Barkhausen AI

CENSUS · BA-D-2026-04

Who is in the corpus pipeline's front door: a Common Crawl coverage census of 2,000 domains

On 2026-07-10 the Common Crawl index for CC-MAIN-2026-25 — the June 2026 monthly crawl — was queried for 2,000 domains: the four documented frames of 500 universities, news outlets, e-commerce sites, and U.S.

2026 · CC-BY-4.0

“Every leap begins as a crackle.”

DOCUMENT	BA-D-2026-04
SERIES	Census
DATE	2026
SUBJECTS	Corpora & training data; Crawlers & access
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

On 2026-07-10 the Common Crawl index for CC-MAIN-2026-25 — the June 2026 monthly crawl — was queried for 2,000 domains: the four documented frames of 500 universities, news outlets, e-commerce sites, and U.S. federal government domains from crawler-access census BA-D-2026-01. Coverage is reported two ways, because 'is a domain in Common Crawl' has two operationalizations that disagree. Captured — at least one archived record with HTTP 200 or a revisit — holds for 89.0% (1,781 of 2,000); presence — at least one index record of any status — for 95.3% (1,907). Universities are wall-to-wall (100%); news is lowest (82.8% captured, 69 fully absent). Joined to the same domains' robots.txt policy, 354 that root-block CCBot still appear in the June crawl — a timing relationship, not a compliance finding. Presence at this pipeline's front door implies nothing about a model retaining or reproducing the content.

IN BRIEF

Common Crawl is a free, roughly monthly archive of the public web, and it is the shared raw material behind most of the open datasets that large language models are trained on — the front door of the training-data pipeline. This census asked a simple question of one monthly crawl: for 2,000 real websites — 500 each of universities, news outlets, online stores, and U.S. federal government sites, the same lists a companion census used to read robots.txt — is the site in that crawl at all? The answer depends on what 'in' means, and the two reasonable definitions disagree. If 'in' means the archive holds a fetchable page (an HTTP 200), 89.0% of the sites qualify; if 'in' means the site appears in the index in any form — including as a redirect or a blocked-page record — 95.3% do. Universities are in essentially without exception; news sites least often, with 69 of 500 absent entirely. The census also lines the crawl up against each site's stated crawler policy: 354 sites that told Common Crawl's crawler to stay out still appear in the June crawl. That is almost certainly a matter of timing — the policy was read weeks after the crawl ran — and is reported as a timing relationship, not as evidence any crawler ignored the rules. One boundary is stated throughout and matters more than any single number: being in this archive is not the same as being remembered by a model. Presence in a training corpus does not imply a model retains or reproduces the content, and this census makes no such claim.

KEY FINDINGS

1. Coverage of Common Crawl has two defensible operationalizations that give different rates, and the census reports both. Captured — a domain has at least one archived record with HTTP 200 or a warc/revisit mime in CC-MAIN-2026-25 — holds for 89.0% (1,781 of 2,000) of the four frames; presence — at least one index record of any status, including redirect-only and blocked-page records — holds for 95.3% (1,907 of 2,000). The 126-domain gap between them is the report's organizing distinction, not a discrepancy to resolve.

2. The two rates split sharply by sector. Universities are wall-to-wall — 500 of 500 both captured and present (100%). News is lowest on both: 82.8% captured (414 of 500) and 86.2% present (431 of 500), with 69 domains fully absent from the crawl, the most of any frame. E-commerce and government are almost always present (99.0% and 96.2%) but carry a lower captured rate (91.4% and 82.0%), because many appear only as blocked or redirect records.

3. The present-but-no-200 gap is 126 domains (government 71, e-commerce 38, news 17) — present in the index yet with no archived HTTP 200 or revisit. This is a definitive verdict, not scan truncation: 125 of the 126 were scanned to completion. The dominant per-domain status explains it — 58 are served the crawler an HTTP 403 (blocked at the door) and 46 are apex redirects only (HTTP 301), the remainder 302/404/405/202 and a handful of throttle and upstream-error records.

4. Joined to the same domains' 2026-07-09 robots.txt policy from BA-D-2026-01, 354 domains root-block Common Crawl's CCBot yet 301 of them (85.0%) appear in the June 2026 crawl and 293 (82.8%) hold an archived HTTP 200. The robots snapshot was taken weeks after the crawl window, so this is a timing relationship — consistent with the block postdating the crawl — and is not evidence that CCBot disregarded any robots.txt rule; no compliance claim is made.

5. The census was collected by reading the Common Crawl index files directly over HTTP Range requests to data.commoncrawl.org, after the CDX query server index.commoncrawl.org refused the collecting host for over five hours. The two methods were cross-checked on 33 universities: presence agreed 30 of 30 with zero contradictions, and the range path additionally resolved 3 domains the CDX path had failed on, so it is strictly more complete on the overlap.

6. Common Crawl is characterized throughout as the front door of the open-corpus pipeline, not as training itself, reusing the verified chain from the primer BA-W-2026-02: for the one large model with a fully disclosed data composition Common Crawl supplied about 82% of the assembled dataset by token count, a 2024 audit found 64% of 47 surveyed models used at least one Common Crawl–derived corpus, and the most modern openly documented corpus is drawn entirely from it. Presence in that pipeline does not imply a model retains or reproduces the content; no causal claim is made.

CONTENTS

1 1. Summary

On 2026-07-10, the Common Crawl index for **CC-MAIN-2026-25** — the June 2026 monthly crawl [11] — was queried for the presence of 2,000 web domains. 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 [2], 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 over a fully enumerated frame, and no figure is generalized to "all universities" or "all news sites." The unit of analysis is the registered domain — the apex host and all its subdomains — as it appears in one monthly crawl.

The organizing decision is that "is a domain in Common Crawl" has **two operationalizations**, and they disagree. **Captured** means the crawl holds at least one archived, fetchable record for the domain — a capture with HTTP status 200 or a `warc/revisit` mime. **Presence** means the domain appears in the index at all — at least one record of *any* status, including a record that is only a redirect or a blocked-page response. Captured is the stricter reading (the archive has content); presence is the looser (the domain is in the index). Across the 2,000 domains, captured holds for 89.0% (1,781) and presence for 95.3% (1,907) — a 126-domain gap that is the subject of Section 4. Both are reported, side by side, throughout; neither is the "true" rate, and which one a claim should cite is a measurement choice that has to be disclosed, the same discipline the minimum-disclosure convention BA-C-4 requires of any rate and that the note BA-DI-3 draws out for a different two-rate signal.

Two further disciplines run through the report. First, the honesty boundary is stated at the outset and repeated in Section 6: Common Crawl is the *front door* of the open-corpus pipeline, not training itself, and **presence in a training corpus does not imply the model retains or reproduces the content; no causal claim is made.** Second, the census joins its coverage result to the same domains' robots.txt policy from BA-D-2026-01 (Section 5), and does so with explicit time semantics — the crawl ran in June 2026, the robots snapshot was taken on 2026-07-09 — so a domain that blocks CCBot today appearing in last month's crawl is reported as a timing relationship, never as a compliance finding.

2 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 [3]), 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 [2] 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 [4], plus 200 non-U.S. institutions from the Hipolabs university-domains list [5] ordered by Tranco rank.
- **News** (**500**). The palewire news-homepages source list [6], 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 [7], 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 [8], 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, 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. Because both censuses load the same frame CSVs through the same normalizer, the `(sector, domain)` key of this census matches BA-D-2026-01 row-for-row, which is what makes the join in Section 5 exact.

2.2 The crawl, and reading its index by SURT range

Common Crawl publishes, for each monthly crawl, a URL index in the form of ~300 SURT-sorted gzip shards (`cdx-00000.gz … cdx-002NN.gz`) plus a single plaintext second-level index, `cluster.idx`, that gives — one line per ~190 KB gzip block — that block's first SURT key and its `(shard, byte offset, byte length)` [10]. The whole set lives under `data.commoncrawl.org/cc-index/collections/CC-MAIN-2026-25/indexes/`, and Common Crawl documents that a client can read an individual capture by taking the `filename`, `offset`, and `length` from the index and issuing an HTTP Range request against the bucket [9][10]. This census does exactly that at the domain level: for each domain it binary-searches `cluster.idx` for the blocks whose SURT range covers the domain, Range-fetches just those blocks (a few hundred kilobytes each), gunzips them, and scans the CDX lines inside. The `cluster.idx` for this crawl (101,131,403 bytes, 852,779 blocks) was downloaded once and cached; block bodies were not persisted, and each result row records the exact shards it read so any row is replayable.

A method switch is disclosed rather than hidden, because it shaped the collection. The census was first attempted against the CDX **query server** `index.commoncrawl.org`, which answers domain queries directly. That endpoint refused the collecting host's IP for over five hours — a connection-level cooldown, not an HTTP error — so no query completed. `data.commoncrawl.org`, the storage front end that serves the raw index *files*, answered Range requests normally, so the census was re-collected by reading the index files directly. It is the same underlying index, reached through a different door; the query server was not touched again. The two approaches were cross-checked on the 33 universities the query-server path resolved before it failed (Section 7): presence agreed on 30 of 30 with zero contradictions, and the file-reading path additionally resolved 3 domains the query path had failed on, so on the overlap it is strictly more complete.

The correctness of the range is a SURT-ordering argument, stated here in brief and in full in the dataset's methodology. A capture's index key is its host label-reversed and comma-joined, with `)` closing the host before the path — `example.com/` becomes `com,example)/`, and `sub.example.com/` becomes `com,example,sub)/`. Matching a registered domain and all its subdomains — the semantics the CDX query server calls `matchType=domain` — is therefore exactly the set of keys with prefix `com,example)` (the apex and any path) or `com,example,` (any subdomain). Because `)` sorts immediately before `,`, that set is one contiguous run bounded below by `com,example)` and above by `com,example-` (the character after `)`), and the two bounds are tight: the next distinct registered domain, `example2.com` → `com,example2)`, sorts at or after the upper bound and is correctly excluded. The one known narrowing is that captures on non-standard ports serialize with a `:port` that sorts outside the range and are not counted; ports 80 and 443 are stripped during index canonicalization, so ordinary http and https captures are counted, and port records never decide whether a domain is captured. The collector's self-test asserts the SURT construction and its boundary cases offline before any network call.

2.3 Captured and presence, defined precisely

The two coverage fields are computed over the records scanned for each domain, and they mean exactly this:

- **`captured`** is true when at least one matching record has HTTP status `200` or the mime `warc/revisit`. This is the census's headline measure: the crawl holds archived, fetchable content for the domain. A `warc/revisit` record is Common Crawl's deduplicated re-observation of content identical to an earlier capture and is counted as archived content, following the crawl's own record model.
- **`captured_any_presence`** (reported as *presence*) is true when at least one matching record exists at all, regardless of status. A domain that appears only as a `301` redirect record, or only as a `403` blocked-page record, is *present* but not *captured*. This field equals the "the domain is in the index" semantics of the query server's block count, and it is the field on which the two collection paths were cross-checked.

A domain that is neither captured nor present — no record of any status in the crawl — is **fully absent**, and those are reported on their own. The two fields are nested by construction (`captured` implies `captured_any_presence`), so the difference between the two rates is precisely the *present-but-not-captured* set analyzed in Section 4.

3 3. Results: coverage by sector

The headline is the two coverage rates over each frame's 500 domains. Universities are wall-to-wall; news is lowest on both measures; e-commerce and government are almost always present but carry a lower captured rate.

SECTOR	N	CAPTURED (200/REVISIT)	PRESENCE (ANY RECORD)	FULLY ABSENT
Universities	500	100.0% (500)	100.0% (500)	0
News	500	82.8% (414)	86.2% (431)	69
E-commerce	500	91.4% (457)	99.0% (495)	5
Government	500	82.0% (410)	96.2% (481)	19
Total	**2000**	**89.0% (1,781)**	**95.3% (1,907)**	**93**

Three features survive excerpting. First, the choice of operationalization moves the headline by more than six points overall — 89.0% versus 95.3% — and by far more within a sector: government reads 82.0% captured but 96.2% present, a 14-point spread, because a large share of federal domains appear in the index only as blocked or redirect records rather than as archived pages. Second, universities are the ceiling on both measures (500 of 500), consistent with their being large, openly crawlable, heavily linked `.edu`/academic hosts. Third, news is the floor on both and carries the most fully-absent domains (69 of 500): many smaller, regional, or hard-paywalled titles are simply not in this crawl at all — an absence that is upstream of any robots.txt or blocking behavior, since a domain the crawler never reached leaves no record to block.

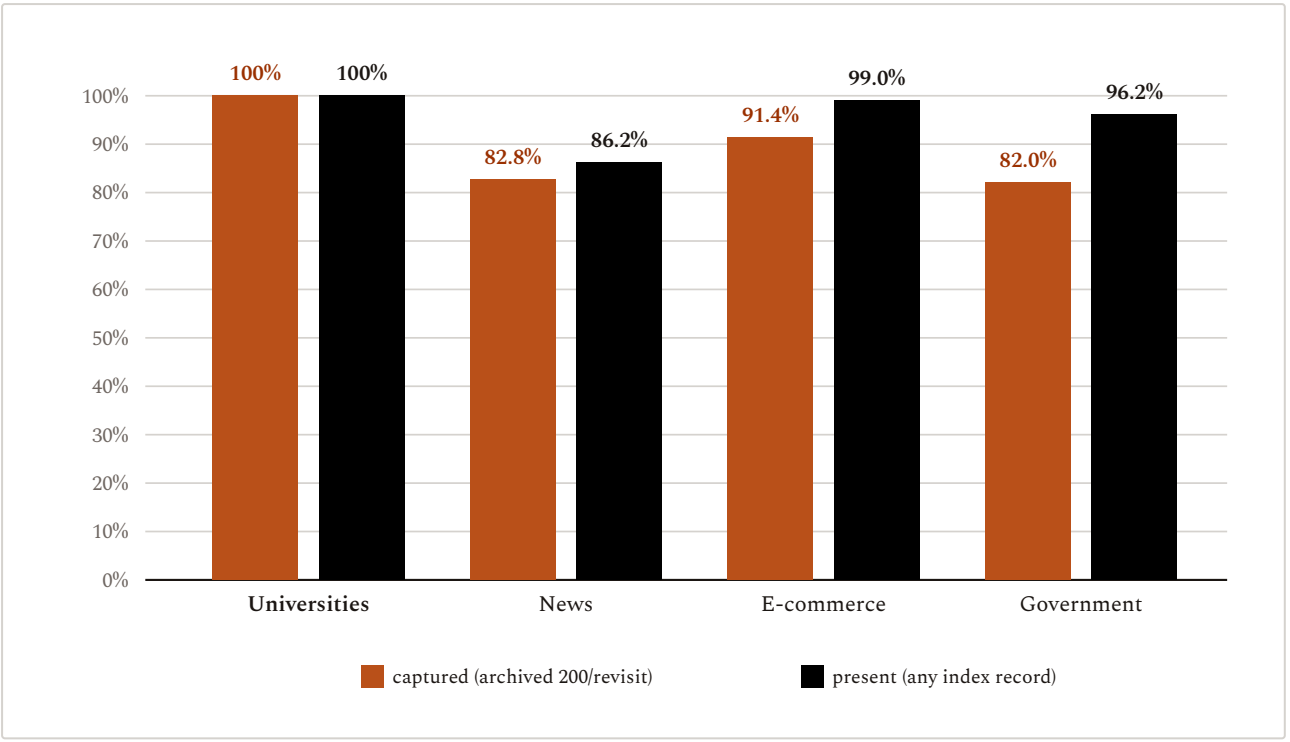

Figure 1. Two coverage rates for each frame's 500 domains in CC-MAIN-2026-25: captured (at least one archived HTTP 200 or revisit record) and presence (at least one index record of any status). Bars are exact frame counts, not estimates; the two measures are nested, so captured is always at or below presence. Universities reach 100% on both.

BA-D-2026-04, dataset cc-coverage-2026 · 500 domains per frame; 2,000 total · CC-MAIN-2026-25 (June 2026 crawl)

4 4. The present-but-no-200 gap

The 126 domains that are present but not captured are the difference between the report's two rates, and they are a real signal rather than an artifact of incomplete scanning. The scan reads a domain's index blocks in order and stops early only once both reported metrics are locked; for this set the verdict is definitive because **125 of the 126 were scanned to completion** — their whole index run fit within the collector's per-domain block budget — so "no archived 200" is a fact about the crawl, not a scan that gave up. The single exception (ozon.ru) spans more blocks than were scanned and is a best-effort no-200; it is flagged as such in the dataset. The gap is concentrated by sector and explained by the dominant per-domain HTTP status.

DOMINANT STATUS AMONG A DOMAIN'S RECORDS	DOMAINS	WHAT IT MEANS
403	58	The crawler is refused at the door (a bot wall or a robots/WAF block)
301	46	The apex is a permanent redirect only (domain consolidation, or http→canonical elsewhere)
302	8	A temporary redirect only
404 / 405 / 410	5	Not-found / method-not-allowed / gone records only
202	3	An async anti-bot accept (e.g. a site that returns 202 across its pages)
429 / 502 / 503 / 522 / 307	6	Throttled, upstream error, or a redirect captured at crawl time

By sector the gap is government 71, e-commerce 38, and news 17 — universities contribute none. The two mechanisms are the two dominant rows: 58 domains where the crawler met an HTTP 403, and 46 where the apex served only a redirect. Both are consistent with what the companion robots census found for the same frames — government and e-commerce carry the heaviest HTTP-layer blocking, and many federal `.gov` entries are registry records whose apex is a redirect or carries no live page. The interpretive consequence is the reason the report keeps both rates: for these 126 domains, "in Common Crawl" is true under *presence* and false under *captured*, and a single headline that picked one would either overstate archived coverage or understate index presence.

5 5. Common Crawl coverage against CCBot policy

The census key matches BA-D-2026-01 exactly, so each domain's coverage in the June crawl can be placed beside its own robots.txt policy toward **CCBot**, Common Crawl's crawler [12], as that policy stood on 2026-07-09. CCBot's `root_blocked` state — whether the file's rules, under the Robots Exclusion Protocol [1], forbid it the site root — is determinable only for a domain with a readable robots.txt (the `parsed` or `no_robots_file` categories of BA-D-2026-01); where the file was a 403 or a challenge page, the policy is unknowable and the domain is held out of the crosstab. That partition reproduces BA-D-2026-01's knowable denominators exactly (news 432, universities 423, e-commerce 311, government 297), a useful confirmation that the join is aligned.

The load-bearing caution is temporal and is stated before the numbers. **CC-MAIN-2026-25 was crawled in June 2026** — the index timestamps observed while collecting fall in mid-June — while the robots.txt policy was read on **2026-07-09**, three to four weeks *after* the crawl window. A domain that root-blocks CCBot on 2026-07-09 and nonetheless appears in the June crawl is therefore consistent with several innocent readings: the block may have been added after the crawl ran; the operator may block future training crawls while leaving past archived captures in place; or the rule may have changed in the interval. This census measures a crawl and a policy at two different times and reports their *relationship*; it is not a test of whether CCBot honored any robots.txt file, and nothing here should be read as a compliance or non-compliance finding.

With that frame, the crosstab reports, for each sector's robots-determinable domains, how many root-block CCBot and how many of those blockers are nonetheless in the June crawl.

SECTOR	DETERMINABLE ROBOTS	CCBOT ROOT-BLOCKED	…OF BLOCKED, CC-PRESENT	…OF BLOCKED, CC-CAPTURED
News	432	250	197 (78.8%)	192 (76.8%)
E-commerce	311	66	66 (100.0%)	64 (97.0%)
Government	297	19	19 (100.0%)	18 (94.7%)
Universities	423	19	19 (100.0%)	19 (100.0%)
Total	**1,463**	**354**	**301 (85.0%)**	**293 (82.8%)**

Read with the timing caveat, two patterns stand out. First, blocking CCBot and being in the crawl are not mutually exclusive: of the 354 determinable domains that root-block CCBot, 301 (85.0%) still appear in the June index and 293 (82.8%) hold an archived 200 — because, most plausibly, the block is more recent than the crawl, or governs a future the archive does not reach back into. Second, the complementary group is even more uniformly present: among the 1,109 determinable domains that *allow* CCBot, 99.6% (1,105) are

present and 96.5% (1,070) are captured, a higher captured rate than the blocked group's 82.8%. The direction of that difference is what one would expect if blocking has any forward effect, but this census cannot attribute it — the blocked frame is dominated by news, whose absence and redirect-only rates are elevated for reasons (paywalls, consolidation) unrelated to CCBot policy, so the gap between 82.8% and 96.5% mixes any blocking effect with frame composition and cannot be decomposed here.

Universities make the timing point cleanly: only 19 of 423 determinable university domains block CCBot at all, all 19 are captured, and the frame is 100% present — a sector that overwhelmingly leaves the crawler alone and is, correspondingly, wall-to-wall in the corpus front door. News is the opposite corner: 250 of 432 block CCBot, the highest rate of any frame, yet 197 of those blockers are still in last month's crawl. Both observations are relationships between a June crawl and a July policy, not claims about the crawler's conduct.

6 6. What presence in the index does and does not mean

This census measures coverage of one monthly crawl. It is worth stating plainly what that does and does not establish, because Common Crawl sits at a point in the training-data pipeline that invites over-reading.

Common Crawl is the front door of the open-corpus pipeline, not training. The reason coverage of a single web archive is worth measuring at all is that most open training corpora are not independent; they descend from this common upstream. The verified chain assembled in the primer BA-W-2026-02 is reused here without extension: for the one large model with a fully disclosed data composition, Common Crawl supplied about 82% of the assembled training dataset by token count and about 60% of the tokens actually sampled during training [13]; a 2024 audit found that 64% of the 47 models it examined had used at least one Common Crawl–derived corpus in pretraining [14]; and the most modern openly documented corpus is drawn *entirely* from Common Crawl, built from dozens of snapshots totaling on the order of 15 trillion tokens [15]. The practical reading is not that any one model is knowable from this census, but that the pipeline has a single front door: what is absent from Common Crawl can be absent from many downstream corpora at once, and what is present is at least *eligible* to reach many models over time. Coverage here is presence at that front door — one crawl of it — and nothing further along the chain.

Presence is not memory, and no causal claim is made. Between "the crawl archived this domain" and "a model reproduces a fact about this entity" lie corpus filtering, deduplication, sampling, training, and the model's own generalization — none of which this census observes. **Presence in a training corpus does not imply the model retains or reproduces the content; no causal claim is made.** Corpus entry is a demonstrated upstream mechanism; a model's memory of a specific fact is an unproven downstream step. The correct reading of a captured result is that the domain cleared the pipeline's first gate for this crawl — not that it is remembered, and not that its absence guarantees it is forgotten. This is the same boundary the primer draws for the parametric path, and this census neither strengthens nor weakens it.

Coverage is not a visibility metric. Nothing here measures whether an assistant names an entity, cites it, or retrieves it; those are the visibility metrics defined in the conventions and estimated by sampling, not by index coverage. A domain's presence or absence in Common Crawl is a property of the corpus pipeline, one input among many to one of the two paths by which an entity becomes known, and it is reported as exactly that.

7 7. Reproducibility appendix

The census is reproducible from public data. **Environment:** Python 3.13 with `httpx` 0.28.1 (HTTP/1.1) reading the Common Crawl index files directly; no CDX query server is involved. **Crawl:** CC-MAIN-2026-25, pinned in every result row [11]. **Index:** the crawl's `cluster.idx` (101,131,403 bytes, 852,779 blocks) is downloaded once from `data.commoncrawl.org/cc-index/collections/CC-MAIN-2026-25/indexes/cluster.idx` and cached; a re-run with the cache present makes no network call for the index [9][10]. **Per domain:** the registered host is reversed to a SURT key, `cluster.idx` is binary-searched for the blocks covering the half-open range `[host), host-)` (apex, all subdomains, all paths on standard ports — the `matchType=domain` equivalence proven in the dataset methodology), those blocks are Range-fetched and gunzipped, and their CDX lines are scanned; at most eight blocks per domain are read, with an early stop once the scanned count has reached its 1,000-record cap and a 200 has been seen. **Fields:** `captured` is set by any 200/revisit record; `captured_any_presence` by any record; `count` is the matching records scanned, capped at 1,000, so it is a coverage-scale proxy and not a full capture count; `count_200_revisit` is the 200/revisit tally among scanned records and is not capped, so it can exceed `count`; `status_counts` is the full HTTP-status histogram that explains every non-captured domain; each row records the exact shards read. **Cross-check:** a retained 33-domain sample collected via the CDX query server before it failed agrees on presence 30 of 30 with zero contradictions and identical capped counts, and the range path additionally resolves the 3 the query path failed on. Every derived field is recomputable from the crawl and the recorded shards; the collector's self-test asserts the SURT construction and its boundary cases offline. The full method-of-record, including the query-server throttling behavior and the SURT-range proof, is in the collection's limitations document and the dataset methodology.

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.

One crawl, one snapshot. This is a single monthly crawl, CC-MAIN-2026-25. Common Crawl's coverage of a domain varies from crawl to crawl — a domain absent here may be captured in an adjacent month, and vice versa — so a coverage result is specific to this crawl and should not be read as "Common Crawl has never archived this domain." The result is perishable in the same way a single crawl is a single sample of the crawler's reach.

Registered domain is the unit, and coverage is not depth. Coverage is measured at the registered-domain level (apex plus all subdomains); it is a presence measure, not a measure of how much of a site is archived. The `count` field is capped at 1,000 records and is a coarse scale proxy, not a true capture count, and `count_200_revisit` — though uncapped and able to exceed `count` — is a tally over the records scanned within the per-domain block budget, not a site-wide capture total. No figure here should be read as "how many pages of this domain are in Common Crawl."

Time semantics of the CCBot join. The crosstab in Section 5 joins a June 2026 crawl to a 2026-07-09 robots.txt snapshot. The two were observed weeks apart, and the join reports their relationship, not causation: a domain that blocks CCBot in July appearing in the June crawl is consistent with the block postdating the crawl and is not evidence that CCBot ignored any rule. The captured-rate difference

between CCBot-blocked and CCBot-allowed domains mixes any blocking effect with frame composition and is not decomposed.

Best-effort scan bound and port narrowing. A domain's index run is scanned up to a fixed block budget; for the two domains that exceed it, `count` is a lower bound and, for the one that is not captured (ozon.ru), the no-200 verdict is best-effort rather than definitive. Captures on non-standard ports sort outside the SURT range and are not counted; standard http/https captures are counted, and this narrowing never decides whether a domain is captured. Both are documented in the dataset methodology.

Capture timestamps are not published in this release. The dataset's `earliest_capture_ts` and `latest_capture_ts` fields are null for every row: the collector reads each CDX record's JSON object, in which the capture timestamp is not a member (it is the index line's separate leading field), so no per-domain time distribution is derived here. The crawl window is stated from Common Crawl's own crawl identity (June 2026), not measured from the dataset; a within-window time distribution would require re-deriving timestamps from the index lines.

Presence is not consumption or memory. The census measures coverage of one crawl and nothing downstream of it. It makes no measurement of any corpus that derives from Common Crawl, of any model trained on such a corpus, or of any assistant's behavior. Presence in the crawl does not imply a model retains or reproduces the content, and absence does not guarantee it is forgotten; the census is a measurement of the pipeline's front door, reported as exactly that.

REFERENCES

1. M. Koster, G. Illyes, H. Zeller, and L. Sassman, IETF. *RFC 9309: Robots Exclusion Protocol* (2022). https://www.rfc-editor.org/rfc/rfc9309.html Accessed 2026-07-09. [archived]
2. 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.
3. 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]
4. 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]
5. 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]
6. 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]
7. 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]
8. 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]
9. Common Crawl Foundation. *Get Started — accessing the data (s3://commoncrawl and data.commoncrawl.org)* (2026). https://commoncrawl.org/get-started Accessed 2026-07-10. [archived]
10. Common Crawl Foundation. *The CDXJ Index (index file structure, SURT urlkey, filename/offset/length Range access)* (2026). https://commoncrawl.org/cdxj-index Accessed 2026-07-10. [archived]
11. Common Crawl Foundation. *Crawl collection registry (collinfo.json) — CC-MAIN-2026-25, June 2026 Index* (2026). https://index.commoncrawl.org/collinfo.json Accessed 2026-07-10. [archived]
12. Common Crawl Foundation. *CCBot* (2026). https://commoncrawl.org/ccbot Accessed 2026-07-09. [archived]

13. Brown et al. (NeurIPS). *Language Models are Few-Shot Learners* (2020). https://arxiv.org/abs/2005.14165 Accessed 2026-07-08. [archived]

14. Baack, Mozilla Foundation. *Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI* (2024). https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/ Accessed 2026-07-08. [archived]

15. Penedo et al. (NeurIPS Datasets & Benchmarks). *The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale* (2024). https://arxiv.org/abs/2406.17557 Accessed 2026-07-08. [archived]

HOW TO CITE

Barkhausen AI (2026). Who is in the corpus pipeline's front door: a Common Crawl coverage census of 2,000 domains. https://barkhausen.ai/research/common-crawl-coverage-census-2026/

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