



CENSUS · BA-D-2026-03

Structured data on the homepage: a JSON-LD census of 2,000 domains across four sectors

On 2026-07-10 the server-delivered homepage HTML of 2,000 domains — four documented frames of 500 universities, news outlets, e-commerce sites, and U.S.

DOCUMENT	BA-D-2026-03
SERIES	Census
DATE	2026
SUBJECTS	Crawlers & access; Engines & documentation
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

On 2026-07-10 the server-delivered homepage HTML of 2,000 domains — four documented frames of 500 universities, news outlets, e-commerce sites, and U.S. federal government domains, the same frames as the crawler-access census BA-D-2026-01 — was parsed for JSON-LD, Microdata, and RDFa markup. Every signal is read from the raw, unrendered HTML a non-rendering crawler receives; JSON-LD injected by client-side script is invisible by design. Among each frame's analyzable homepages, JSON-LD ranged widely: 80.6% (333 of 413) of news, 56.6% (151 of 267) of e-commerce, 33.6% (144 of 429) of universities, and 19.1% (63 of 329) of government. Where present, the dominant types are generic — WebSite, Organization, WebPage — not the entity's own kind. It measures deployment only: presence of markup, not consistency across pages or languages, nor whether marked-up facts appear in visible text. Non-response is reported separately; university bytes are reused from 2026-07-09, the rest fetched 2026-07-10.

IN BRIEF

Many websites embed a small block of machine-readable data — most often JSON-LD, following the shared schema.org vocabulary — that labels what the page is about: an organization, a website, a product. This census read the raw HTML that 2,000 real homepages sent on a single day — 500 each of universities, news outlets, online stores, and U.S. federal government sites, every list drawn from a named public source — and checked whether that labeling was present in the bytes a crawler receives before any JavaScript runs. Deployment varied sharply by sector: four in five news homepages carried JSON-LD (80.6%), against fewer than one in five government homepages (19.1%), with e-commerce (56.6%) and universities (33.6%) in between. But where the markup was present, it mostly described the page in generic terms — 'a website', 'an organization', 'a web page' — rather than declaring the specific kind of entity the site is; only 14% of university homepages labeled themselves a college or university, and only 24% of news homepages a news organization. Two cautions frame every number. First, this measures deployment, not quality: it does not check whether the labels are consistent across a site's pages, or whether the facts in the markup also appear in the visible text a reader sees — both of which a full assessment requires. Second, deploying this markup is not evidence of being more visible in AI assistants: the markup is consumed by traditional search, but the answer-engine documentation reviewed elsewhere in this series treats it as unnecessary for AI features. This is a single-day snapshot; sites change without notice, and every row is in the open dataset.

KEY FINDINGS

1. Across four documented frames of 500 domains each, JSON-LD structured data in the raw, server-delivered homepage HTML was most common in news and least in government: 80.6% (333 of 413 analyzable news homepages) carried at least one JSON-LD block, against 56.6% (151 of 267) of e-commerce, 33.6% (144 of 429) of universities, and 19.1% (63 of 329) of U.S. federal government homepages — deployment rates over each frame's analyzable domains, for homepages fetched 2026-07-10 (universities reuse 2026-07-09 bytes).
2. Where JSON-LD is present, the most common top-level types are generic site-description nodes rather than the entity's own kind: WebSite, Organization, and WebPage lead every sector, while entity-specific self-typing is far rarer — CollegeOrUniversity appears on 14.0% (60 of 429) of university homepages, NewsMediaOrganization on 24.0% (99 of 413) of news homepages, GovernmentOrganization on 3.3% (11 of 329) of government homepages, and Product on 1.1% (3 of 267) of e-commerce homepages, product markup being a property of product pages rather than the homepage.
3. Deployment is not attainment of level BL-2 in the Barkhausen Ladder (BA-C-1). BL-2 requires that key facts be expressed as JSON-LD, that the entity representation be consistent across pages and languages, and that every marked-up fact also appear in the visible text. This census measures only the first clause, on a single page — the homepage — read from raw HTML, so the deployment rates above are an upper bound on BL-2 attainment, not a measure of it: a homepage that carries JSON-LD may still fail the consistency and visible-text criteria this census does not test.
4. The census measures presence, not consumption, and no result should be read as evidence that structured data raises AI visibility. Within this series' framework, schema.org markup is characterized in BA-W-2026-03 as consumed by search for rich results but a documented negative for AI features — Google's AI-features documentation states there is no special structured data needed for AI Overviews or AI Mode, and OpenAI's crawler documentation is silent on structured data as a ranking or eligibility signal, each scoped to that vendor's own documentation as sampled 2026-07-09.
5. JSON-LD is the dominant structured-data vocabulary in every frame; the older Microdata and RDFa syntaxes are sparse and unevenly distributed — a Microdata attribute appears on 2.7–10.7% of analyzable homepages across the four sectors, and RDFa is concentrated in government (12.2%, 40 of 329) and otherwise at most 6.3%. Counting any of the three, structured data of some kind appears on 83.3% (344 of 413) of news, 58.4% (156 of 267) of e-commerce, 44.1% (189 of 429) of university, and 31.6% (104 of 329) of government analyzable homepages.
6. Rates are reported only over analyzable homepages — HTTP 200 with a non-empty body — and non-response is reported separately, because it is heavy in two frames: e-commerce yielded 267 of 500 analyzable, with 170 domains returning HTTP 403 to the homepage request, and government 329 of 500, with 88 HTTP 403 responses plus 75 transport failures (38 DNS, 26 timeout, 7 TLS, 4 other). A frame's deployment rate therefore describes its analyzable remainder, not all 500, as fetched 2026-07-10.

CONTENTS

1	1. Summary	6
2	2. Scope and method	6
3	3. Results: deployment by sector	8
4	4. What the markup declares	9
5	5. Non-response and analyzable denominators	10
6	6. What deployment does and does not mean	11
7	7. Reproducibility appendix	12
8	Limitations	12

1. Summary

On 2026-07-10, the server-delivered homepage HTML of 2,000 web domains was parsed for structured data — JSON-LD, Microdata, and RDFa. The domains form four frames of 500 each — universities, news outlets, e-commerce sites, and U.S. federal government domains — and are the identical frames used by the crawler-access census BA-D-2026-01: each frame is a top-N selection from a cited public source under a documented ordering (Tranco traffic rank [1], or reported enrollment for the U.S. university portion), so the exact list is reproducible. This is a census of those four documented frames, not a sample of any larger population: every figure below is an exact count and percentage of a fully enumerated frame, and no figure is generalized to “all universities” or “all news sites.” The unit of analysis is the presence of structured-data markup in one page — the homepage — as served on one day.

Every signal is read from the **raw, unrendered HTML** the server returned, before any JavaScript executes — the bytes a non-rendering crawler receives. This is the same measurement stance as the companion censuses BA-D-2026-02 and BA-D-2026-05, and it is a deliberate choice: JSON-LD injected by client-side script (through a tag manager, a hydration step, or a `setAttribute` call) is invisible here by design, so a “no JSON-LD” result means “no JSON-LD in the server-delivered HTML,” not “the site has no JSON-LD.” The direction of that bias is known and one-sided: for a consumer that does not render, the raw-HTML count is the operative count; for any consumer that does render, these figures undercount.

The organizing finding is a wide sector gradient in deployment. Among each frame’s analyzable homepages, JSON-LD is present on 80.6% (333 of 413) of news, 56.6% (151 of 267) of e-commerce, 33.6% (144 of 429) of universities, and 19.1% (63 of 329) of government (Figure 1). Two disciplines run through the report. First, every rate is computed only over a frame’s *analyzable* homepages — those that returned HTTP 200 with a non-empty body — and non-response is reported on its own in Section 5 rather than folded into a denominator. Second, the report is careful about what deployment does and does not establish: it measures the *presence* of markup on the homepage, not its consistency across a site’s pages or languages, not whether marked-up facts also appear in visible text, and not whether any answer engine consumes it. Section 6 states that boundary against the maturity model of BA-C-1 and the documentation review of BA-W-2026-03.

2. Scope and method

2.1 The four frames

The frames are those of BA-D-2026-01, unchanged: 2,000 rows, 2,000 distinct registrable domains (eTLD+1, computed with the Public Suffix List [2]), disjoint across the four sectors. Each is a top-500 selection from an authoritative public source imposed on a documented ordering so the selection is reproducible — the Tranco research list (list ID 46XZX) [1] as the universal traffic ordering, except for the U.S. university portion, which is ordered by reported enrollment.

- **Universities (500).** 300 U.S. institutions ordered by reported enrollment from the College Scorecard [3], plus 200 non-U.S. institutions from the Hipolabs university-domains list [4] ordered by Tranco rank.
- **News (500).** The palewire news-homepages source list [5], reduced to registrable domains, de-duplicated, intersected with Tranco, and taken in rank order. The source list is US-heavy.
- **E-commerce (500).** The “shopping” category of the UT1 web-filtering blacklists [6], restricted to bare registrable domains, intersected with Tranco, taken in rank order after a CDN/infrastructure exclusion list.

- **Government (500).** The CISA `current-federal.csv` registry of U.S. federal `.gov` domains [7], intersected with Tranco and taken in rank order. This frame is U.S.-only and federal-only.

The frames are traffic-biased by construction and combine two orderings (enrollment for U.S. universities, traffic rank elsewhere), so they are not uniform samples of their sectors; cross-sector comparisons are comparisons of like method on unlike frames. The full selection rules are in BA-D-2026-01 and in the dataset’s methodology.

2.2 What was fetched, and the raw-HTML-only stance

For each domain, `https://<domain>/` was requested once (GET, up to five redirects, 20-second timeout, user-agent Mozilla/5.0 (compatible; research-fetch), a single `http://` retry on a connect-level failure), and the exact response bytes and headers were stored. A homepage is **analyzable** when the stored response is HTTP 200 with a non-empty body; the same rule as BA-D-2026-02. Non-analyzable domains still receive a result row carrying fetch status only. One collection-window difference is load-bearing and is disclosed everywhere it matters: **university homepage bytes are reused from the 2026-07-09 crawler census** (the identical stored responses BA-D-2026-02 analyzed), while the news, e-commerce, and government homepages were **fetched fresh on 2026-07-10**. The university column therefore describes 2026-07-09; the other three describe 2026-07-10.

All structured-data signals are extracted from the raw, unrendered HTML. The research question is what a non-JavaScript-executing crawler or answer engine can read, and the crawler taxonomy BA-C-6 documents training, retrieval, and user-fetch crawlers as fetching pages rather than driving a rendering engine. JSON-LD delivered only after client-side rendering is not counted, by design; the companion census BA-D-2026-05 measures the size of that raw-versus-rendered gap directly.

2.3 What “has JSON-LD” and each type mean

JSON-LD is a JSON-based serialization for linked data; on the web it is embedded in a page as a `<script type="application/ld+json">` block whose body is a JSON document [8], and its type vocabulary — Organization, WebSite, Product, and the rest — is schema.org [9]. The census parses the raw HTML with a standard HTML parser and applies these operationalizations, each stated in full in the dataset’s methodology so a reader can reproduce or contest them:

- **has_jsonld** is true when the page carries at least one `<script>` block whose (lower-cased) type attribute contains `ld+json`. The parser decodes HTML character references in attribute values, so a type written `application/ld+json` is correctly recognized. This is a *block-presence* test; it does not require the block to be valid JSON.
- **block_count / invalid_block_count.** Each block’s inner text is parsed with `json.loads` as-is, with **no repair**. A block whose JSON does not parse — for example a CMS that HTML-escapes the quotes inside the script — is counted invalid, because that is what a conformant JSON-LD consumer reading the raw script sees. This is a deliberate lower bound on validity, not a claim that the markup is well-formed.
- **Top-level types and distinct_types.** For each valid block, the top-level `@type` values are collected — flattening a top-level array of nodes, the members of a `@graph`, and an `@type` written as an array — and normalized by stripping a schema.org URL or compact-IRI prefix while preserving case (`https://schema.org/Organization` → `Organization`). Only top-level and `@graph`-level nodes are counted, not entities nested arbitrarily deep inside another, because on a homepage the top-level nodes

are what the page declares about itself. `distinct_types` is the per-domain sorted set of these normalized names.

- **Presence flags** (`has_organization`, `has_collegeoruniversity`, and so on) are exact normalized-name matches over that top-level set: a subtype such as `BlogPosting` does not set `has_article_or_newsarticle`. The one deep-scan flag is `has_searchaction_sitelinks`, set when a `SearchAction` type appears anywhere (the Google `sitelinks-searchbox` construct, always nested under `WebSite.potentialAction`).
- **Microdata and RDFa** are recorded as **presence only**: `has_microdata` is true if an `itemscope` attribute appears on any tag, `has_rdfa` if a `typeof` attribute appears on any tag. No item is extracted and no validity is judged; these are coarse “is any of this syntax present at all” signals.

3. Results: deployment by sector

The headline measure is `has_jsonld` over each frame’s analyzable homepages. The denominators — the analyzable counts — differ by frame because non-response differs by frame (Section 5): universities 429, news 413, e-commerce 267, government 329.

SECTOR	ANALYZABLE	JSON-LD	MICRODATA	RDFA	ANY STRUCTURED DATA
News	413	80.6% (333)	10.7% (44)	1.0% (4)	83.3% (344)
E-commerce	267	56.6% (151)	7.1% (19)	0.4% (1)	58.4% (156)
Universities	429	33.6% (144)	10.3% (44)	6.3% (27)	44.1% (189)
Government	329	19.1% (63)	2.7% (9)	12.2% (40)	31.6% (104)

Three features survive excerpting. First, JSON-LD deployment spans a wide range across the frames — from four in five news homepages to fewer than one in five government homepages — and JSON-LD is the dominant structured-data syntax in every frame. Second, Microdata and RDFa are sparse and do not compensate for a low JSON-LD rate: the only place the older syntaxes are conspicuous is government RDFa (12.2%, 40 of 329), plausibly a content-management-system default rather than a deliberate entity declaration, and the census’s presence-only measurement cannot distinguish the two. Third, “any structured data” tracks JSON-LD closely everywhere, so the sector gradient is a JSON-LD gradient.

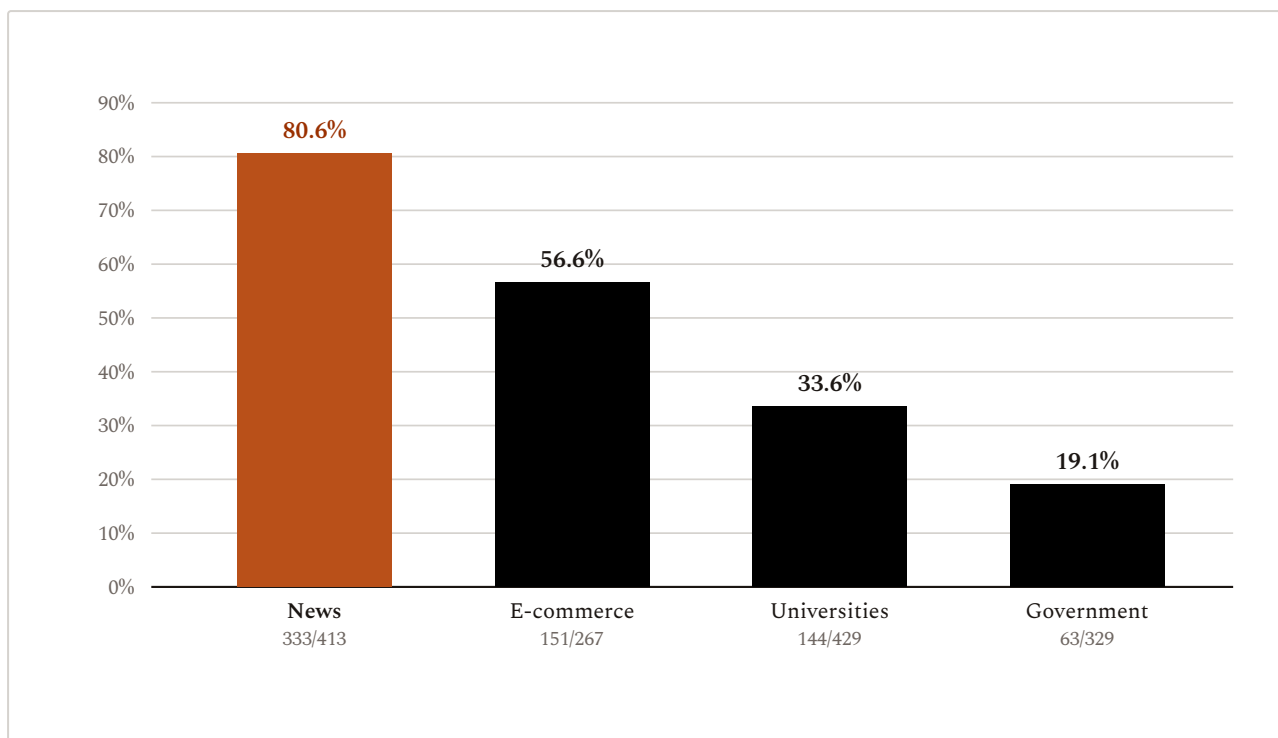


Figure 1. Share of each sector's analyzable homepages that carry at least one JSON-LD block in the raw, server-delivered HTML. Bars are exact frame counts, not estimates; denominators are each frame's analyzable homepages. University bytes are from 2026-07-09; the other three frames from 2026-07-10.

BA-D-2026-03, dataset structured-data-2026 · news n=413; e-commerce n=267; universities n=429; government n=329 · 2026-07-09 (universities) / 2026-07-10 (news, e-commerce, government)

A small validity note belongs with the deployment figures. Even where JSON-LD is present, a block occasionally fails to parse as-is: 15 domains across the four frames (universities 4, news 5, e-commerce 6, government 0) carry at least one block that does not parse with `json.loads` without repair, typically because the CMS HTML-escaped the quotes inside the script. Because the census does not entity-decode-then-retry, these are counted invalid — the state a raw JSON-LD consumer encounters — so the deployment figures count presence of a block, and a small number of those blocks are not machine-readable JSON.

4. What the markup declares

Deployment says a block is present; the types say what the page declares itself to be. The table below reports, per sector, the most common **top-level** `@type` values by the number of analyzable homepages that declare each. The dedup convention is explicit: each domain contributes each distinct normalized top-level type **at most once** (the count is domains, not blocks or occurrences), and only top-level or `@graph`-level nodes are counted, not types nested inside another entity. Percentages are over the frame's analyzable homepages.

SECTOR	MOST COMMON TOP-LEVEL TYPES (DOMAINS DECLARING, % OF ANALYZABLE)
Universities (n=429)	WebSite 76 (17.7%), CollegeOrUniversity 60 (14.0%), WebPage 57 (13.3%), Organization 51 (11.9%), BreadcrumbList 41 (9.6%), ImageObject 20 (4.7%), EducationalOrganization 11 (2.6%)
News (n=413)	WebSite 186 (45.0%), WebPage 163 (39.5%), Organization 134 (32.4%), NewsMediaOrganization 99 (24.0%), BreadcrumbList 79 (19.1%), ItemList 32 (7.7%), ImageObject 28 (6.8%)
E-commerce (n=267)	Organization 97 (36.3%), WebSite 92 (34.5%), WebPage 25 (9.4%), OnlineStore 8 (3.0%), Corporation 6 (2.2%), BreadcrumbList 6 (2.2%), ItemList 5 (1.9%), Product 3 (1.1%)
Government (n=329)	WebSite 43 (13.1%), WebPage 31 (9.4%), Organization 24 (7.3%), BreadcrumbList 19 (5.8%), GovernmentOrganization 11 (3.3%), ImageObject 11 (3.3%), Article 6 (1.8%)

The consistent pattern is that generic site-description types dominate entity-specific ones. `WebSite`, `Organization`, and `WebPage` — the nodes that describe the page *as a page and a site* — lead every frame, while the type that names *what the entity is* is rarer: `CollegeOrUniversity` on 14.0% of university homepages, `NewsMediaOrganization` on 24.0% of news homepages, `GovernmentOrganization` on 3.3% of government homepages. Two of these gaps have straightforward mechanical readings that the data supports without further inference. `Product` markup is nearly absent from e-commerce homepages (1.1%) because product markup is a property of a product detail page, not of the storefront homepage this census reads. And `NewsArticle` is rare on news homepages (declared by 7 of 413) for the same structural reason: a homepage is an index, not an article. The `SearchAction` `sitelinks-searchbox` construct — the one non-top-level signal measured — appears on 37.5% (155 of 413) of news homepages, 32.6% (87 of 267) of e-commerce, 13.3% (57 of 429) of university, and 9.4% (31 of 329) of government analyzable homepages, roughly tracking the JSON-LD gradient itself.

5 5. Non-response and analyzable denominators

Every rate above is computed over analyzable homepages, and the analyzable count is well below 500 in two frames. A domain that did not return an HTTP 200 with a body has an *undeterminable* structured-data state — a 403 challenge page or a DNS failure is neither “has JSON-LD” nor “has none” — so those domains are excluded from the deployment denominators and reported here as their own result. The composition differs by frame.

SECTOR	ANALYZABLE	NON-ANALYZABLE (OF 500)	HTTP 403	TRANSPORT FAILURES (DNS / TIMEOUT / TLS / OTHER)	WINDOW
Universities	429	71	33	38 (19 / 11 / 6 / 2)	2026-07-09
News	413	87	56	20 (1 / 18 / 0 / 1)	2026-07-10
E-commerce	267	233	170	24 (4 / 19 / 0 / 1)	2026-07-10
Government	329	171	88	75 (38 / 26 / 7 / 4)	2026-07-10

Each frame’s non-analyzable total is the sum of HTTP 403 responses, transport failures, and other non-200 responses (202, other 4xx/5xx); that last, un-tabulated component is 0 for universities, 11 for news, 8 for

government, and 39 for e-commerce — the e-commerce figure including its one HTTP 200 with a zero-byte body, which is counted 200 but not analyzable.

E-commerce and government are the two low-analyzable frames, for different reasons. E-commerce is dominated by active blocking at the HTTP layer: 170 of its 500 domains returned HTTP 403 to the homepage request — bot-protection infrastructure refusing the non-browser fetch — plus a residue of 202/4xx/5xx responses and one HTTP 200 with a zero-byte body (counted 200 but not analyzable, so the status-200 count of 268 exceeds the analyzable 267 by one). Government mixes HTTP-layer blocking (88 × 403) with a large transport-failure count (75), of which 38 were DNS failures — many federal .gov entries are registry records whose apex host carries no address, an artifact of enumerating a domain registry rather than a set of live web servers, the same apex-only-fetch pattern documented in BA-D-2026-01. The consequence for interpretation is explicit: government’s 19.1% and e-commerce’s 56.6% describe 329 and 267 analyzable homepages respectively, not 500, and the domains that answer cleanly may differ systematically from those behind a challenge wall.

6 6. What deployment does and does not mean

This census measures one thing — whether structured-data markup is present in a homepage’s raw HTML — and it is worth stating plainly what that does not establish, because the surrounding claims are easy to over-read.

Deployment is necessary, not sufficient, for BL-2. The Barkhausen Ladder (BA-C-1) places structured data at level BL-2, and its criteria are three: an entity **MUST** express its key facts as schema.org JSON-LD; its entity representation **MUST** be consistent across pages and across languages; and every fact expressed in markup **MUST** also be present in the visible, rendered text of the page. This census tests only the first criterion, and only on a single page — the homepage — read from raw HTML. It does not check whether the same entity is represented consistently on other pages or in other languages, and it does not diff marked-up facts against visible text. A homepage counted `has_jsonld` here may therefore still fall short of BL-2 on the two criteria this census does not measure, which is why the deployment rates are best read as an **upper bound** on BL-2 attainment rather than a placement at it. This is the same “necessary but not sufficient” caution BA-C-1 states directly: markup aids disambiguation and traditional search, but a fact that exists only in JSON-LD is not reliably available to an answer engine that reads the page.

Presence is not consumption, and this is not a visibility result. Nothing in this census measures whether any engine reads the markup or whether deploying it changes an entity’s AI visibility, and no figure here should be read that way. The documented position, assembled in BA-W-2026-03, is that schema.org markup is consumed by search — Google’s structured-data documentation describes using it to understand a page and to display rich results [10] — but is a documented negative on the answer-engine side: Google’s own AI-features documentation states there is no special structured data needed to appear in AI Overviews or AI Mode [11], and OpenAI’s crawler documentation, sampled 2026-07-09, does not list structured data among the signals its crawlers use [12]. That is a documented negative for one vendor’s AI features and a verified silence for another, each scoped to that vendor’s documentation — not a general finding that answer engines ignore structured data, and equally not any evidence that deploying it raises AI visibility. This census adds a deployment baseline to that documentation review; it does not revise its direction.

7 Reproducibility appendix

The census is reproducible offline from stored bytes. **Environment:** Python 3.13, httpx 0.28.1 (HTTP/1.1 only), and the standard-library `html.parser` and `json` — no third-party HTML or JSON-LD library, so the parse is the plain one a minimal consumer would run. **Frames:** identical to BA-D-2026-01 (Section 2.1); the domain key (`sector`, `domain`) matches that census row-for-row. **Fetch:** GET `https://<domain>/`, up to five redirects, 20-second timeout, user-agent `Mozilla/5.0 (compatible; research-fetch)`, one `http://` retry on a connect-level failure, concurrency 48; exact response bytes and headers stored.

University homepage bytes are the stored 2026-07-09 responses reused unchanged; news, e-commerce, and government were fetched 2026-07-10. **Analysis** (`jsonld_census.py`) performs no network I/O: it re-derives every field from the stored bytes, reading HTTP status and content-type from the stored headers as the source of truth. `has_jsonld` is `block-count > 0` for `<script>` blocks whose lower-cased type contains `ld+json`; block validity is `json.loads` as-is with no repair; top-level `@type` collection flattens top-level arrays, `@graph` members, and `@type` arrays, normalizes by stripping the `schema.org` prefix while preserving case, and matches presence flags as exact normalized names; `SearchAction` is the one deep-scan flag; `Microdata` (`itemscope`) and `RDFa` (`typeof`) are presence booleans. Every operationalization, including the raw-HTML-only stance and the deliberate validity lower bound, is recorded in the dataset’s methodology and in the collection’s `LIMITATIONS-2.md`.

8 Limitations

Frame, not population. Every figure is an exact count over a documented frame of 500 domains, not an estimate for a sector. The frames are traffic-biased top-N selections from specific public sources and combine two orderings (enrollment for U.S. universities, traffic rank elsewhere), so no figure here should be read as “X% of universities” or “X% of news sites.” Cross-sector differences are differences between these particular frames built by like method, not between representative populations.

Raw HTML only. All signals are read from the server-delivered HTML before any JavaScript executes. JSON-LD, Microdata, or RDFa injected by client-side rendering — through a tag manager, a hydration payload, or a `setAttribute/dangerouslySetInnerHTML` call — is invisible here by design. A “no JSON-LD” result means “none in the raw HTML,” which is exactly what a non-rendering crawler sees, and the direction of the resulting undercount is known and one-sided. BA-D-2026-05 measures the size of the raw-versus-rendered gap directly.

Homepage only, and deployment only. This census reads one page per domain — the homepage — and measures the presence of markup, not its quality. It does not assess entity consistency across a site’s pages or languages, does not diff marked-up facts against visible text, and does not validate the markup against any `schema.org` shape (beyond recording whether each JSON-LD block parses as JSON at all). The BL-2 criteria that this census does not test are stated in Section 6; the deployment rates are an upper bound on BL-2 attainment.

Two collection windows. University homepage bytes are reused from 2026-07-09; news, e-commerce, and government were fetched 2026-07-10. The two days are adjacent, but a cross-sector comparison that puts the university column beside the other three is comparing a 2026-07-09 snapshot to a 2026-07-10 one. Homepages change without notice; every figure should be assumed perishable and checked against a current fetch before it is relied upon.

Undeterminable is not absent. Non-response is uneven and heavy in two frames — e-commerce yielded 267 of 500 analyzable, government 329 of 500 — so those frames’ deployment rates rest on analyzable

remainders that may differ systematically from the domains behind a 403 or a DNS failure. A structured-data state that could not be read is missing data with a plausibly non-random mechanism, not an absence of markup, and the rates should be read with that selection in mind.

Presence measured, not consumption. The census records what is deployed, not what any engine does with it. It makes no visibility measurement and no engine-behavior claim; the documented state of structured-data consumption is reviewed in BA-W-2026-03 and is summarized, not extended, here.

REFERENCES

1. V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, Proceedings of the Network and Distributed System Security Symposium (NDSS). *Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation* (2019). <https://doi.org/10.14722/ndss.2019.23386> Accessed 2026-07-09.
2. Public Suffix List (Mozilla Foundation). *Public Suffix List (public_suffix_list.dat)* (2026). https://publicsuffix.org/list/public_suffix_list.dat Accessed 2026-07-09. [archived]
3. U.S. Department of Education, College Scorecard. *College Scorecard institution-level data (public API)* (2026). <https://collegescorecard.ed.gov/data/> Accessed 2026-07-09. [archived]
4. Hipolabs. *university-domains-list (world_universities_and_domains.json)* (2026). https://raw.githubusercontent.com/Hipo/university-domains-list/master/world_universities_and_domains.json Accessed 2026-07-09. [archived]
5. B. Welsh, palewire, news-homepages project. *news-homepages source list (sites.csv)* (2026). <https://raw.githubusercontent.com/palewire/news-homepages/main/newshomepages/sources/sites.csv> Accessed 2026-07-09. [archived]
6. F. Prigent, Université Toulouse 1 Capitole, web-filtering blacklists. *UT1 blacklists, shopping category (shopping.tar.gz)* (2026). <https://dsi.ut-capitole.fr/blacklists/download/shopping.tar.gz> Accessed 2026-07-09. [archived]
7. Cybersecurity and Infrastructure Security Agency (CISA), dotgov-data. *Federal .gov domain registry (current-federal.csv)* (2026). <https://raw.githubusercontent.com/cisagov/dotgov-data/main/current-federal.csv> Accessed 2026-07-09. [archived]
8. W3C (JSON-LD Working Group). *JSON-LD 1.1: A JSON-based Serialization for Linked Data (W3C Recommendation) — §Embedding JSON-LD in HTML Documents* (2020). <https://www.w3.org/TR/json-ld11/> Accessed 2026-07-10. [archived]
9. Schema.org (W3C Schema.org Community Group). *Schema.org — shared vocabulary for structured data on the web* (2026). <https://schema.org/> Accessed 2026-07-10. [archived]
10. Google Search Central (Google for Developers documentation). *Introduction to structured data markup in Google Search* (2025). <https://developers.google.com/search/docs/appearance/structured-data/intro-structured-data> Accessed 2026-07-09. [archived]
11. Google Search Central (Google for Developers documentation). *AI features and your website* (2025). <https://developers.google.com/search/docs/appearance/ai-features> Accessed 2026-07-09. [archived]
12. OpenAI (developer documentation). *Overview of OpenAI Crawlers* (2026). <https://developers.openai.com/api/docs/bots> Accessed 2026-07-09. [archived]

HOW TO CITE

Barkhausen AI (2026). Structured data on the homepage: a JSON-LD census of 2,000 domains across four sectors. <https://barkhausen.ai/research/structured-data-census-2026/>

Published under the Creative Commons Attribution 4.0 International (CC-BY-4.0).

