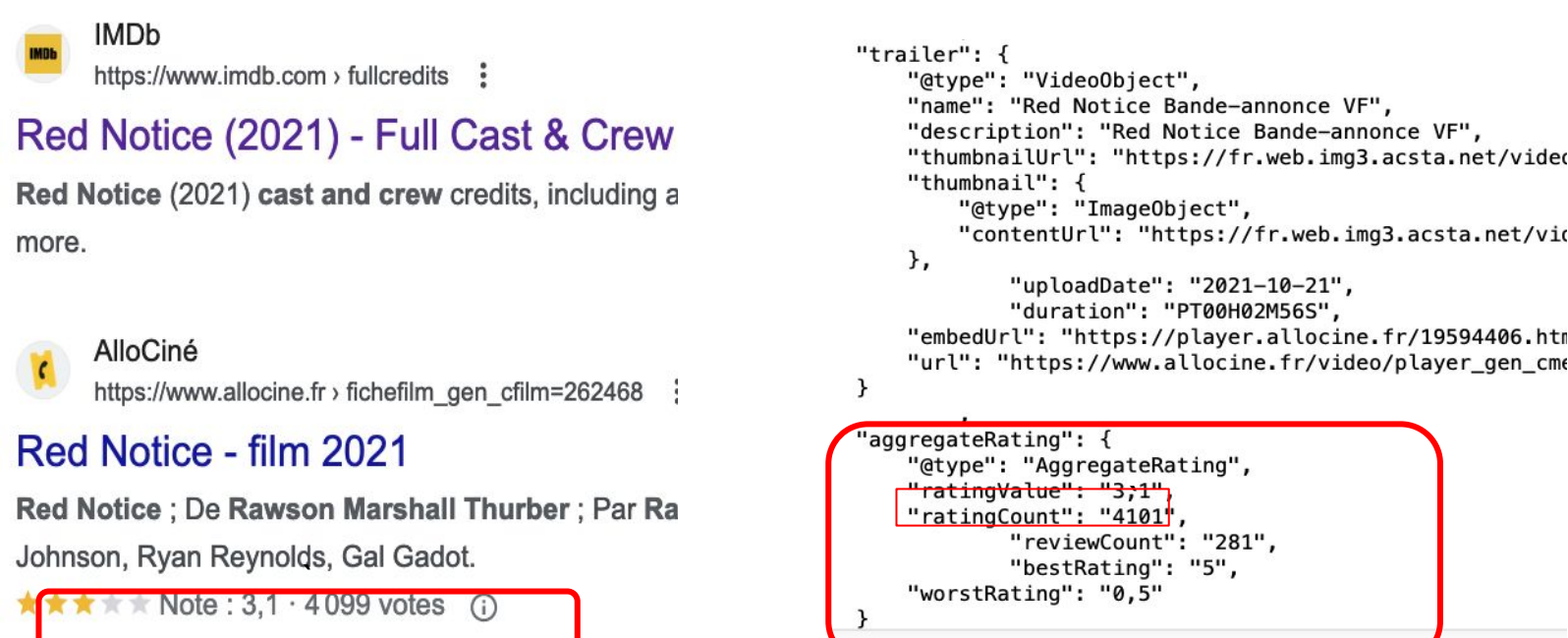


# Schema.org: How is it used?

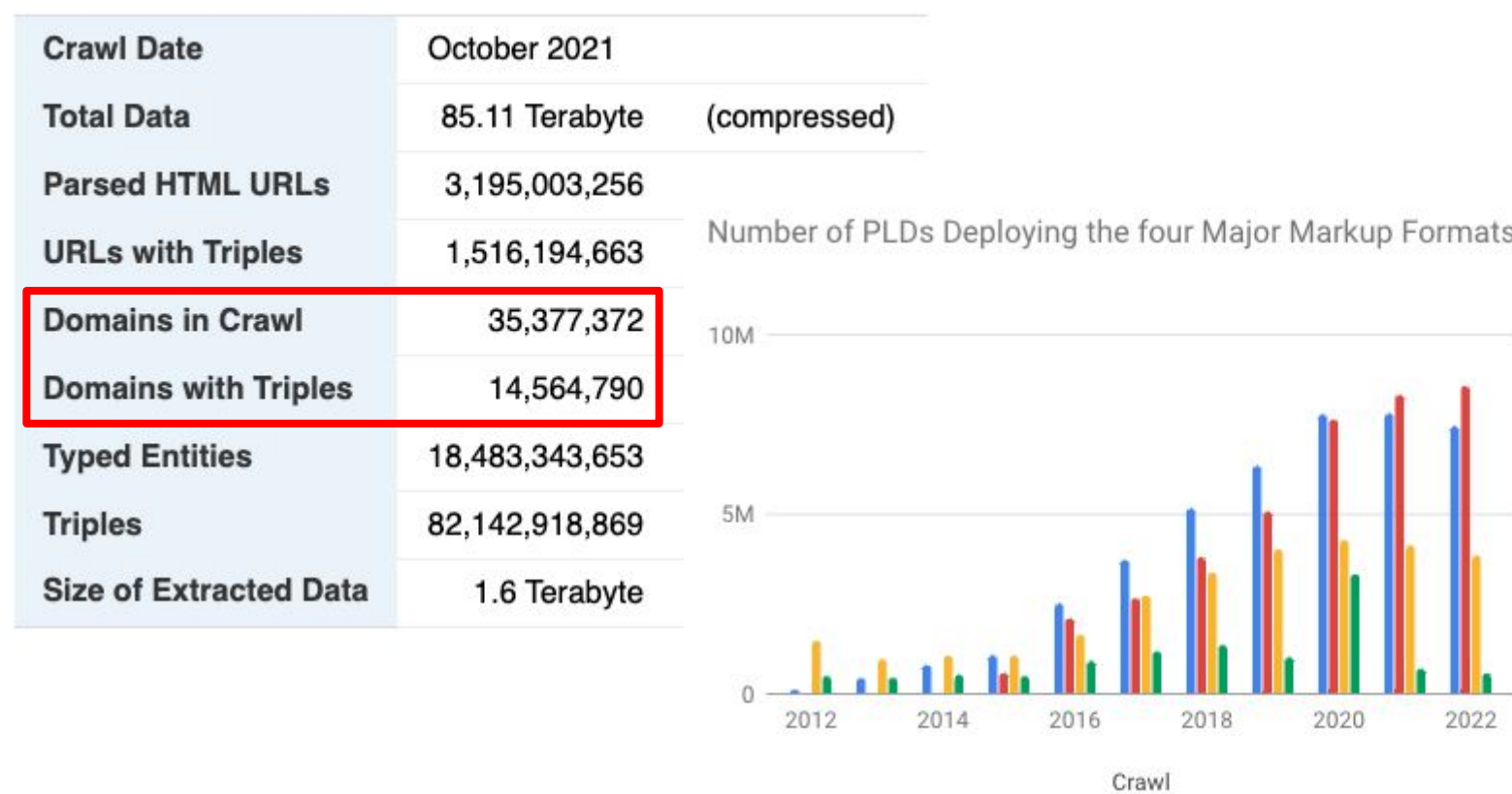
## What is Schema.org?

- 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].



[1] WebDataCommons: <https://webdatacommons.org/structureddata/2020-12/stats/stats.html>

## 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

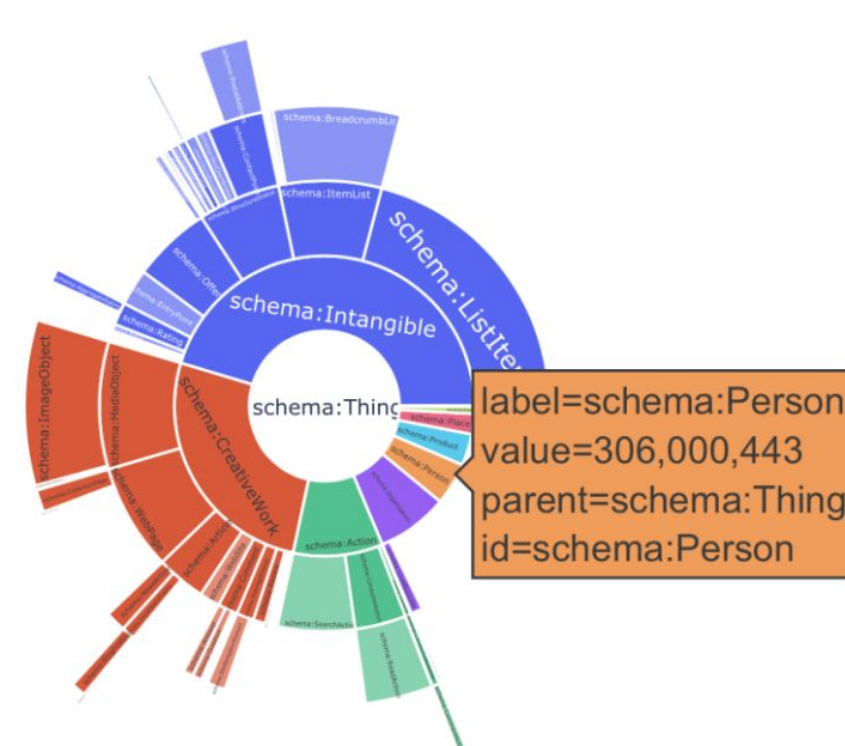
(a) Cakes entities at four different websites

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

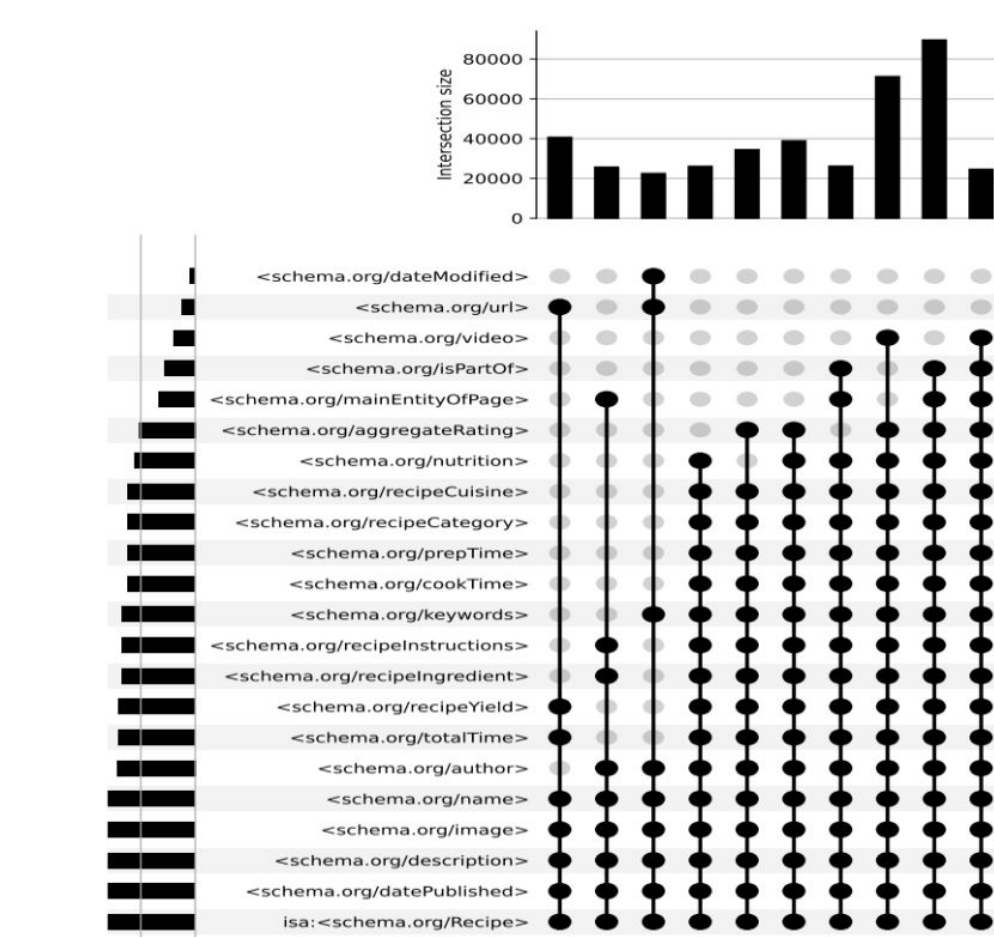
### Experimental study



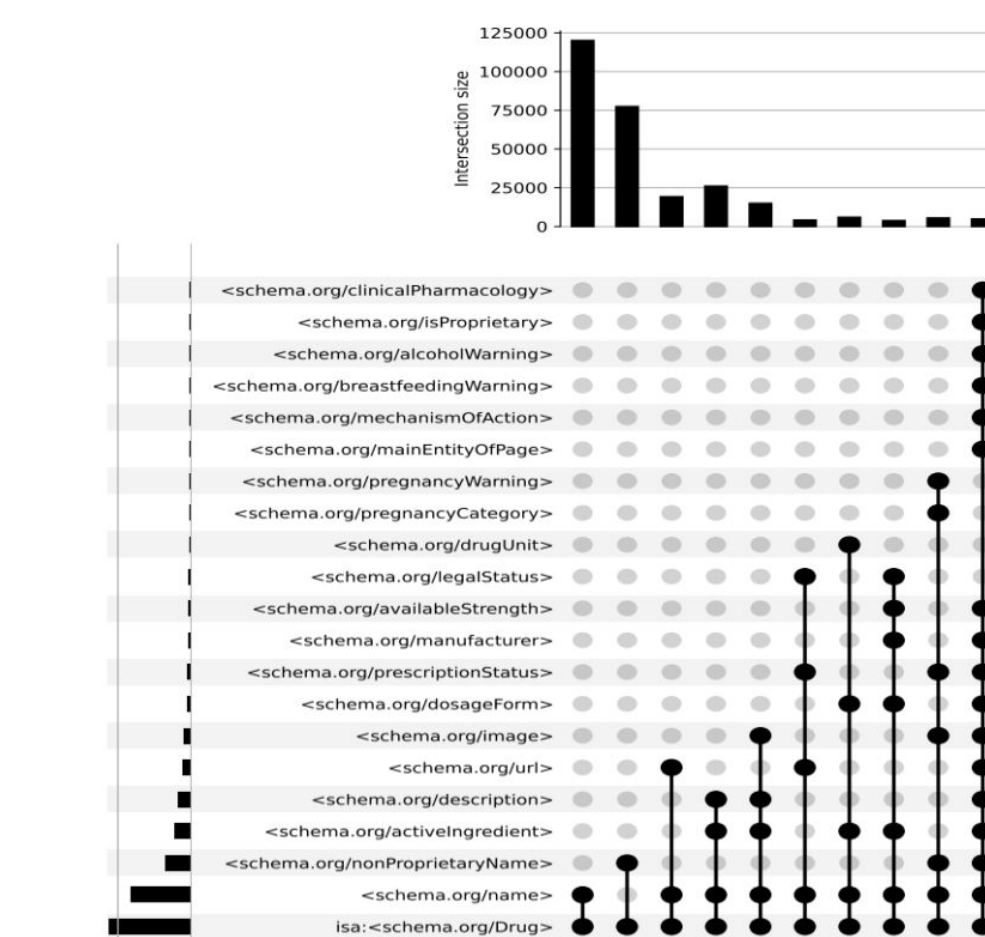
(a) Hierarchy of used classes

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



(a) Top-10 properties combination of Recipe



(b) Top-10 properties combination of Drug

Rank	Class	Coverage
1	BorrowAction	0.99
2	DepartmentStore	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

### 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?

- 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?

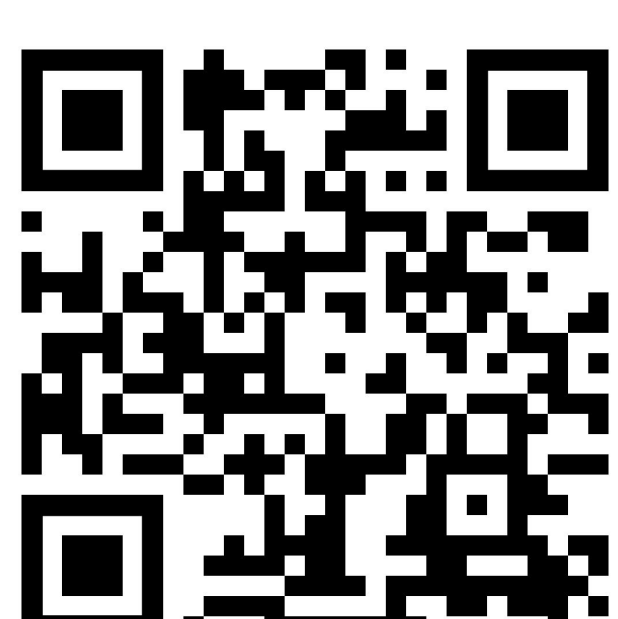
- The greater the use of properties, the greater the **coverage** of this class.
- The coverage metric favors classes with fewer properties.
- Average number of used properties (AvP):
  - AvP(Recipe)** = 14.08 on 144 defined properties,
  - AvP(Product)** = 7.3 on 68 defined properties.
- This reveals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

## Conclusion

- Only a small number of properties have actually been instantiated.
- The combination of properties differs considerably from class to class.
- Upset plots enable webmasters and Schema.org maintainers to know which properties are commonly used.

## Future works

- 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) : <https://project.inria.fr/mikrolog/>



ISWC  
2023  
NOVEMBER 6-10  
ATHENS-GREECE

Checkout the  
demonstration

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

