The Web Data Commons Schema.org Table Corpora

Ralph Peeters
ralph.peeters@uni-mannheim.de
University of Mannheim
Mannheim, Germany

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

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

ABSTRACT

The research on table representation learning, data retrieval, and data integration in the context of data lakes requires large table corpora for the training and evaluation of the developed methods. Over the years, several large table corpora such as WikiTables, GitTables, or the Dresden Web Table Corpus have been published and are used by the research community. This paper complements the set of public table corpora with the Web Data Commons Schema.org table corpora, two table corpora consisting of 4.2 (Release 2020) and 5 million (Release 2023) relational tables describing products, events, local businesses, job postings, recipes, movies, books, as well as 37 further types of entities. The feature that distinguishes the corpora from all other publicly available large table corpora is that all tables that describe entities of a specific type use the same attributes to describe these entities, i.e. all tables use a shared schema, the schema.org vocabulary. The shared schema eases the integration of data from different sources and allows training processes to focus on specific types of entities or specific attributes. Altogether the tables contain ~653 million rows of data which have been extracted from the Common Crawl web corpus and have been grouped into separate tables for each class/host combination, i.e. all records of a specific class that originate from a specific website are put into a single table. This paper describes the creation of the WDC Schema.org Table Corpora, gives an overview of the content of the corpora, and discusses their use cases.

CCS CONCEPTS

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

KEYWORDS

table corpora; information extraction; schema.org; semantic web

ACM Reference Format:
Ralph Peeters, Alexander Brinkmann, and Christian Bizer. 2024. The Web Data Commons Schema.org Table Corpora. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651441

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651441

1 INTRODUCTION

The schema.org[1] community effort defines a shared vocabulary for describing entities such as products, local businesses, events, job offers, questions and answers, as well as many other types of entities [7, 10]. Schema.org terms are used together with the Microdata and RDFa syntaxes to annotate structured data within the BODY of HTML pages. Alternatively, the terms are used in combination with the JSON-LD syntax to embed structured data in the HEAD section of HTML pages. Since 2011, the search engines Google[2], Bing[3], and Yandex[4] use schema.org data to display rich snippets in search results, display info boxes next to search results, and entities on maps. Other applications that use schema.org data include Google Shopping, Google for Jobs, and Google Dataset Search[5]. In order to have their content displayed in these applications, millions of websites have started to use the schema.org vocabulary to annotate structured data within their pages and today approximately 50% of all web pages contain schema.org annotations [2]. The Web Data Commons (WDC) project[6] regularly extracts schema.org data from the Common Crawl[7], the largest public web corpus. The Common-Crawl is released monthly and typically contains around 3 billion HTML pages originating from over 30 million different websites (hosts). The WDC project uses the extracted data to calculate statistics about the adoption of schema.org on the Web [2] and publishes the extracted data in the form of N-Quads[8], a provenance-enabled graph data format.

In order to allow the extracted data to be directly used for applications that require tabular data, as well as to prevent users from needing to deal with duplicate data resulting from the structure of the original websites, we generate the WDC Schema.org Table Corpora from the extracted data by (i) grouping the data by website, (ii) removing incomplete entities that were extracted from listing pages and (iii) deduplicating it. We represent the resulting tables in a JSON format that can directly be read by the pandas[9] library. We publish two releases of the WDC Schema.org Table Corpus: The 2020 release consists of 4.2 million relational tables that together contain ~292 million rows of data. The corpus was generated from the WDC 2020 JSON-LD and Microdata extraction. The second corpus contains 5 million tables with together ~361 million rows of

[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] http://webdatacommons.org/structureddata/
[7] https://commoncrawl.org/
[8] https://www.w3.org/TR/n-quads/
[9] https://pandas.pydata.org/

data and was generated from the WDC 2023 extraction. The tables in the corpora belong to 44 different schema.org classes with product, local business, and event being the most widely used classes. As the tables are generated from schema.org annotations, all tables that describe entities of a specific type use the same set of attributes to describe these entities, i.e. all tables use a shared schema. However, as the data originates from over 4.3 million different websites (hosts) from all across the Web, the actual data values are heterogeneous concerning value format, unit of measurement, and language.

This paper is structured as follows: Section 2 gives an overview of the creation process of the WDC Schema.org Table Corpora. Sections 3 and 4 describe the data format and provisioning of the corpora. Section 5 presents profiling statistics about the content of the 2023 release of the corpus. Section 6 compares the table corpora to other table corpora from related work. Section 7 discusses use cases of the WDC Schema.org Table Corpora, such as training general-propose or task-specific embedding models, benchmarking data integration methods, analyzing the deployment of Semantic Web technologies, and using the corpora as sources of domain-specific data.

2 CREATION PROCESS

This section describes the process of creating the WDC Schema.org Table Corpora. We use the 2023 release to illustrate the number of tables and rows and the amount of computation used.

1. Extracting Data from the Common Crawl. The WDC project has developed a parsing framework for extracting structured data from the Common Crawl [14]. The framework runs in the AWS cloud and supports the parallel processing of multiple (W)ARC files. To extract JSON-LD, Microdata, RDFa, and Microformats data from the HTML pages contained in the (W)ARC files, the framework uses the Any23 parser library[10]. For the 2023 release, we used 250 AWS spot instances with 8 × 3.2 GHz CPUs and 16 GB RAM for the extraction which altogether required 4,602 machine hours. The extracted corpus consists of 97 billion RDF quads (N-Quads). Webmasters primarily use JSON-LD and Microdata syntaxes to annotate web pages with schema.org terms. Therefore, we merge the extracted JSON-LD and Microdata data to form class-specific subsets for selected schema.org classes. The subsets consist of all entities of a specific class along with entities of other classes present on the same page and contain 39 billion RDF quads[11]. It took 5 days of compute time on a local shared server equipped with 96 × 3.6 GHz CPU cores and 1024 GB RAM to create the schema.org subsets.

2. Group by Host. Next, all entities, corresponding attributes, and attribute values are converted from RDF quads into tabular form and grouped by host. If an attribute contains child entities instead of a literal value, all child entities and their attributes are extracted as a list. However, only literal values are considered for the attributes of child entities, dismissing any child entity attributes further down the hierarchy. For example, a web page about a movie might annotate the name of the movie and details of the actors who appear in the movie including their names and their spouses. In this case, the movie's name and a list of actors are extracted. For each actor, only the actor's name is extracted because it is a literal value. Child entities further down the hierarchy such as the actor's spouse are omitted. After this step, the 2023 version of the table corpus contains ∼7.5 million tables with overall ∼1.4 billion rows.

3. Removal of Listing Pages and Sparse Entities. Listing pages contain concise information about entities that are described in more detail on other pages. In order to have attribute-rich entity descriptions in the corpus, we want to exclude descriptions originating from listing pages. Other pages provide detailed descriptions of one entity and brief descriptions of other entities as part of navigation elements or advertisements. Our objective is to extract only the main entity from such pages. We apply the following heuristic to exclude these sparse entities: If a web page contains only one relevant entity with at least three attributes, the entity is extracted. For web pages that contain multiple entities, we concatenate the attribute values of each entity and calculate the mean absolute deviation (MAD) of each entity based on the length of the concatenated attribute values. Entities with at least three attributes and concatenated attribute value lengths greater than the median plus three times the MAD (positive outliers) are extracted. If a web page marks up multiple entities without outliers, those entities are dismissed as originating from a listing page. Applying the heuristics reduces the size of the corpus to ∼5 million tables and ∼429 million rows.

4. Content-based Deduplication. Content-based deduplication removes exactly equal entity descriptions that originate from different web pages of the same host. This process is applied to all attributes except for schema.org/url, which is excluded for top and second-level attributes as it may differ and lead to false positives during deduplication. We keep only attributes with a density above 25% for the final table for each host and dismiss all sparser attributes. After content-based deduplication, the 2023 release of the WDC Schema.org Table Corpus contains ∼5 million tables containing alltogether ∼361 million rows of data. It took 10 days of compute time on a local shared server equipped with 96 × 3.6 GHz CPU cores and 1024 GB RAM to create the 2023 release from the extracted RDF quads resulting from Step 1.

3 DATA FORMAT

The tables are encoded in JSON line data format and can be read by the pandas Python library. An example table of the class Movie is shown in Figure 1. Each record represents an entity annotated using the schema.org vocabulary. The 'row_id' is an identifier for the extracted entity that is created during the extraction process. The 'name' column contains literal values of the schema.org/name attribute of the extracted entity, while the 'actor' column shows an example of extracted child entities of type schema.org/actor. These are represented as lists containing all literal attribute names and values of the respective child entity. Due to space limitations, only three actors are shown in Figure 1.

4 DATA PROVISIONING

Both the 2023[12] and the 2020[13] release of the WDC Schema.org Table Corpus are provided for public download on the WDC website. The 2023 release has a size of 71GB in zipped format. The 2020

[10] https://github.com/apache/any23
[11] https://webdatacommons.org/structureddata/schemaorg/
[12] https://webdatacommons.org/structureddata/schemaorgtables/2023/
[13] https://webdatacommons.org/structureddata/schemaorgtables/

```
{   "row_id":001,
    "name":"Ant-Man",
    "actor":   [{"name":"Michael Douglas"},
               {"name":"Evangeline Lilly"}] }
{   "row_id":002,
    "name":"Into the Wild",
    "actor":   {"name":"Sean Penn"} }
```

Figure 1: JSON serialization of a table describing movies.

release has a size of 51GB. For each class of entities, we provide three download files: (1) the top 100 tables containing the largest number of rows, (2) tables containing at least 3 rows and, (3) the tail tables containing the remaining smaller tables. The user may choose to use one or any combination of these files, depending on the intended application.

5 CONTENT OF THE 2023 TABLE CORPUS

This section presents profiling information for the 2023 release of the WDC Schema.org Table Corpus. The corpus comprises around 5 million tables, containing over 361 million rows of data in total. The tables cover 42 schema.org classes and originate from over 4.33 million websites (hosts).

Tables by Class. The statistics on the number of tables per class, their rows, and the average number of attributes for a selection of schema.org classes are presented in Table 1. The schema.org classes selected demonstrate the breadth of the corpus, ranging from classes with extensive tables, such as `Product`, to those with fewer tables, such as `Dataset`. In addition to the statistics for the complete corpus (Overall), Table 1 provides separate statistics for the largest 100 tables and tables with at least three rows (Minimum 3). For example, the corpus contains over one million `LocalBusiness` tables describing altogether 8 million business entities with on average 4 attributes, such as name, address, telephone, or average rating. By distinguishing between the Top 100 and Minimum 3 tables, we can see that the Top 100 tables account for 1% to 11% of all rows for the three most popular classes: `Product`, `LocalBusiness`, and `Event`.

Attributes. The WDC Schema.org Table Corpus is constructed using schema.org annotations. As a result, all tables share the same set of attributes, while the attributes present in a specific table depend on the annotations included by the corresponding host in its web pages. Table 2 shows a selection of attributes appearing in tables for the classes `Product`, `LocalBusiness` and `Movie`. Common attributes such as `name` and `description` are present in many tables across multiple classes. Other attributes, such as `productID` and `genre`, are more class-specific and less frequently used, indicating that a long-tail distribution for such attributes exists in the corpus. Nevertheless, for both head and long-tail attributes, the tables exhibit a high average value density of 95%. This shows that if hosts use a schema.org term, they do so consistently. Some attributes are entity identifiers that can be used to link entities across tables for example to derive training data for entity matching tasks. Examples of such attributes are `SKU, productID, MPN` and `GTIN13` for the class `Product` as well as the `telephone` number for `LocalBusiness`.

Table 1: Number of tables and rows for selected schema.org classes in millions (M) and thousands (k).

	Overall			Top100	Minimum 3	
Class	**Tables**	**Rows**	**Avg. Attr.**	**Rows**	**Tables**	**Rows**
Product	3M	288M	5	4M	2M	283M
LocalBusiness	1M	8M	4	903k	65k	6M
Event	368k	15M	8	1M	261k	14M
Restaurant	60k	716k	7	348k	7k	313k
JobPosting	58k	3M	7	543k	37k	3M
Recipe	41k	4M	11	619k	33k	3M
Question	38k	4M	6	2M	21k	2M
Hotel	22k	2M	6	794k	9k	880k
Book	14k	3M	6	944k	10k	2M
Movie	6k	2M	7	481k	5k	1M
SportsEvent	4k	579k	6	296k	3k	281k
Hospital	2k	40k	6	31k	396	7k
Dataset	2k	364k	7	286k	1k	77k

Table 2: Fraction of tables containing specific attributes.

Product		LocalBusiness		Movie	
Attribute	**in % of tables**	**Attribute**	**in % of tables**	**Attribute**	**in % of tables**
name	100	name	99	name	96
offers	96	address	97	description	73
description	86	telephone	91	director	63
sku	57	aggregaterating	19	datecreated	55
brand	33	geo	16	aggregaterating	38
image	28	pricerange	12	duration	36
category	11	email	11	actor	36
aggregaterating	9	description	11	genre	26
productid	7	openinghoursspec	10	datepublished	25
mpn	7	url	8	image	15
gtin13	3	image	7	url	13

6 RELATED WORK

Various table corpora have been created in recent years. Table 3 lists table corpora and shows statistics on their number of tables (Tabs.), the average number of rows (Avg. # Rows) and attributes (Avg. # Attr.), and whether all tables in the corpus use a single shared schema. The WDC Web Table Corpus [12] and the Dresden Web Table corpus [6] extract relational HTML tables from web pages in the Common Crawl. These web tables [17] have been used in related work on table search [4] and table augmentation [3]. The WikiTables table corpus contains tables that were extracted from Wikipedia [1]. VizNet [8] consists of tables that were chosen for benchmarking visualization methods. Open Data Portal Watch [15] contains tabular data that was collected from open data portals. Table 3 shows that the tables in the WDC Web Table Corpus, the Dresden Web Table Corpus, WikiTables and VizNet have a relatively small number of rows. Compared to the web tables corpora, the WDC schema.org table corpora and GitTables [9] contain on average more rows. GitTables [9] consists of tables that are extracted from CSV files shared on GitHub. The tables in all the referenced table corpora do not use a single shared schema but each table uses a different, proprietary schema. As a post-processing step, the tables in GitTables are annotated with semantic types from DBpedia and schema.org using an automated annotation method which might misinterpret table semantics [9]. The WDC Schema.org Table

Corpora are generated from schema.org annotations. As a result, the tables in our table corpora follow a single shared schema.

Table 3: Related table corpora.

Table Corpus	Single Schema	Tabs.	Avg. # Rows	Avg. # Attr.
Dresden Web Tables Corpus [6]	✗	59M	17	6
WDC Web Tables Corpus 2015 [12]	✗	90M	14	5
WikiTables [1]	✗	15M	15	6
VizNet [8]	✗	31M	17	3
Open Data Portal Watch [15]	✗	1M	17	14
GitTables [9]	✗	2M	209	25
WDC Schema.org Corpus 2020	✓	4M	89	5
WDC Schema.org Corpus 2023	✓	5M	72	5

7 USE CASES

This section describes various use cases of the WDC Schema.org Table Corpus.

Benchmarking. The table corpus is a useful resource for evaluating table annotation, schema matching, entity matching, and data retrieval methods due to its shared schema and the presence of entity identifiers such as GTINs, MPNs, or phone numbers in many tables. For example, the SOTAB Table Annotation Benchmark [11] was constructed by selecting a subset of the tables from the 2020 version of the corpus, removing the attribute labels from the tables, and having the annotation systems predict the removed labels. The SOTAB Benchmark was used in the 2023 edition of the SemTab challenge[14]. A second benchmark that uses tables from the 2020 version of the corpus is the WDC Schema Matching Benchmark[15] which requires instance-based schema matching methods to discover correspondences between table columns. A benchmark that uses schema.org entity identifiers as ground truth is the WDC Products entity matching benchmark [16].

Source of Training Data. Its structuredness, the common schema, and shared entity identifiers also make the corpus an interesting source of (pre-)training data for table representation learning as well as data integration, e.g. entity matching [16] and table annotation [11]. The corpus further contains 4 million question-answer pairs originating from 38,000 websites (see row `Question` in Table 2) which could be used for fine-tuning LLMs or as background knowledge for retrieval-augmented question answering. Besides such task-specific training, the corpus can also be used as a structured pre-training resource for table representation learning methods [5] or for fine-tuning LLMs for structured data tasks [13].

Analyzing the Adoption of Semantic Web Technologies. We publish detailed statistics about which host uses which schema.org terms together with the corpora. From a web science perspective, these statistics together with the data itself can be used to analyze the adoption of the schema.org vocabulary within specific application domains as well as on the Web in general [2, 7].

Source of Domain Data. Last but not least, the corpus can be used as a large source of domain data. For example, if a user wants to assemble a list of shops or hotels in a city, the 1 million local business tables in the corpus with together 8 million rows could be a useful starting point. Or, if the user wants to analyze the skills that are currently in demand on the job market, they could use the 3 million job postings in the corpus for their analysis.

[14]https://sem-tab-challenge.github.io/2023/
[15]https://webdatacommons.org/structureddata/smb/

8 CONCLUSION

This paper presented the WDC Schema.org Table Corpora that we extracted from the Common Crawl. To the best of our knowledge, the corpora are the largest public table corpora that contain tables from many different sources which all share a common schema. We hope that the WDC Schema.org Table Corpora will be useful for a variety of use cases beyond those already described.

REFERENCES

[1] Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. 2013. Methods for Exploring and Mining Tables on Wikipedia. In *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics*. 18–26.

[2] 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*. ACM, Austin TX USA, 136–139.

[3] Michael Cafarella, Alon Halevy, Hongrae Lee, et al. 2018. Ten Years of Webtables. *Proceedings of the VLDB Endowment* 11, 12 (Aug. 2018), 2140–2149.

[4] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset Search: A Survey. *The VLDB Journal* 29, 1 (2020), 251–272.

[5] Xiang Deng, Huan Sun, Alyssa Lees, et al. 2020. TURL: Table Understanding through Representation Learning. *Proceedings of the VLDB Endowment* 14, 3 (Nov. 2020), 307–319.

[6] Julian Eberius, Katrin Braunschweig, Markus Hentsch, et al. 2015. Building the Dresden Web Table Corpus: A Classification Approach. In *2015 IEEE/ACM 2nd International Symposium on Big Data Computing*. 41–50.

[7] R. V. Guha, Dan Brickley, and Steve Macbeth. 2016. Schema.org: Evolution of Structured Data on the Web. *Commun. ACM* 59, 2 (Jan. 2016), 44–51.

[8] Kevin Hu, Snehalkumar 'Neil' S. Gaikwad, Madelon Hulsebos, et al. 2019. VizNet: Towards A Large-Scale Visualization Learning and Benchmarking Repository. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk, 1–12.

[9] Madelon Hulsebos, Çagatay Demiralp, and Paul Groth. 2023. GitTables: A Large-Scale Corpus of Relational Tables. *Proceedings of the ACM on Management of Data* 1, 1 (May 2023), 30:1–30:17.

[10] Samantha Kanza, Alex Stolz, Martin Hepp, et al. 2018. What Does an Ontology Engineering Community Look Like? A Systematic Analysis of the Schema.Org Community. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018*. 335–350.

[11] Keti Korini, Ralph Peeters, and Christian Bizer. 2022. SOTAB: The WDC Schema.org table annotation benchmark. In *CEUR Workshop Proceedings*, Vol. 3320. RWTH Aachen, Aachen, Germany, 14–19.

[12] Oliver Lehmberg, Dominique Ritze, Robert Meusel, et al. 2016. A Large Public Corpus of Web Tables containing Time and Context Metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*. 75–76.

[13] Peng Li, Yeye He, Dror Yashar, et al. 2023. Table-GPT: Table-tuned GPT for Diverse Table Tasks. *arXiv preprint arXiv:2310.09263* (2023).

[14] Robert Meusel, Petar Petrovski, and Christian Bizer. 2014. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In *The Semantic Web – ISWC 2014*. 277–292.

[15] Johann Mitlöhner, Sebastian Neumaier, Jürgen Umbrich, et al. 2016. Characteristics of Open Data CSV Files. In *2016 2nd International Conference on Open and Big Data*. 72–79.

[16] Ralph Peeters, Reng Chiz Der, and Christian Bizer. 2024. WDC Products: A Multi-Dimensional Entity Matching Benchmark. In *Proceedings of the 27th International Conference on Extending Database Technology, Paestum, Italy*. 22–33.

[17] Shuo Zhang and Krisztian Balog. 2020. Web Table Extraction, Retrieval, and Augmentation: A Survey. *ACM Trans. Intell. Syst. Technol.* 11, 2 (March 2020), 1–35.