

The Shape of Schema.org at Scale: A Seven-Month Analysis of Structured Data Across 14.78 Billion Web Pages

Brie Moreau¹, Ilmi Sejfula¹, and Taha Shah¹

¹White Light Digital Marketing, AI Research Lab

September 2026

Abstract

Schema.org markup has become the primary vocabulary through which the open web declares machine-readable structure, yet public reporting on its aggregate shape at web scale remains limited. This paper presents a seven-month empirical study of schema.org usage across all Common Crawl monthly snapshots from January 2026 through July 2026, covering 14,784,536,534 pages and 127,990,580,002 individual type occurrences. Across the combined corpus, 58.32% of pages carry schema markup (weighted mean), coverage rises from 55.27% in the lowest month (February) to 60.05% in the highest (July), monotonically from March onward, and vocabulary usage is concentrated to a degree that decays faster than pure Zipf: the top 100 types account for approximately 98% of all instances, while more than 62% of distinct types appear only once. A stratified sample-based format analysis shows JSON-LD page share growing from 74.53% to 77.84% and microdata share falling from 49.40% to 45.14% across the period, monotonically each month. Implementation errors (casing duplicates, property names used as `@type`, URLs used as `@type`, and malformed strings) form a stable shadow layer at 0.63% of all occurrences. Seven co-occurrence templates (breadcrumb, FAQ, product, enhanced e-commerce, search box, local business, article) are recovered every month at conditional probabilities above 87%. A full-corpus Wikidata linking scan finds 5,382,130 pages (0.261% of all pages) carrying at least one Wikidata QID reference. Compared with Google’s Schema.org Usage Statistics Dataset, which measures domain breadth, our instance-volume ranking of the top schema types shows a Spearman rank correlation of 0.86, indicating that the two orthogonal signals broadly agree on the relative prominence of the schema.org vocabulary.

1 Introduction

Schema.org is a shared vocabulary maintained by a consortium originally founded by Google, Microsoft, Yahoo, and Yandex [2]. It provides a fixed set of type and property names that publishers embed in their HTML so that search engines, assistants, and other consumers can extract structured information without site-specific parsers. Since its launch in 2011 the vocabulary has grown to cover several hundred types, and its adoption has become central to how e-commerce, editorial, and local business content is surfaced by contemporary retrieval systems.

Despite the vocabulary’s practical importance, public aggregate measurements of how schema.org is actually deployed at web scale are sparse. Web Data Commons [4] has published periodic extractions from Common Crawl since 2012, and Google publishes a Schema.org Usage Statistics Dataset [3] that reports how many distinct domains use each type. These are the two dominant public reference points. Neither, however, captures the full instance-level distribution across a large modern corpus, and neither publishes a monthly time series that would allow adoption trends,

format shifts, or implementation error patterns to be tracked at short horizons.

This paper reports the results of processing every Common Crawl monthly snapshot from CC-MAIN-2026-04 through CC-MAIN-2026-30, a contiguous seven-month period. Each crawl was extracted independently under a fixed analytical shape and the results were combined post-hoc. The corpus is treated as a single object of study; individual months are broken out only where they reveal a meaningful trend. The paper’s contribution is empirical rather than methodological: the goal is to describe, with concrete numbers, what schema.org looks like on the open web in the first half of 2026, and to make that description reproducible.

The remainder of the paper is organised as follows. Section 2 reviews prior structured-data measurements. Section 3 describes the extraction pipeline and the six-part analysis shape applied to every month. Section 4 presents the aggregate findings under seven headings: coverage, vocabulary shape, category composition, format evolution, implementation errors, co-occurrence templates, and Wikidata linking. Section 5 interprets the results and compares them against Google’s

dataset. Section 6 concludes.

2 Related Work

Web Data Commons. The most direct precedent for this paper is the Web Data Commons project [4], which has extracted structured data from Common Crawl on a semi-regular basis for over a decade. WDC publishes counts of entities and quads for RDFa, microdata, microformats, and (since 2014) JSON-LD. WDC’s methodological focus has historically been on producing downloadable dumps that support downstream research; its aggregate summary statistics are typically framed around markup formats and top-level types rather than the full distribution shape or implementation-error surface.

Google’s Schema.org Usage Statistics. In June 2026 Google published a public Schema.org Usage Statistics Dataset [3] that reports, for each schema.org type, the number of distinct root domains observed to use it and coarse frequency bands. Because it is derived from Google’s production crawl index, it has substantially wider domain coverage than any Common Crawl based measurement. However, it reports domain breadth rather than instance volume: a type used ten million times on a handful of sites will look smaller than a type used a handful of times on a million sites. It also does not include a time series.

Schema.org itself. The schema.org project maintains documentation of the vocabulary and a set of examples [2], but not empirical usage measurements.

Common Crawl. Common Crawl [1] is the substrate on which nearly all public open-web structured-data measurements are built. Each monthly crawl produces on the order of 2–3 billion pages packaged into 100,000 WARC files. The consistency of the crawl format across months is what makes longitudinal comparison possible.

The present paper differs from prior work in three respects. First, it operates on seven contiguous monthly crawls processed under an identical extraction pipeline, producing a directly comparable time series. Second, it publishes the full concentration distribution and long tail composition rather than only headline counts. Third, it quantifies the shadow layer of implementation errors as a first-class object rather than filtering them out.

3 Methodology

3.1 Corpus

The corpus consists of every monthly Common Crawl snapshot released in the first seven months of 2026: CC-MAIN-2026-04 (January), 08 (February), 12 (March), 17 (April), 21 (May), 25 (June), and 30 (July). Each crawl contains 100,000 WARC files. Across the seven months, 700,000 WARC files were processed successfully; 127 files in the March crawl returned S3 throttling errors and were treated as zero-page entries in per-file statistics but excluded from coverage histograms. The total number of pages examined is 14,784,536,534.

3.2 Extraction pipeline

For each crawl, WARC files were streamed from Common Crawl’s public S3 bucket into a fleet of 200 spot compute instances. Each page’s HTML was parsed with two independent extractors, one for JSON-LD blocks and one for microdata. Every extracted `@type` value was normalised to strip surrounding whitespace, prefixes such as `http://schema.org/` and `https://schema.org/`, and enclosing quotation marks. The normalised value was then written to a per-file type-counts CSV. Casing was preserved rather than folded, so that casing duplicates could be measured downstream.

Per-crawl aggregation was performed as a single reduce step over the 100,000 per-file CSVs. The pipeline emits six analyses per month: a distribution-shape analysis, a long-tail composition analysis, a category grouping, an error-pattern analysis, a per-file variance analysis, and a co-occurrence analysis. These are defined below.

3.3 Six-part analysis shape

Every month is required to produce exactly the same six analyses in exactly the same shape, so that data from different months can be combined mathematically without ambiguity. Coverage is defined as the percentage of pages carrying at least one recognised schema.org type. Density is defined as the mean number of distinct type instances per schema-bearing page.

The category grouping uses a fixed ten-category map spanning Navigation & Structure, Content & Media, People & Organizations, E-commerce, Places & Geo, Data & Properties, Events & Actions, Interaction & Social, Technology & Software, and Vehicles. The map is applied only to the top 200 types by frequency in each month, since the long tail contains too many idiosyncratic strings to categorise reliably.

Error patterns are defined as four disjoint categories: (i) casing duplicates, where the same concept appears

in more than one casing (**Website** versus **WebSite**); (ii) property-as-type errors, where a schema.org property name is used in the `@type` slot (`mainEntityOfPage`); (iii) URL-as-type errors, where a full URL, typically from a non-schema.org vocabulary, is placed in the `@type` slot; and (iv) malformed types, where the value contains whitespace, slashes, or other special characters that make it a non-parseable string.

Co-occurrence is measured on a stratified sample of 10 WARC files per month, on the schema-bearing pages within them (approximately 115,000 pages per month). For each ordered pair of types (A, B), we compute $P(B \mid A)$ and $P(A \mid B)$ by counting the pages on which each type appears. Highly correlated pairs are those where both conditionals exceed 80%; asymmetric pairs are those where $P(B \mid A) \geq 90\%$ but $P(A \mid B) < 40\%$.

3.4 Sample-based sub-analyses

Two sub-analyses in this paper are sample-based rather than full-corpus, and this is disclosed here so that their numbers are read correctly:

- The JSON-LD versus microdata format split is computed from a stratified random sample of 200 WARC files per month, evenly spaced across the 100,000 path indices, using a fixed random seed of 42. Approximately 1,000,000 pages are examined per month for the format analysis.
- The domain-level analysis (unique domains carrying schema markup) is computed from a stratified random sample of 500 WARC files per month, again with seed 42, approximately 0.5% of the corpus. Domain rankings from this sample are reliable and comparable; raw domain counts should be treated as directional.

The Wikidata linking analysis in Section 4.7 is full-corpus, run over all 100,000 WARC files of the June 2026 crawl. All other headline numbers in this paper (coverage, distribution shape, category grouping, shadow layer, per-file variance) are full-corpus.

3.5 Combination rules

Some metrics can be summed across months (total pages, total pages with schema, total occurrences of a specific `@type`, total occurrences in a specific error category). Some can only be weighted-averaged (coverage percentage, density, distinct types per WARC), with pages processed as the weight. Some are per-month only and are not combined (top-N rankings may reorder between months; category percentages may shift with content mix; co-occurrence conditional probabilities are month-specific). Every combined view reported here follows these rules.

Metric	7-month total
Pages processed	14,784,536,534
Pages with schema markup	8,620,770,647
Total type occurrences	127,990,580,002
Weighted mean coverage	58.32%
Weighted mean density	14.84
WARC files processed	700,000
Total compute cost	\$368

Table 1: Cumulative totals across the seven-month corpus (CC-MAIN-2026-04 through CC-MAIN-2026-30).

3.6 Cost and compute

Total compute across the seven months amounted to approximately 38 wall-clock hours of extraction and roughly \$368 in aggregate cost. All source data is public. Aggregated per-crawl outputs are retained in S3 buckets under the `wldm-schema-extraction-cc-2026-*` naming scheme.

4 Results

4.1 Corpus scale and coverage

Table 1 gives the cumulative totals for the seven-month corpus. Across 14,784,536,534 pages, 8,620,770,647 carry at least one schema.org type, and the parser emitted 127,990,580,002 individual type occurrences after normalisation.

Weighted mean coverage across the full corpus is 58.32%. Per-month coverage rises monotonically from a low of 55.27% in February to a high of 60.05% in July, an increase of 4.78 percentage points over the seven months. January registers 57.31%, then February dips to 55.27% before the monthly figure rises through March (57.69%), April (59.00%), May (59.19%), June (59.73%), and July (60.05%). The February dip is a real crawl-composition effect rather than a data-quality artefact; the aggregation was re-run against the raw per-file logs to confirm no error mass was hidden. The trend is shown in Figure 1.

Per-file variance is remarkably tight. Standard deviation of coverage across the 100,000 WARC files of each month is below 1.25 percentage points in every month, and below 0.75 percentage points in five of the seven months. Distinct-type counts per WARC file cluster around a mean of 509 (weighted) with standard deviation of approximately 21 in each month. This uniformity implies that the underlying phenomenon (which pages carry schema markup, and how many types they carry) does not depend strongly on which subset of the crawl is sampled.

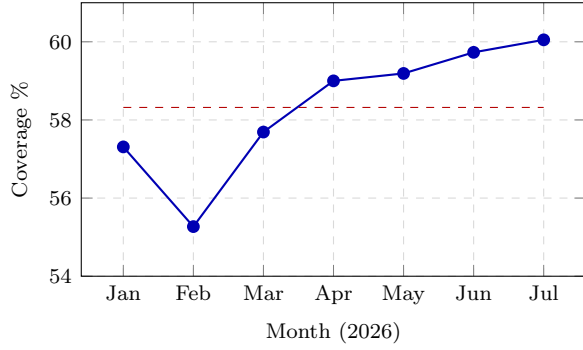


Figure 1: Monthly coverage percentage across the seven-month corpus. Dashed line marks the weighted mean (58.32%). February registers the low; July registers the high. Overall trend is monotonically upward from March onward.

Top N types	Occurrences	% of all hits
1	3,104,312,299	16.74%
5	7,877,813,237	42.47%
10	11,463,617,419	61.80%
20	14,310,372,675	77.15%
50	17,420,794,254	93.91%
100	18,154,304,308	97.87%
500	18,520,815,653	99.84%
1000	18,541,490,623	99.96%

Table 2: Cumulative concentration of type occurrences in the top N types (June 2026, CC-MAIN-2026-25). The top 100 types capture 97.87% of all occurrences.

4.2 Distribution shape

Vocabulary usage is extremely concentrated. Table 2 gives the cumulative percentage of total occurrences accounted for by the top N types, using the June 2026 crawl as a representative month; the shape is essentially unchanged in every other month.

A single type, `ListItem`, accounts for approximately one in every six occurrences in every month examined (16.67% to 17.05%). The top 10 types account for approximately 62% of occurrences, and the top 100 for approximately 98%. Types with a rank between 500 and 1000 contribute less than 0.15% cumulatively.

The long tail is correspondingly heavy. In each month, more than 62% of distinct types are singletons (appearing exactly once). In January the singleton share peaks at 64.9% (114,765 out of 176,912 distinct types). Types appearing 10 times or fewer account for approximately 85–90% of distinct types in every month, and types appearing 100 times or fewer for approximately 95%. The Zipf test (Table 3) confirms that $\text{rank} \times \text{frequency}$ is far from constant: it climbs from 3.10 billion at rank 1 to 5.17 billion at rank 10 before collapsing to 490 million by rank 100 and 13.7

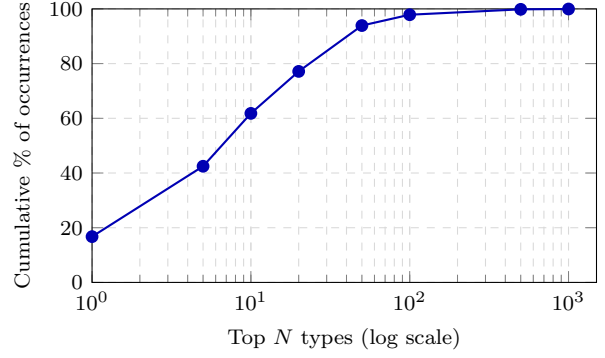


Figure 2: Cumulative concentration curve for the top N schema.org types (June 2026). The curve is steeper than a pure Zipf distribution: $\text{rank} \times \text{frequency}$ falls sharply beyond rank 100 rather than remaining constant.

Rank	Frequency	$\text{rank} \times \text{freq.}$
1	3,104,312,299	3.10×10^9
10	517,005,418	5.17×10^9
100	4,901,007	4.90×10^8
1,000	13,744	1.37×10^7
10,000	46	4.60×10^5
100,000	1	1.00×10^5

Table 3: Zipf test for June 2026. If the distribution were Zipfian, $\text{rank} \times \text{frequency}$ would be roughly constant. It is not.

million by rank 1,000. The distribution decays substantially faster than pure Zipf.

A qualitative sample of 40 random singletons per month (seed = 42) shows that the long tail is dominated by non-vocabulary noise. Approximately 60–70% of singletons are full URLs placed in the `@type` slot, typically from non-schema.org vocabularies (purl.org/goodrelations, microformats.org, w3.org/2001/XMLSchema) or from confused implementations. Approximately 10–15% are foreign-language content strings (Bengali, Russian, Japanese, Portuguese, Spanish, Georgian, Thai, Arabic, Greek, German, Vietnamese, and other languages appear in the sampled singletons). Article titles or product names, custom vocabularies (`TreePruningAndRemovalCompany`, `FinancialPlan`, `BreakfastService`), version-suffixed real types (`EntryPoint 1.20.3`), typos of real types, and complete garbage or word fragments make up the remainder. The long tail is therefore a noise floor rather than a genuine vocabulary long tail.

4.3 Category composition

Grouping the top 200 types per month into ten fixed categories yields the composition shown in Figure 3.

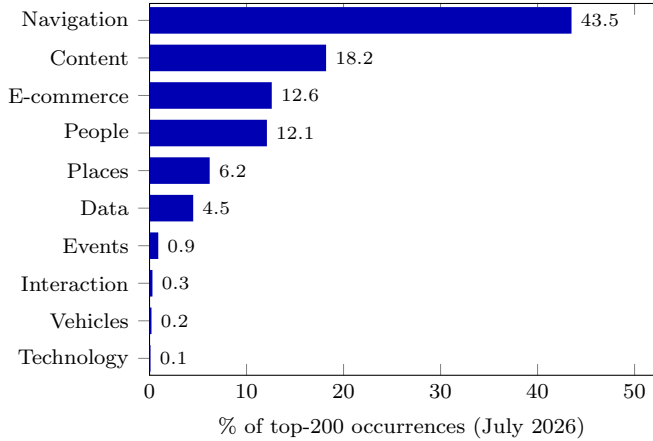


Figure 3: Category composition in July 2026 as a percentage of the top 200 types’ total occurrences. Navigation & Structure (`ListItem`, `BreadcrumbList`, `SiteNavigationElement`, `WebPage`, `WebSite`) dominates. Shape is essentially unchanged in every other month.

The picture is stable across all seven months: no category shifts by more than 3.1 percentage points across the seven-month window, and the overall shape is dominated by structural markup.

Navigation & Structure sits between 43.5% and 46.6% of top-200 occurrences across the seven months, with a slight downward trend (46.6% in January to 43.5% in July). It is populated overwhelmingly by `ListItem`, `BreadcrumbList`, `SiteNavigationElement`, `WebPage`, and `WebSite`. Content & Media sits stably at approximately 18%, driven by `ImageObject` (which alone contributes 1.6–2.0 billion occurrences per month), `Article`, and blogging types. E-commerce (`Offer`, `Product`, `Brand`, `Rating`, `Review`) rises slightly across the period, from 11.2% to 12.6%, consistent with growing commercial schema adoption; Places & Geo rises from 5.2% to 6.2% over the same period.

Events & Actions, Interaction & Social, Vehicles, and Technology & Software collectively contribute less than 1.5% of top-200 occurrences in every month. This is not because the underlying types are rare in absolute terms (Vehicles alone accounts for tens of millions of occurrences per month), but because they are dwarfed by the structural and content categories.

4.4 Format evolution: JSON-LD versus microdata

The stratified 200-WARC per-month sample shows a consistent, monotonic migration from microdata toward JSON-LD across the seven-month period (Figure 4). JSON-LD page share rises from 74.53% in January to 77.84% in July, a gain of 3.31 percentage points. Microdata page share falls from 49.40%

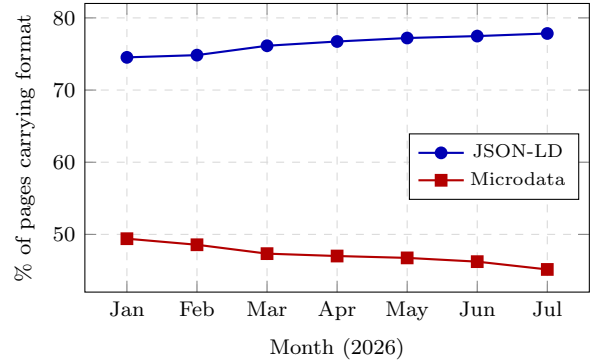


Figure 4: Format share across all seven months (stratified sample, 200 WARC files per month, seed=42). Movement is monotonic in both directions. Overlap (pages carrying both formats) remains stable at 22.98–23.95%.

Error category	7-month total	% of shadow
Property-as-type	399,653,859	49.2%
Casing duplicates	351,673,905	43.3%
URL-as-type	42,691,296	5.3%
Malformed strings	17,927,213	2.2%
Total shadow layer	811,946,273	100%

Table 4: Composition of the shadow layer of implementation errors across the seven-month corpus. Total shadow layer represents 0.63% of 127,990,580,002 total occurrences.

to 45.14% over the same period, a loss of 4.26 percentage points. Both directions are monotonic: every month shows JSON-LD ticking up and microdata ticking down relative to the prior month.

The overlap between the two, that is, the percentage of pages carrying both formats, is stable throughout at approximately 23–24% (22.98% in July, 23.95% in May, 23.93% in January). The migration is not therefore a matter of dual-format pages dropping one; it is microdata-only pages converting to JSON-LD or being replaced with JSON-LD-only equivalents.

The shift is more pronounced when measured by occurrences rather than pages. JSON-LD’s share of total type occurrences rises from 70.49% in January to 75.66% in July, meaning JSON-LD-carrying pages are becoming denser in type occurrences relative to microdata pages as well as more numerous.

4.5 Shadow layer of implementation errors

Combining the four error categories yields a stable “shadow layer” representing invalid or non-conforming markup. Table 4 presents the totals per month; Figure 5 shows the composition.

The shadow layer represents 0.63% of all type oc-

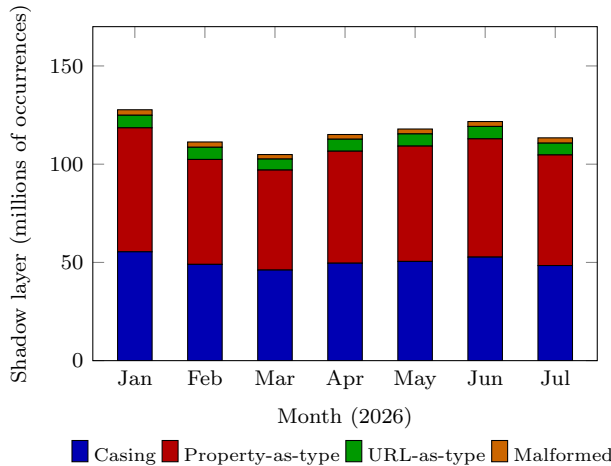


Figure 5: Composition of the shadow layer across all seven months. Values in millions of occurrences. Property-as-type and casing errors together account for 92.5% of the shadow layer.

currences across the seven-month corpus (811,946,273 out of 127,990,580,002). Per-month share ranges from 0.58% (July) to 0.67% (January), a spread of 0.09 percentage points. No month exceeds the 0.5-percentage-point variance threshold from the mean, and there is a weak but consistent downward trend consistent with improving implementation quality over time.

Property-as-type errors are the largest single category at 49.2% of the shadow layer. These are cases where a schema.org property name (`mainEntityOfPage`, `dateModified`, `itemPage`, `organization`, `offer`) is placed in the `@type` slot. Because JSON property casing conventionally begins with a lowercase letter and schema.org type casing conventionally begins with uppercase, this error is easily detectable and has a clear rewriting fix. Casing duplicates are the second largest category at 43.3%; the top offenders are `Website` (versus `WebSite`), `organization` (versus `Organization`), `offer` (versus `Offer`), and `person` (versus `Person`).

Two error patterns emerged as high-frequency in July that had not been present at that scale in earlier months: `ImageObject` (with a zero substituted for the letter O) at approximately 5.6 million occurrences, and `http://schema.org/Review` (missing the colon in the URL scheme) at approximately 202 thousand occurrences. Both appear to have propagated through a single templating library or CMS plugin update.

4.6 Co-occurrence templates

Every month yields seven stable co-occurrence templates: patterns of two or more types that reliably appear together on the same page with conditional probabilities above 87%. These are:

1. **Breadcrumb:** `BreadcrumbList` + `ListItem`. Recovered in every month at $P \geq 99.7\%$ in both directions.
2. **FAQ:** `FAQPage` + `Question` + `Answer`. Recovered in every month at $P \geq 97\%$.
3. **Product:** `Product` + `Offer`. Recovered in every month at 87–91% in both directions.
4. **Enhanced e-commerce:** `Product` + `Offer` + `OfferShippingDetails` + `MerchantReturnPolicy` + `DefinedRegion`. The shipping-details and return-policy types are asymmetric riders that almost always require the product-offer pair.
5. **Search box:** `WebSite` + `SearchAction` + `EntryPoint` + `PropertyValueSpecification`. Recovered in every month at $P \geq 99\%$.
6. **Local business:** `PostalAddress` + `GeoCoordinates` + `Organization`. Recovered in every month at 92.4–94.2%.
7. **Article:** `Article` + `Person` + `Organization` + `ImageObject`. Recovered as an asymmetric cluster in every month.

All seven templates hold across all seven months at the specified thresholds; no template fell below the 80%/80% symmetric threshold in any month for its core pair. The `DefinedRegion` ↔ `OfferShippingDetails` symmetric pair fluctuated: it dropped below 80%/80% in April, then recovered above threshold in subsequent months. The enhanced e-commerce template as a whole (which depends on the pair asymmetrically) remained intact throughout.

These templates correspond directly to Google’s stated rich-result eligibility patterns for breadcrumbs, FAQ pages, product listings, sitelinks search boxes, local business panels, and article rich cards. Their stability over seven months and 700,000 WARC files suggests that publisher behaviour is strongly shaped by the search-engine rich-result surface: when the surface prescribes a co-occurrence pattern, that pattern is what appears on the web.

4.7 Wikidata linking

A separate scan of the June 2026 corpus (CC-MAIN-2026-25) was run to measure how often schema.org markup carries a link to a Wikidata entity via the `sameAs`, `mainEntity`, `mainEntityOfPage`, or `@id` property. The full-corpus run examined 2,065,085,607 pages across 100,000 WARC files.

Metric	Value
Pages scanned	2,065,085,607
Pages with Wikidata links	5,382,130
Wikidata rate	0.261% of all pages
Unique Wikidata QIDs	1,181,408
Total QID occurrences	12,468,148

Table 5: Wikidata linking in the June 2026 corpus (full-corpus scan).

Only 0.261% of all pages, or about one in every 384 pages, carry a Wikidata reference. Among schema-bearing pages the rate is higher but still small. This finding is significant because Wikidata linking is the most direct signal that a page is not just marking up its own content, but is grounding the entities on that page in a shared, resolvable knowledge base.

QID concentration is heavy but less extreme than schema.org type concentration. The top 1 QID accounts for 1.06% of QID occurrences; the top 100 for 18.86%; the top 1,000 for 40.75%; the top 10,000 for 62.48%. The top ten most-referenced entities are dominated by media brands and consumer products: Apple Books (Q284115) with 132,632 occurrences, Apple TV (Q270285) with 122,282, Xataka (Q110313255) with 83,467, and a cluster of Fox-affiliated US television stations. This is consistent with app-store, TV listing, and media-catalog pages emitting per-item markup that references a central Wikidata entity.

QIDs are carried predominantly on the `sameAs` property (the correct usage) rather than `mainEntity` or `@id`. In the smaller June sample run of 200 files, `sameAs` accounted for 4,058 of 5,344 property-hit instances (76%). This suggests that publishers who link to Wikidata generally know what they are doing.

Note on sample versus full corpus: an initial 200-WARC stratified sample of the June crawl reported a Wikidata rate of 0.546% on JSON-LD pages, approximately double the full-corpus rate of 0.261% on all pages. The sample over-represented pages with Wikidata links. The full-corpus figure is authoritative.

5 Discussion

5.1 Comparison with Google’s Schema.org Usage Statistics

Google’s Schema.org Usage Statistics Dataset [3], released June 2026, reports the number of distinct root domains observed to use each schema.org type. Our study reports the number of instance occurrences. These are two orthogonal signals: a type used ten million times on a handful of large sites will rank high in our study and low in Google’s, while a type used a

handful of times on a million sites will rank the opposite way.

The two signals nonetheless broadly agree on which schema.org types are most prominent. Spearman rank correlation on the top 100 types (using our instance ranking as one signal and Google’s domain ranking as the other) is 0.86. Types such as `Organization`, `WebSite`, `WebPage`, `BreadcrumbList`, `ImageObject`, `Person`, and `ListItem` appear in the top tier of both rankings.

The disagreements are informative. Types associated with a small number of very high-traffic sites (for example, template-heavy `SiteNavigationElement` usage) rank higher on instance volume than on domain breadth. Types associated with wide but shallow adoption (for example, `LocalBusiness` used by tens of thousands of small local sites) rank higher on domain breadth than on instance volume. Both signals are informative and together they give a fuller picture of the vocabulary’s shape than either does alone.

Our domain-level analysis (stratified sample of 500 WARC files per month) shows a 45.8% increase in the number of distinct schema-bearing domains from January (132,525) to July (193,179), and confirms that the top 7 types by domain breadth (`ListItem`, `Organization`, `BreadcrumbList`, `ImageObject`, `Person`, `WebPage`, `WebSite`) are the same in every month and match Google’s highest domain-breadth tier exactly. From March 2026 onward, `Offer` and `Product` enter the top 10 by domain count, an e-commerce adoption signal that is visible in domain breadth before it fully registers in instance volume.

5.2 What the numbers mean

Three observations deserve emphasis.

First, schema.org has effectively achieved majority coverage of the crawlable web at the page level: 58.32% of the 14.78 billion pages examined carry markup, and the coverage is rising. The web is more machine-readable in July 2026 than it was in January 2026, and the growth is monotonic once the February crawl-composition dip is set aside.

Second, the vocabulary is heavily concentrated in structural and content types, with a shape substantially steeper than pure Zipf. The top 10 types account for roughly 62% of instances; the top 100 account for approximately 98%. The long tail is mostly noise (URLs, foreign language strings, custom vocabularies, typos), not vocabulary richness.

Third, the shadow layer of implementation errors, though small at 0.63%, is remarkably stable. Property-as-type and casing errors together consistently account for over 90% of it. A publisher-side tool that normalises casing and rejects property names in

the `@type` slot would eliminate most of the shadow layer at negligible cost.

5.3 Limitations

Two limitations should be flagged.

First, the corpus is Common Crawl, not the full web. Common Crawl has known coverage biases: it underrepresents behind-login content, some very large sites limit its access, and it is not uniform in its geographic or linguistic coverage. Our findings should be read as “schema.org markup on the crawlable open web” rather than “schema.org markup on the web”.

Second, the format split and domain-level analyses are sample-based (200 and 500 WARC files per month respectively) rather than full-corpus. This is disclosed here and in Section 3. The samples are stratified, evenly spaced, and use a fixed random seed, and per-file uniformity in the full-corpus analyses gives us confidence that sample-based figures are representative. Nonetheless the format-split percentages carry sampling uncertainty that the full-corpus figures do not.

6 Conclusion

This paper has presented a seven-month empirical study of schema.org usage across the first-half-of-2026 Common Crawl snapshots. The corpus totals 14,784,536,534 pages and 127,990,580,002 type occurrences. Coverage stands at 58.32% weighted mean, is rising monotonically month over month, and is remarkably uniform across WARC files within each crawl. Vocabulary usage is heavily concentrated: the top 100 types capture approximately 98% of all occurrences, and the distribution decays substantially faster than pure Zipf. The web is migrating from microdata toward JSON-LD at a rate of approximately 0.5 percentage points per month, monotonically. A stable shadow layer of implementation errors sits at 0.63% of all occurrences, dominated by property-as-type and casing mistakes. Seven publisher-side templates (breadcrumb, FAQ, product, enhanced e-commerce, search box, local business, article) recur in every month at conditional probabilities above 87%. Wikidata linking, though sparse at 0.261% of pages, appears where publishers are grounding named entities in a shared knowledge base.

Compared with Google’s June 2026 Schema.org Usage Statistics Dataset, which measures domain breadth rather than instance volume, our rankings show a Spearman correlation of 0.86 at the top 100. The two datasets are orthogonal and complementary rather than redundant. Full aggregated outputs are

retained in Common Crawl-aligned S3 buckets to support future reproduction and extension.

References

- [1] Common Crawl Foundation. *Common Crawl: An open repository of web crawl data*. <https://commoncrawl.org>, accessed September 2026.
- [2] Schema.org Community Group. *Schema.org: A collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet*. <https://schema.org>, accessed September 2026.
- [3] Google. *Schema.org Usage Statistics Dataset*. BigQuery public dataset, released June 2026. <https://developers.google.com/search/blog/2026/schema-org-usage-statistics>.
- [4] Web Data Commons. *Extraction reports for JSON-LD, microdata, RDFa and microformats from the Common Crawl web corpus*. <https://webdatacommons.org/structureddata/>, accessed September 2026.
- [5] Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J. Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis. *International Semantic Web Conference (ISWC)*, 2013.
- [6] Meusel, R., Petrovski, P., Bizer, C. The Web-DataCommons Microdata, RDFa and Microformat Dataset Series. *International Semantic Web Conference (ISWC)*, 2014.
- [7] Guha, R.V., Brickley, D., Macbeth, S. Schema.org: Evolution of Structured Data on the Web. *Communications of the ACM*, 59(2):44–51, 2016.
- [8] Vrandečić, D., Krötzsch, M. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.