# The Web Data Commons Schema.org Data Set Series

Alexander Brinkmann
alexander.brinkmann@uni-mannheim.de
University of Mannheim
Mannheim, Germany

Anna Primpeli
anna.primpeli@gmail.com
University of Mannheim
Mannheim, Germany

Christian Bizer
christian.bizer@uni-mannheim.de
University of Mannheim
Mannheim, Germany

## ABSTRACT

Millions of websites have started to annotate structured data within their HTML pages using the schema.org vocabulary. Popular entity types annotated with schema.org terms are products, local businesses, events, and job postings. The Web Data Commons project has been extracting schema.org data from the Common Crawl every year since 2013 and offers the extracted data for public download in the form of the schema.org data set series. The latest release in the series consists of 106 billion RDF quads describing 3.1 billion entities. The entity descriptions originate from 12.8 million different websites. From a Web Science perspective, the data set series lays the foundation for analyzing the adoption process of schema.org annotations on the Web over the past decade. From a machine learning perspective, the annotations provide a large pool of training data for tasks such as product matching, product or job categorization, information extraction, or question answering. This poster gives an overview of the content of the Web Data Commons schema.org data set series. It highlights trends in the adoption of schema.org annotations on the Web and discusses how the annotations are being used as training data for machine learning applications.

## CCS CONCEPTS

• **Information systems → Web data description languages**; **Data extraction and integration**.

## KEYWORDS

information extraction, semantic annotations, schema.org, web science

**ACM Reference Format:**
Alexander Brinkmann, Anna Primpeli, and Christian Bizer. 2023. The Web Data Commons Schema.org Data Set Series. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587331

## 1 INTRODUCTION

The schema.org vocabulary[1] defines terms for describing entities such as persons, products, events, organizations, reviews, job offers, questions and answers, and a total of 790 other types of entities [4]. The schema.org vocabulary is maintained using a community process [5]. Schema.org terms are used together with the Microdata syntax to annotate data elements in the BODY of HTML pages. Alternatively, the terms are used with JSON-LD syntax to embed structured data in the HEAD of HTML pages. Since 2011, the search engines Google[2], Bing[3], Yahoo, and Yandex[4] have been asking webmasters to use schema.org terms to annotate structured data within their pages. The search engines use the annotated data to display rich snippets in search results, info boxes next to search results, and entities on maps. Other applications that use scehma.org data include Google Shopping, Google for Jobs, and Google Dataset Search[5]. These applications have motivated many webmasters to add schema.org annotations to their pages.

As starting point for extracting large amounts of schema.org data, a comprehensive web corpus is required. The CommonCrawl[6] is the largest publicly available web corpus. The corpus is released monthly and typically contains about 3 billion HTML pages from more than 30 million different pay-level domains (PLDs). The Web Data Commons project has been extracting structured data from the Common Crawl every year since 2010[7]. Since 2013, we have made the schema.org data we extract from the crawl available for public download in the form of the schema.org data set series.

In this poster, we profile the schema.org data set series and identify trends in the adoption of schema.org vocabulary over the last 10 years. Afterwards, we exemplify the potential of the data set series to be used as a large pool of training data for tasks such as product matching, product or job categorization, or information extraction. The poster is structured as follows. Section 2 gives an overview of the extraction framework that has been used to create the schema.org data set series. Section 3 describes how the data is made available for public download. Section 4 analyses the adoption of schema.org annotations over the period 2013 to 2022. Section 5 profiles the content of the latest data set in the series. Section 6 discusses the use of selected schema.org classes as training data for various machine learning tasks. Section 7 discusses related work.

## 2 EXTRACTION PROCESS

The Web Data Commons project has developed a parsing framework for extracting structured data from the Common Crawl. The

[1] https://schema.org/

[2] https://developers.google.com/search/docs/appearance/structured-data/intro-structured-data

[3] https://www.bing.com/webmasters/help/marking-up-your-site-with-structured-data-3a93e731

[4] https://yandex.com/support/webmaster/schema-org/what-is-schema-org.html

[5] https://developers.google.com/search/docs/appearance/structured-data/search-gallery

[6] https://commoncrawl.org/

[7] http://webdatacommons.org/structureddata/

*WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA*

ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587331

framework runs in the AWS cloud and supports the parallel processing of multiple (W)ARC files. The framework uses the Any23 parser library [8] to extract JSON-LD, Microdata, RDFa, and Microformats data from the HTML pages contained in the (W)ARC files. The extracted data is represented in the form of RDF quads (N-Quads[9]). An RDF quad consists of an RDF statement plus a fourth element that contains the URL of the web page from which the data was extracted [7]. Since webmasters primarily use the JSON-LD and Microdata formats to annotate web pages with schema.org terms, we merge the extracted JSON-LD and Microdata data and afterwards create class-specific subsets for a selection of schema.org classes. The subsets contain all the entities of a particular class together with the entities of other classes that are present on the same page. For example, a page containing data about a product might also contain reviews and offers for that product; a page containing data about an event might also contain data about the location of the event and the people who are performing at the event.

## 3 DATA PROVISIONING

The schema.org data sets series is available for public download on the WDC website[10]. The complete data set series is 7.6 TB in size, with an average size of 770 GB per data set. The data sets are split into multiple files of 1.5 GB each. For users who prefer formats other than N-Quads, we provide code[11] to convert the download files into CSV and JSON formats. The December 2020 release of the WDC schema.org data set series is also available as a table corpus[12], which has been created by grouping the data into separate tables for each class/host combination, e.g. all records of a specific class extracted from a specific website are put into a single table. The resulting corpus consists of 4.2 million relational tables, which are available for download in a JSON format that can be read by the pandas[13].

## 4 EVOLUTION OF THE SCHEMA.ORG ADOPTION FROM 2013 TO 2022

This section discusses the evolution of the adoption of the schema.org vocabulary on the Web based on statistics that we have derived from the data set series. When interpreting the statistics, it needs to be considered that the Common Crawl only covers a sample of approximately 30 million popular pay-level domains (PLDs) while the full Web contains more websites.

**Number of Websites using the schema.org Vocabulary.** Figure 1 shows the overall number of PLDs in the Common Crawl that offer schema.org annotations (blue line). We see that the number of PLDs has grown significantly from 400,000 in 2013 to over 12 million in 2022. If we relate these numbers to the overall number of PLDs covered by the Common Crawl, we see that in 2013 only 3.1% of all websites use schema.org annotations, while the percentage grows to 37.9% in 2022. The red line in Figure 1 shows the number of PLDs that use the Microdata syntax for embedding schema.org annotations into their pages. The orange line visualizes the number of PLDs that use the JSON-LD syntax for embedding data into the HEAD of HTML pages. We see that since 2020, more PLDs use the JSON-LD syntax than Microdata. One factor that likely contributed to this development is Google's recommendation in 2016 to prefer the JSON-LD syntax over the Microdata syntax for annotating HTML pages [14].

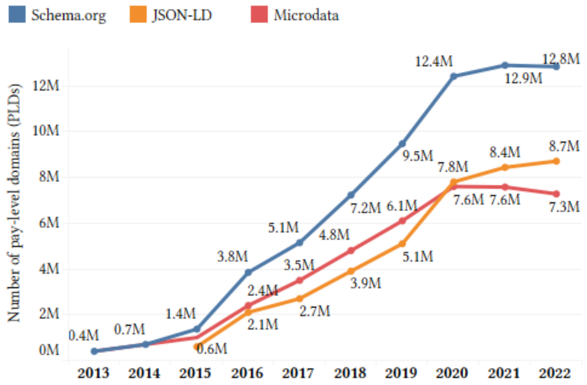

**Figure 1: Number of pay-level domains (PLDs) using schema.org terms (blue) together with either the Microdata (red) or the JSON-LD (orange) syntax.**

**Growth of Popular Schema.org Classes.** Figure 2 shows the evolution of the number of websites that annotate data describing products, local businesses, events, and job postings over the period 2013 to 2022. The figure uses a log scale for displaying the number of PLDs. Over the last five years the number of websites providing `Product` annotations increased from 594K to 2.6M (430% growth), the number of websites annotating `LocalBusiness` entities increased from 386K to 1.2M (310% growth), while the adoption of the `JobPosting` class increased from 7K websites to 50K (721% growth).

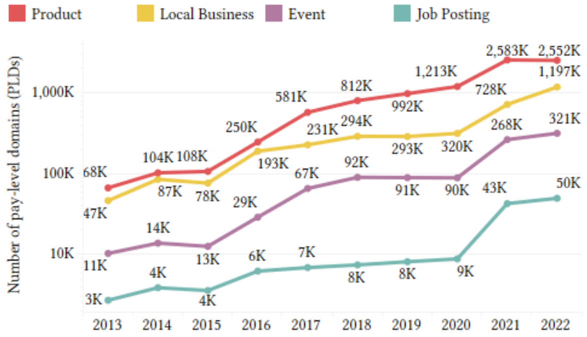

**Figure 2: Growth of the number of PLDs that annotate specific schema.org classes in the period from 2013 to 2022**

[8] https://github.com/apache/any23
[9] https://www.w3.org/TR/n-quads/
[10] https://webdatacommons.org/structureddata/schemaorg/
[11] https://github.com/wbsg-uni-mannheim/StructuredDataProfiler
[12] http://webdatacommons.org/structureddata/schemaorgtables/
[13] https://pandas.pydata.org/
[14] https://web.archive.org/web/20160529043847/https://developers.google.com/search/docs/guides/intro-structured-data#common-use-cases

**Richness of the Descriptions.** On average, we extracted 32 Microdata RDF quads from each webpage in 2013. The number of RDF quads per page increased to 36 RDF quads per webpage in 2022. For JSON-LD annotations the average number of RDF quads per webpage increased from 10 in 2015 to 52 in 2022. The increased number of quads per page, especially for JSON-LD annotations, indicates that the level of detail of the schema.org data has increased over the last decade.

## 5 CONTENT OF THE 2022 RELEASE

This section provides an overview of the content of the 2022 release of the WDC schema.org data set series. The release consists of 106 billion RDF quads describing 3.1 billion entities belonging to 44 different classes. Table 1 lists the number of entities for selected classes as well as the number of PLDs from which the descriptions originate. For instance, the data set contains descriptions of 600 million products originating from over 2 million different PLDs. The data set contains 70 million records describing events, 5 million records describing job postings, and 47 million questions from Q&A pages together with 52 million answers. Table 2 lists the properties that are used in the data set to describe products. The first column shows the percentage of PLDs that use a specific property to describe products. The density column contains the percentage of the entities of a class for which the specific property is filled. The properties `name` , `image` , `offers` and `description` are the most popular properties used by more than 85% of all PLDs that annotate product data. The percentage of PLDs offering specific properties drops quickly down the list. For instance, only 12% of all PLDs annotate categorization information for the products that they offer (`category`). Table 3 and Table 4 list the properties that are used to describe job postings and events. The classes show a similar pattern where some properties are widely used while others are hardly used. Comparing the widely used properties with the properties that are required for content to appear in Google's schema.org applications[15] reveals a clear correlation, supporting the hypothesis that appearing in these applications is a major motivation for annotating web content.

**Table 1: Number of PLDs and number of entities per schema.org class found in the 2022 data set**

| schema.org Class | #PLDs | #Entities |
|---|---:|---:|
| Person | 4,342 K | 932 M |
| Product | 2,552 K | 600 M |
| Offer | 2,376 K | 635 M |
| LocalBusiness | 1,197 K | 63 M |
| BlogPosting | 905 K | 144 M |
| AggregateRating | 814 K | 90 M |
| Review | 331 K | 110 M |
| Event | 314 K | 70 M |
| Question | 200 K | 47 M |
| Answer | 199 K | 52 M |
| JobPosting | 50 K | 5 M |

**Table 2: `Schema.org/Product`: Relative PLD usage, density and relative entity usage of selected properties.**

| Product Property | % PLDs | % Density |
|---|---:|---:|
| name | 98.67 | 96.84 |
| image | 87.14 | 77.37 |
| offers | 86.16 | 80.1 |
| description | 85.75 | 67.01 |
| url | 64.4 | 51.34 |
| sku | 60.52 | 41.66 |
| aggregateRating | 15.85 | 10.76 |
| category | 12.18 | 8.93 |
| mpn | 8.59 | 11.32 |
| productID | 5.37 | 7.27 |
| gtin13 | 4.06 | 4.71 |
| gtin8 | 1.33 | 0.73 |

**Table 3: `Schema.org/JobPosting`: Relative PLD usage, density and relative entity usage of selected properties**

| JobPosting Property | % PLDs | % Density |
|---|---:|---:|
| description | 96.61 | 95.22 |
| title | 96.59 | 97.2 |
| datePosted | 95.75 | 94.53 |
| hiringOrganization | 94.49 | 95.13 |
| jobLocation | 92.42 | 93.94 |
| employmentType | 74.2 | 72.79 |
| baseSalary | 37.97 | 40.00 |
| industry | 17.91 | 32.27 |
| skills | 9.47 | 15.31 |
| occupationalCategory | 6.27 | 15.29 |

**Table 4: `Schema.org/Event`: Relative PLD usage, density and relative entity usage of selected properties**

| Event Property | % PLDs | % Density |
|---|---:|---:|
| startDate | 97.71 | 96.62 |
| name | 97.61 | 98.23 |
| description | 89.02 | 78.98 |
| location | 85.75 | 75.09 |
| image | 66.94 | 51.71 |
| performer | 42.57 | 47.83 |
| eventStatus | 38.94 | 22.22 |

## 6 USING SCHEMA.ORG AS TRAINING DATA

The wide adoption of certain properties makes schema.org data valuable training data for various machine learning tasks. In the following, we give an overview of some of these use cases.

**Product Matching.** The goal of product matching is to identify offers that refer to the same real-world product. Among the `schema.org/ Product` properties in Table 2, the following properties contain product identifiers: `gtin13`, `gtin8`, `mpn`, `productID`,

[15] https://developers.google.com/search/docs/appearance/structured-data/sd-policies

and sku. The identifiers allow offers for the same real-world product from multiple e-shops to be grouped into clusters [10]. Pairs of matching and non-matching product descriptions derived from these clusters can be used to train product matchers. The WDC Product Data Corpus and Gold Standard for Large-Scale Product Matching[16] is a benchmark that uses pairs of offers from the schema.org data set series to train and test product matchers [10].

**Product Categorization.** schema.org/Product annotations can also be used as training data for learning how to categorize products into standard product hierarchies like GS1 GPC[17] [2, 8]. Meusel et al. [8] used the annotated category values as features for training a classifier to predict product category labels given product descriptions with no category information [8]. Brinkmann and Bizer [2] tackle the task of hierarchical product categorization by pre-training a ROBERTa transformer model with the values of the schema.org/Product properties title, description, and category in order to inject domain-specific knowledge into the model. The model is afterwards fine-tuned for hierarchical product categorisation using manually labelled product offers.

**Job Categorization.** The schema.org/ JobPosting annotations can be exploited to train a classifier for categorizing job offers. The occupationalCategory property often contains values from HR taxonomies BLS O*NET-SOC[18] and ISCO-08[19]. A classifier trained to predict the occupational category given values of schema.org/JobPosting properties such as title, description and skills can be used to categorize job postings where the occupational category is missing.

**Information Extraction.** Not all events published on the Web are annotated using the schema.org vocabulary. To extract information about non-annotated events, schema.org/Event annotations can be used to train event extractors. Foley et al. [3] used the schema.org/Event properties name, startDate, and location to train an extractor for collecting data about local events, such as small venue concerts, theatre performances, and garage sales. They showed that using the annotations improves the recall of the system.

## 7 RELATED WORK

The major search engine companies Google, Yahoo!, Microsoft, and Yandex extract structured data from their web crawls, but do not provide public access to the data for commercial reasons. However, they have published a number of studies on the adoption of markup languages: Mika and Potter analyse the adoption of the languages based on web crawls of the Bing search engine from 2011 and 2012 [9]. Guha et al. present an updated analysis of the deployment of Microdata with a particular focus on the Schema.org vocabulary [4]. Alrashed et al. state that 61% of the hosts in a Google crawl that provide Schema.org/Dataset annotations do not actually describe datasets and propose a classifier to identify such low-quality annotations [1]. Kanza et al. [5] analyze the social processes around creating and maintaining the schema.org vocabulary. Meusel et al. studied the adoption of the schema.org vocabulary over the period 2012 to 2014 using early extractions from the Web Data Commons project [6]. Sections 4 and 5 of this poster can be seen as a continuation and update to this work. The Web Almanac[20] also tracks the adoption of semantic annotations on the Web based on a crawl by the HTTP archive. The major difference to our work is that the HTTP archive crawls only the front page of a website while the Common Crawl contains multiple pages per website. This difference in the crawl depth is likely the main reason for the difference in the reported statistics. Other efforts to extract training data from CommonCrawl include the OSCAR text corpus[21] which is used for training language models [12], as well as the LAION-5B dataset[22], which consists of 5.8 billion image-text pairs for training vision models [11].

## 8 CONCLUSION

This poster presented a series of publicly available schema.org data sets which have been extracted from the Common Crawl over the course of the last ten years. To the best of our knowledge, the Web Data Commons schema.org data set series is the only publicly available data set of its kind. Schema.org annotations are used as training data for tasks such as product matching, product or job categorization, and event information extraction. We hope that the Web Data Commons schema.org data set series will prove useful for even more applications in the future.

## REFERENCES

[1] Tarfah Alrashed, Dimitris Paparas, Omar Benjelloun, and al. 2021. Dataset or Not? A study on the veracity of semantic markup for dataset pages. In *Proceedings of the 20th International Semantic Web Conference*. 338–356.

[2] Alexander Brinkmann and Christian Bizer. 2021. Improving Hierarchical Product Classification using Domain-specific Language Modelling. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* (2021), 14–25.

[3] John Foley, Michael Bendersky, and Vanja Josifovski. 2015. Learning to Extract Local Events from the Web. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 423–432.

[4] R. V. Guha, Dan Brickley, and Steve Macbeth. 2016. Schema.org: evolution of structured data on the web. *Commun. ACM* 59, 2 (2016), 44–51.

[5] Samantha Kanza, Alex Stolz, Martin Hepp, and Elena Simperl. 2018. What Does an Ontology Engineering Community Look Like? A Systematic Analysis of the schema.org Community. In *Proceedings of the 15th European Semantic Web Conference*. 335–350.

[6] Robert Meusel, Christian Bizer, and Heiko Paulheim. 2015. A Web-scale Study of the Adoption and Evolution of the schema.org Vocabulary over Time. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*. 1–11.

[7] Robert Meusel, Petar Petrovski, and Christian Bizer. 2014. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In *Proceedings of the 13th International Semantic Web Conference*. 277–292.

[8] Robert Meusel, Anna Primpeli, Christian Meilicke, Heiko Paulheim, and Christian Bizer. 2015. Exploiting Microdata Annotations to Consistently Categorize Product Offers at Web Scale. In *Proceedings of the 16th International Conference on Electronic Commerce and Web Technologies*. 83–99.

[9] Peter Mika and Tim Potter. 2012. Metadata Statistics for a Large Web Corpus. In *Proceedings of the Linked Data Workshop (LDOW) at the International World Wide Web Conference*.

[10] Anna Primpeli, Ralph Peeters, and Christian Bizer. 2019. The WDC Training Dataset and Gold Standard for Large-Scale Product Matching. In *Companion Proceedings of The 2019 World Wide Web Conference*. 381–386.

[11] Christoph Schuhmann, Romain Beaumont, Richard Vencu, and al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 [cs].

[12] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora*. 9–16.

[16] http://webdatacommons.org/largescaleproductcorpus/v2/

[17] https://www.gs1.org/standards/gpc

[18] https://www.onetcenter.org/taxonomy.html

[19] https://www.ilo.org/public/english/bureau/stat/isco/isco08/

[20] https://almanac.httparchive.org/en/2022/structured-data

[21] https://oscar-project.org/

[22] https://nips.cc/virtual/2022/poster/55659