the schema.org dataset composed of 6.7B web entities, we used an 8 nodes HPC cluster (8 CPU threads, 32 GB of RAM, 20 GB of local storage per node). The code is written in Apache Spark. We computed in total $4,638,824$ CSets, which took around 30 hours.

Most used classes. The sunburst diagram of Figure 2a presents the class hierarchy and the corresponding instance count per class. Table 2b shows the most common classes, with

Rank	Class	Coverage
1	BorrowAction	0.99
2	DepartementStore	0.84
3	PlanAction	0.81
4	SportActivityLocation	0.79
5	Event	0.74
6	Product	0.73
7	LiveBlogPosting	0.72
8	Recipe	0.72
9	PostalAddress	0.72
10	SaleEvent	0.69

(a) Top-10 classes ranked by coverage

Rank	Class	AvP
1	ReviewNewsArticle	15.34
2	ReportageNewsArticle	14.21
3	Recipe	14.08
4	Car	13.96
5	AnalyseNewsArticle	13.22
6	AdvertiserContentArticle	13.15
7	SocialEvent	12.90
8	LearningResource	12.82
9	VideoGallery	12.79
10	TechArticle	12.49

(b) Classes ranked by average properties

Figure 4: Ranking of Schema.org types through average property and coverage metrics.

schema:Person ranking 8th with around 306 million entities. It also reveals that the classes are not uniformly employed, indicating varying degrees of usage across the web.

Most commonly used combination of properties per class. Figures 3a and 3b illustrate, respectively, the Upset plot for the top-10 characteristic sets of the classes *Recipe* and *Drug*. The 10 columns represent the top-10 combination of properties, ordered by the number of combined properties. In the case of *Drugs*, the vast majority of instances (top left histograms) are annotated with only *name* and *nonProprietaryName* properties. The last column of this plot shows very few instances annotated with the largest combination of properties. Conversely, in the case of *Recipes*, starting from column 8, we observe that a large number of instances are better annotated. More generally, the characteristic sets and their visual representation through an upset plot provide interesting insights into the class's latent soft schema and the poorly used properties.

Classes coverage. To compare class descriptions, we use the coverage metric defined in [7]. A high coverage (near 1) indicates that class entities use most properties defined in the Schema.org type specification. For readability, we computed the coverage for the properties contained in the top-10 CSets per class. As shown in Table 4a, *Recipe* has a high coverage of 0.72, whereas we computed a low coverage of 0.14 for *Drug* (104 defined properties), which confirms the results shown in Figures 3a and 3b.

Classes average properties (AvP). We computed the average number of used properties (AvP) per class as indicated in Table 4b. On schema.org specification, a class definition has an AvP of 70.47, but on class instances, we observe an AvP of 5. This means that most of the properties defined in the schema.org are not used when instantiating the classes. We observed that *Recipe* obtained a good ranking with an AvP of 14.08 on 144 defined properties, whereas *Product* has a lower AvP of 7.3 on 68 defined properties. This indicates that the cooking community may better populate the available *Recipe* properties than the e-commerce community with the class *Product*.

More generally, we observed (i) no correlation between the rate of type usage and its coverage, (ii) classes with the best coverage are those with fewer properties, (iii) the rankings with AvP and coverage metrics return different top-10. The coverage metric seems to be biased towards classes specified with few properties.

4. Conclusion and future works

We analyzed how webmasters effectively use Schema.org types and properties to annotate web pages extracted from the WebDataCommons dataset. For each of the 776 analyzed types, we computed the number of instances, the characteristic sets, the average number of properties per type, and their coverage.

Thanks to the characteristic sets, we could graphically display the instantiated schema with Upset plots. Compared to the Schema.org specifications, we observed that very few properties are effectively instantiated, and there is a great diversity in the combination of used properties. The Upset plots allow webmasters and Schema.org maintainers to know which properties are effectively and commonly used.

In future works, we aim to define new quality metrics and leverage this Schema.org observatory to study: (i) the per-class and per-web domain use of properties such as *sameAs*, (ii) the temporal evolution of Schema.org by analyzing the yearly published WebDataCommons datasets and (iii) the adoption of emerging community-specific profiles (e.g., Bioschemas) promoted in the context of FAIR, reproducible, and open sciences.

Acknowledgments

This work is supported by the French ANR project DeKaloG (Decentralized Knowledge Graphs), ANR-19-CE23-0014, CE23 - Intelligence artificielle, and the French CominLabs project MikroLog (The Microdata Knowledge Graph) grant no. 2019-05655. We are especially grateful to the Nantes University master students Houda Bourefis, Julien Brochard, Farah Farouh, Hajar Lazrak Senhaji, Mohammed Ali Ahmed Ragaa, Mohammed-Amine Bouzid, Yacine-Fadl Bestaoui, and Oswaldo Andres Jimenez Hidalgo, who contributed to the early stages of this study.

References

[1] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: Evolution of structured data on the web, Queue 13 (2015) 10 – 37.

[2] A. Brinkmann, A. Primpeli, C. Bizer, The web data commons schema.org data set series, in: Companion Proceedings of the ACM Web Conference, 2023, pp. 136–139.

[3] H. Mühleisen, C. Bizer, Web data commons - extracting structured data from two large web corpora, in: LDOW, 2012.

[4] R. Meusel, P. Petrovski, C. Bizer, The webdatacommons microdata, rdfa and microformat dataset series, in: The Semantic Web – ISWC 2014, Springer International Publishing, Cham, 2014, pp. 277–292.

[5] R. Meusel, C. Bizer, H. Paulheim, A web-scale study of the adoption and evolution of the schema.org vocabulary over time, in: 5th International Conference on Web Intelligence, Mining, and Semantics, 2015.

[6] T. Neumann, G. Moerkotte, Characteristic sets: Accurate cardinality estimation for rdf queries with multiple joins, 27th International Conference on Data Engineering (2011).

[7] S. Duan, A. Kementsietsidis, K. Srinivas, O. Udrea, Apples and oranges: a comparison of RDF benchmarks and real RDF datasets, in: SIGMOD, 2011.