

#### Schema.org: How is it used? Minh-Hoang Dang, Alban Gaignard, Hala Skaf-Molli, Pascal Molli

#### ▶ To cite this version:

Minh-Hoang Dang, Alban Gaignard, Hala Skaf-Molli, Pascal Molli. Schema.org: How is it used?. International Semantic Web Conference (ISWC) 2023, Nov 2023, Athens, France. CEUR Workshop Proceedings, 2023. hal-04250523v1

#### HAL Id: hal-04250523 https://hal.science/hal-04250523v1

Submitted on 17 Nov 2023 (v1), last revised 29 Apr 2024 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Schema.org: How is it used?

NantesUniversité

Minh-Hoang Dang<sup>1</sup> , Alban Gaignard<sup>2</sup> , Hala Skaf-Molli<sup>1</sup> and Pascal Molli<sup>1</sup>



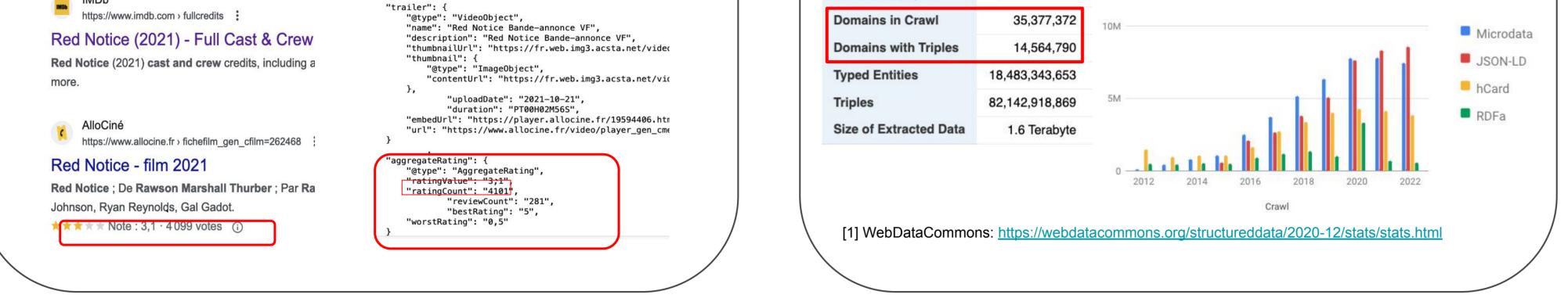
<sup>1</sup> Nantes Université, LS2N, Nantes, France <sup>2</sup> Nantes Université, CNRS, INSERM, L'Institut du Thorax, F-44000 Nantes, France



l'institut

## What is Schema.org?

- Schema.org provides a standardized vocabulary for annotating web pages, describing entities and their relationships.
- Major search engines like Google, Bing, Yahoo, and Yandex have been encouraging webmasters to use Schema.org since 2011.



Schema.org is largely used

- By October 2021, 42% of websites were annotated [1].
- More and more webpages employ JSON-LD as a markup format [1].

Crawl Date	October 2021			
Total Data	85.11 Terabyte	(compressed)		
Parsed HTML URLs	3,195,003,256			
URLs with Triples	1,516,194,663	Number of PLD:	s Deploying the four Maj	or Markup Formats
Domains in Crawl	35,377,372	10M		
Domains with Triples	14,564,790			
Typed Entities	18,483,343,653			

## How is it used?

- Schema.org defines vocabulary, but we don't know **how people actually use** it:
  - The class *Person* is defined, but *how many instances* are there?
  - There are 60 properties for the class
     Person, but are they all used ?

 Are web pages dedicated to Recipe better annotated than those dedicated to Drug?

## We do not know how people use Schema.org !

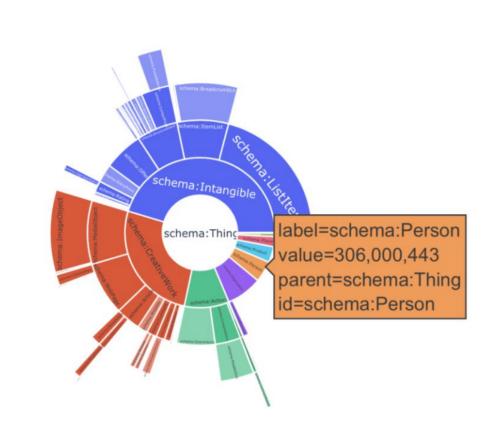
#### **Objectives**

- Get insights about the commonly used combinations of class-specific properties.
- Evaluate the quality of class descriptions.

## Methodology

Characteristic sets describe semantically similar entities by grouping them according to the set of shared properties.
We consider a class well-described if many entities in that class share a large combination of properties.
We computed characteristic sets (CSets) for the WebDataCommons JSON-LD dataset (October 2021)
6.7B web entities
4, 638, 824 CSets

### **Experimental study**



#### (a) Hierarchy of used classes



Rank	Туре	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

75000 -

# What are the most frequently used classes?

- The top 3 most common classes are automatically generated by SEO tools.
- Classes are not used uniformly e.g. there are twice as many
   Organizations as Products.

# What are the typical combinations of properties for each class?

(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) N	umber of entities of char	acteristi	С
	Combinaison of properties	Count	

description + totalTime

name + image

name + image + totalTime 2

sets

#### Conclusion

- Only a small number of properties have actually been instantiated.
- The combination of properties differs considerably from class to class.

<schema.org datemodified=""></schema.org>		•					•					-	-	-	-	-	-	-	
<schema.org url=""></schema.org>	• •	•			•		•				<schema.org clinicalpharmacology=""></schema.org>								P
<schema.org video=""></schema.org>					٠		•		-		<schema.org isproprietary=""></schema.org>			-		-	-		2
<schema.org ispartof=""></schema.org>				•			•				<schema.org alcoholwarning=""></schema.org>				•	•	•		p
<schema.org mainentityofpage=""></schema.org>	• •			•		• •	•				<schema.org breastfeedingwarning=""></schema.org>		•	•	•	•	•	P 1	<u>P. 1</u>
<pre><schema.org aggregaterating=""></schema.org></pre>	0 0	0 0	•	• •	•	• •	•				<schema.org mechanismofaction=""></schema.org>	•		•	•	•	•	0 9	p
<schema.org nutrition=""></schema.org>	• •	• •		• •	•	• •	•				<schema.org mainentityofpage=""></schema.org>		•	•	•	•	•	0 9	D. 19
<schema.org recipecuisine=""></schema.org>	0 0	•	•	• •	•	+ +	•			1	<schema.org pregnancywarning=""></schema.org>	•	•	•	•	•	•	D (	)     •
<pre><schema.org recipecategory=""></schema.org></pre>	• •	• •	•	• •	•	• •	•			1	<schema.org pregnancycategory=""></schema.org>	•	•	۲	•	۲	•	D (	Þ. 🖣
<pre><schema.org preptime=""></schema.org></pre>	• •		•	• •	•	+ +	•			1	<schema.org drugunit=""></schema.org>	• •	•	۲	•	•	•	P (	þ (4
<schema.org cooktime=""></schema.org>	• •		•	• •	•	• •	•			1	<schema.org legalstatus=""></schema.org>	• •	۲	۲	•	۲	•	∲.●	<b>)</b> 4
<schema.org keywords=""></schema.org>	• •	+ +	•	• •	•	+ +	•			1	<schema.org availablestrength=""></schema.org>		•	۲	•	۲	0.0	•	
<schema.org recipeinstructions=""></schema.org>	• •		•	• •	•	• •	•			- 1	<schema.org manufacturer<="" td=""><td></td><td></td><td>•</td><td>•</td><td>•</td><td>0.1</td><td>≬.∢</td><td><b>)</b> (</td></schema.org>			•	•	•	0.1	≬.∢	<b>)</b> (
<schema.org recipeingredient=""></schema.org>	• •	•	•	• •	•	+ +	•			1	<schema.org prescriptionstatus=""></schema.org>			۲	•	•	•	0 (	
<schema.org recipeyield=""></schema.org>	••		•	• •	•	• •	•			1	<schema.org dosageform=""></schema.org>				•	•		∢ ♦	
<schema.org totaltime=""></schema.org>	••	•	•	• •	•	• •	•			- L	<schema.org image=""></schema.org>		•	•	•	•	0	0.0	
<schema.org author=""></schema.org>	• •	• •	•	• •	•	• •	•			- E	<schema.org url=""></schema.org>			•	•	•	•	0.1	6. 3
<pre><schema.org name=""></schema.org></pre>	♦ ♦	+ +	•	• •	•	+ +	•				<schema.org description=""></schema.org>		•		•	•	0	0 0	6
<schema.org image=""></schema.org>	<b>♦</b> ♦	+ +	•	• •	•	+ +	•				<schema.org activeingredient=""></schema.org>				•	•	•	• •	<b>)</b> d
<schema.org description=""></schema.org>	<b>+ +</b>	+ +	•	• •	•	+ +	•			-	<schema.org nonproprietaryname=""></schema.org>		•		•	•	0	0.0	
<schema.org datepublished=""></schema.org>	<b>+</b> +	• •	•	• •	•	+ +	•				<schema.org name=""></schema.org>			•	•	•	•	• •	
isa: <schema.org recipe=""></schema.org>	• •	• •	•	• •	•	• •	•				isa: <schema.org drug=""></schema.org>		•	•	•	•	•		
												-	-	-	-	-	-		
50000 0									25000	0 00									

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

- 776 available upset plots for each Schema.org class
- Very few instances of Drug use a large combination of properties, whereas a majority of Drugs solely annotate their name.

# Which classes have precise descriptions?

- The greater the use of properties, the greater the **coverage** of this class.
- The coverage metric favors classes with fewer properties.
- Average number of used properties (AvP):
  - AvP(Recipe) = 14.08 on 144
     defined properties,
  - **AvP(Product)** = 7.3 on 68 defined properties.
- This reveals that some communities provide better quality semantic annotations than others.

#### **Future works**

• Upset plots enable webmasters and

Schema.org maintainers to know which properties are commonly used.

- Continue exploring Schema.org data with the latest version of the WebDataCommon datasets.
- Observe how Schema.org adoption of has evolved over the years.



#### **Read the poster** https://hal.science/hal-04250523

This work is funded by the French CominLabs project MikroloG (The Microdata Knowledge Graph) : <u>https://project.inria.fr/mikrolog/</u>



# Checkout the demonstration

https://schema-obs-demo. onrender.com/

