

Text-and-data-mining reservations in the EU-27 press, channel by channel

EasyxLab · EasyByte Hub S. Coop. Mad. · Study S16

 Working draft Published 2026-10-03

Abstract

We read what a crawler can read on 1,511 news websites in the 27 EU Member States: `robots.txt` (RFC 9309), TDMRep in its three channels, Content-Signal, the IETF Content-Usage rule, `noai` and `llms.txt`. We could read 1,356 sites in full. Of those, 99 of 1,356 (7.3%), 95% CI 6.0–8.8%, state a reservation of text and data mining that binds a crawler whatever its name. 558 of 1,356 (41.2%) block at least one named AI crawler, and 480 of 558 (86.0%) of those state nothing a newly named crawler would read. TDMRep is on 71 of 1,356 (5.2%), mostly in headers and `<meta>` elements and mostly in France. 218 of 1,510 (14.4%) sites write a reservation or prohibition in `robots.txt` comments, which RFC 9309 parsers discard. Under the access rule frozen before collection the headline is 93 of 1,282 (7.3%). The crawler documentation we could read names `robots.txt` and no other channel.

1. Introduction

The text-and-data-mining (TDM) exception of Article 4 of Directive (EU) 2019/790 applies unless rightholders have reserved their rights "in an appropriate manner, such as machine-readable means in the case of content made publicly available online". Article 53(1)(c) of the AI Act obliges providers of general-purpose AI (GPAI) models to identify and comply with such reservations. It has applied since 2 August 2025; models already on the market have until 2 August 2027, and fines are possible from 2 August 2026.

Most news publishers reacted with `User-agent: GPTBot`-style groups in `robots.txt`, a reservation addressed only to the bots named; others write prose in comments, which parsers ignore, or use TDMRep, Content-Signal or the IETF AI Preferences drafts.

The question: **for EU-27 news websites, how many state a machine-readable reservation that binds any crawler, how many only block named bots, state it only in comments or state none, and how many contradict themselves?** A second panel reads what GPAI providers say they honour.

2. Prior work

Searches (2026-10-03). - OpenAlex, six queries ("TDMRep text and data mining reservation protocol", "robots.txt AI crawlers news publishers", "text and data mining opt-out machine-readable Article 4(3)", "AI training opt-out robots.txt consent", "Content Signals robots.txt ai-train", "news publishers block GPTBot"); GitHub search ("tdmrep", "ai crawler robots.txt census").

- Not queried: export.arxiv.org and `/api` are disallowed for us; arxiv.org allows `/abs` (Crawl-delay 15), but its comment "Indiscriminate automated downloads from this site are not permitted" stops our conservative helper, so arXiv figures below are as checked by the independent reviewer's helper agents on 2026-10-03, which read single `/abs` pages, a targeted and not indiscriminate download, keeping the 15 s crawl delay. Zenodo's `/api` is disallowed; OpenAlex later answered 429. Queries and hits: `data/prior_work_search.json`.

Closest results.

- **Reuters Institute** (2024): "By the end of 2023, 48% of the most widely used news websites across ten countries were blocking OpenAI's crawlers"; 24% blocked Google's AI crawler. Shares ranged from "79% in the USA to just 20% in Mexico and Poland".
- **Ben Welsh's news-homepages tracker** (read 2026-10-03): "596 of 1,157 news publishers" (51.5%) told OpenAI, Google AI or Common Crawl to stop scanning their sites.
- **Steinacker-Olsztyn, Gosain and Dao** (arXiv 2510.10315, 2025): on 4,079 news sites, "60.0% of reputable sites disallow at least one AI crawler".
- **Dinzinger, Heß and Granitzer** (arXiv 2404.02309, 2024), on Common Crawl: about 45 hosts serve `tdmrep.json`, around 60 domains use the `tdm-reservation` tag, "particularly French websites, e.g. lefigaro.fr, appear to be leading", and `noai` is on 82 hosts.
- **AI Crawler Observatory - Italy** (`mxdangelo/aicrawl-census`, dashboard of 2026-10-02): it separates specific from wildcard blocks. In its News sector, 26 of 82 sites (31.7%) block GPTBot by name. 9 of 595 sites reserve rights with TDMRep, read through all three channels.
- **Consent in Crisis** (Longpre et al., 2024): `robots.txt` and terms of service of the sources of C4, RefinedWeb and Dolma. It codes wildcard versus named AI agents and reports "~5%+ of all tokens in C4, or 28%+ of the most actively maintained, critical sources in C4, fully restricted from use", and "nearly 45% of all News website tokens are fully restricted".

What is new here is scale and channel coverage for the EU-27 press. We count reservations written only in comments, with an audited detector. We give the first published counts of Content-Signal and IETF `Content-Usage` in the press. We give TDMRep by channel across 27 countries. The wildcard-versus-named split and the French lead in TDMRep were already known from smaller or differently framed samples.

3. Legal and technical baseline

Legal texts were downloaded from the Publications Office (Cellar) on 2026-10-03. Their sha256 are in `data/sources_manifest.csv`.

- **Directive (EU) 2019/790, Art. 4(3)**. "The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online."
- **Regulation (EU) 2024/1689, Art. 53(1)(c)**. Providers shall "put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790".

- Art. 113(b): "Chapter III Section 4, Chapter V, Chapter VII and Chapter XII and Article 78 shall apply from 2 August 2025, with the exception of Article 101".
- Art. 101(1) allows fines "not exceeding 3 % of their annual total worldwide turnover in the preceding financial year or EUR 15 000 000, whichever is higher".
- Art. 111(3): "Providers of general-purpose AI models that have been placed on the market before 2 August 2025 shall take the necessary steps in order to comply with the obligations laid down in this Regulation by 2 August 2027."
- **Regulation (EU) 2026/1744 (Digital Omnibus on AI).** Point (40) of its Article 1 replaces points (a) and (c) of Art. 113's third paragraph and adds "(d) Articles 102 to 110 shall apply from 27 July 2026". It amends neither Art. 53(1)(c) nor Art. 101. The "Article 53" it amends belongs to Regulation (EU) 2018/1139.
- **GPAI Code of Practice, Copyright chapter, Measure 1.3.**
 - Source: the Commission's PDF lies under `ec.europa.eu/newsroom/dae/`, which `ec.europa.eu's robots.txt` disallows. We quote the unofficial transcription at `code-of-practice.ai`, read 2026-10-03, which describes itself as an unofficial, best-effort service.
 - Text. Signatories commit "to employ web-crawlers that read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt), as specified in the Internet Engineering Task Force (IETF) Request for Comments No. 9309, and any subsequent version ... and to identify and comply with other appropriate machine-readable protocols to express rights reservations pursuant to Article 4(3) of Directive (EU) 2019/790, for example through asset-based or location-based metadata, that have either have been adopted by international or European standardisation organisations, or are state-of-the-art, including technically implementable, and widely adopted by rightsholders, considering different cultural sectors, and generally agreed through an inclusive process based on bona fide discussions to be facilitated at EU level with the involvement of rightsholders, AI providers and other relevant stakeholders as a more immediate solution, while anticipating the development of cross-industry standards."
 - The Code sets no adoption threshold, and we test none. Signatories include Amazon, Anthropic, Google, Microsoft, Mistral AI and OpenAI.

Protocols on the measurement date. TDMRep (W3C CG Final Report, 2024-05-10): the `/well-known/tdmrep.json` file, `tdm-reservation` / `tdm-policy` headers and `<meta>`. IETF AI Preferences: vocab-08 (2026-09-14) defines `train-ai`, `ai-use`, `search = y / n`, and its s. 6.4 says unknown labels must be ignored (in -01, June 2025, the training label was `ai`; -02 renamed it); attach-05 (2026-08-19) defines the `Content-Usage` header and per-group `robots.txt` rule. Content-Signal (Cloudflare): `search`, `ai-input`, `ai-train` = yes/no. `noai` and `notdm` have no specification.

4. Data

Outlets. We took the Wikidata items that are newspapers (with subclasses), online newspapers, news websites or news magazines. Each has an EU-27 country, an official website, and no dissolution or discontinuation date. That gives 2,553 items. We queried QLever (Wikidata copy of 2026-08-12), because `query.wikidata.org` disallows `/sparql`.

Frame. A host is in the frame if it is in the top-100,000 bucket of its own country's Chrome UX Report list for August 2026. We excluded library and archive hosts, platforms, pages inside other hosts, duplicates,

and official gazettes and public-sector publishers. The gazette rule was completed after the scan (deviation D6).

The frame has 1,511 news websites, from Malta (1) to France (196): large-band countries ($n \geq 100$: DE, ES, FI, FR, IT, SE) 908 hosts, medium (30–99, 10 countries) 472, small (< 30 , 11 countries) 131.

5. Method

Access. One helper (`politefetch.py`) reads `robots.txt` first (RFC 9309 parser of study S8), then only the channels a site declares for machines (TDMRep file, home-page headers and `<meta>` , `/llms.txt`), at one request per second per host, storing no article content. A general prohibition of automated access in the comments stops it. Purpose-specific notices (no TDM, no AI training), section labels and inline comments on one agent's rule do not; the first count as reservations. A ban that names TDM/AI purposes but ends in an open clause ("or otherwise", "for any purpose", "not limited to") is general (v4, D7). The frozen rule stopped at any comment the detector flagged; the distinctions above were introduced after collection (D5) and are a change of rule. EasyxLab's lab-wide rule (3 October 2026) treats `robots.txt` as binding at the legal minimum; S16 deliberately keeps this stricter rule because its subject is those reservations.

Classification (frozen before collection). **Agnostic reservation:** a TDMRep `tdm-reservation: 1` (file, header or `<meta>`), Content-Signal `ai-train=no` , Content-Usage `train-ai=n` for an unnamed crawler, or a `*` group disallowing `/` . Otherwise **named bots only** if one of 26 named AI tokens is disallowed `/` ; otherwise **comment-only** if comments carry a reservation or general prohibition; otherwise **none stated**. A **contradiction** is two explicit use-preference channels saying reserve and allow. Sites read on `robots.txt` alone keep their class with the suffix `_robots_only_partial` .

Comment detectors are multilingual regular expressions: v1 in the scan; v2 fixed misses around domain names (D2); v3 (D5) separated general prohibitions from inline per-agent comments, labels and purpose-specific notices, read list notices across blocks and stopped "exa.ai" matching "AI"; the 98 sites v3 freed were re-read once (same fetch mechanics, changed prohibition rule); v4 (D7) treats purpose-naming bans with an open clause as general and short headings as labels, moving 24 re-read sites back to robots-only.

6. Results

6.1 Classes

Of the 1,511 frame hosts, 1,510 answered for `robots.txt` : 1,421 with a file and 89 with a 4xx. One was unreachable. 154 sites were read on `robots.txt` alone, and 1,356 in full.

CLASS (1,356 FULLY READ SITES)	SITES	SHARE	95% CI
agnostic reservation	99	7.3%	6.0–8.8
named bots only	480	35.4%	32.9–38.0
comment-only	0	0.0%	0.0–0.3
none stated	777	57.3%	54.7–59.9

The headline is **99 of 1,356 (7.3%)**, CI 6.0–8.8%: these fully read sites state a reservation that binds a crawler whatever its name. **558 of 1,356 (41.2%)** block at least one named AI crawler at the root. **480 of 558 (86.0%)**, CI 82.9–88.7%, of those have no agnostic reservation, so a GPAI crawler announced under a new token would meet no machine-readable reservation in the channels we read. 777 of 1,356 (57.3%) state no reservation in any channel we read.

Frozen-rule sensitivity. D5 changed the access rule after collection. Without D5 the fully read set is 1,282 sites:

- 93 of 1,282 (7.3%), CI 6.0–8.8%, have an agnostic reservation;
- 486 of 1,282 (37.9%) block named bots, and 413 of 486 (85.0%) of those have no agnostic reservation;
- TDMRep is on 66;
- there are 0 contradictions.

All 1,510 sites.

- Named blocking is exact, because `robots.txt` was read everywhere: 699 of 1,510 (46.3%).
- The agnostic count is a minimum, 110 of 1,510 (7.3%).
- Named blockers without an agnostic reservation are a maximum, 621 of 699 (88.8%).
- Most-named tokens: GPTBot (637), CCBot (633), Bytespider (538); 403 sites block all four of GPTBot, ClaudeBot, Google-Extended and CCBot.

Among fully read sites the agnostic channels overlap. A `*` group disallowing `/` gives 11 (0.8%), TDMRep 71 (5.2%), Content-Signal `ai-train=no` 18 (1.3%) and IETF `train-ai=n` none. Counting the non-standard `noai` would give 105 of 1,356 (7.7%).

6.2 Channels and contradictions

TDMRep is on 71 of 1,356 (5.2%), CI 4.2–6.6%, always with `tdm-reservation` = 1.

- By channel: meta 41, header 22, file 13.
- 58 of the 71 only in a header or a `<meta>` element; a census of `/ .well-known/tdmrep.json` misses those.
- Another 61 sites answer that path with an HTML page and status 200 (soft 404).
- 57 point to a `tdm-policy`.
- By country: France 53, Spain 9, Italy 5, Austria 2, Germany and Poland 1 each. Among fully read French sites, 53 of 192 (27.6%, CI 21.8–34.3%) use it; 52 of 183 under the frozen rule.

Content-Signal lines appear on 21 sites: 18 with `ai-train=no` and 3 with `ai-train=yes` .

Content-Usage appears on 17 sites (13 Finnish, 4 Italian). All 17 write `ai=n` , the -01 label. The template's comments state an agnostic intent; `gazzettadimantova.it` writes "Block any non-specified AI crawlers ... from using content for training AI models". Under the frozen rule, and under `vocab-08`, this is not a reservation. Read as -01's training label, the agnostic share would be 116 of 1,356 (8.6%), CI 7.2–10.2%.

Contradictions. 2 of 1,356 under the current rule, 0 under the frozen rule. Both are `Corriere della Sera` sites (`corriere.it` and its Veneto edition), each sending Content-Signal `ai-train=yes` and `<meta name="tdm-reservation" content="1">` . The scouting pilot had seen this. The first scan had stopped at the inline comment on Corriere's `Yandex` rule, and only the D5 re-scan read the sites.

llms.txt is present on 120 of 1,339 (9.0%) fully read sites that answered for it.

6.3 Reservations written as prose

218 of 1,510 (14.4%) sites carry a reservation of TDM or AI use (196) or a general prohibition of automated access (154) in `robots.txt` comments. Among fully read sites, 64 do. 60 of those have no agnostic machine-readable reservation, and all 60 block named bots, so no fully read site relies on the comment alone. The texts are group templates (a German § 44b UrhG formula, Mediahuis, Bonnier, an Austrian § 42h(6) notice, a Finnish group's AI-training notice); to an RFC 9309 parser none of them exists.

6.4 By band, country and popularity

Agnostic among fully read sites, by band:

- large countries 74 of 800 (9.2%, CI 7.4–11.5);
- medium 15 of 431 (3.5%, CI 2.1–5.7);
- small 10 of 125 (8.0%, CI 4.4–14.1).

All countries with at least 30 fully read sites:

- France 55 of 192 (28.6%, CI 22.7–35.4);
- Poland 8 of 31 (25.8%, CI 13.7–43.2), a small n;
- Spain 9 of 182; Italy 5 of 114; Sweden 3 of 77; the Netherlands 2 of 53; Ireland 1 of 33;
- Greece 1 of 52, Hungary 1 of 52, Romania 1 of 54;
- Germany 1 of 92 (63 German sites were read on `robots.txt` only); Finland 1 of 143;
- Czechia 0 of 36 and Denmark 0 of 65.

Named blocking falls with popularity: 407 (58.2%) of the 699 sites in the top-1,000 CrUX bucket, and 21 of 110 (19.1%) in the 50,001–100,000 bucket.

6.5 What providers say they honour

We archived the crawler documentation of 14 organisations on 2026-10-03. We could read it for 8 of 14 providers: OpenAI, Anthropic, Google, Mistral AI, Apple, Amazon, Perplexity and Common Crawl.

- All 8 name `robots.txt` .

- Amazon also honours a `noarchive` robots meta tag as "do not use the page for model training". Apple mentions `nosnippet`.
- None names TDMRep, Content-Signal, the IETF `Content-Usage` or `noai`.
- Not read: Meta, ByteDance (disallowed), Microsoft (script-rendered), Cohere (404), Aleph Alpha, DeepSeek (no crawler page). These are statements, not observed behaviour.

6.6 Detector audits (all by AI agents)

- **The builder agent's own sample** (100 sites, seed 20261003):
 - reservations right on 54 of 54 flags;
 - prohibitions v1 right on 19 of 44, with 17 missed; v2 right on 35 of 62 (optimistic, because v2 was revised after reading this sample).
- **The first independent reviewer, v1/v2** (44 other sites, seed 777):
 - reservation precision 20 of 20, recall 20 of 21;
 - prohibition precision 15 of 21, recall 15 of 15;
 - it found that per-block matching misses list-style notices (D4).
- **The re-reviewer, v3** (43 further sites, seed 4242):
 - prohibition precision and recall 15 of 15 in the sample, but recall about 84% (130 of 154) once the 24 D7 misses are counted;
 - reservation precision 15 of 16 (a heading), recall 15 of 15.
- **v4** changes exactly these two failure modes. No audit of v4 is claimed.

7. Limitations and deviations

- **Minima and maxima.** For 154 sites we read `robots.txt` only, because a deliberately conservative detector flags their comments as a general prohibition. Over all sites, agnostic, TDMRep and contradiction counts are minima and the "named bots only" share is a maximum. The fully read figures are exact for those 1,356 sites on 2026-10-03, which are not "the EU press". On 12 of them the home page was not observed (timeout, error, or disallowed).
- **D0 (breach).** The scouting pilot read the home pages of up to 14 hosts whose comments prohibit robots.
- **D1 (breach).** 7 requests reached a sign-on host before its `robots.txt`.
- **D2 (breach).** 245 requests went to 92 sites whose general prohibition v1 missed.
- **D3.** The main scan's log cannot show pacing (completion times).
- **D4 (breach).** 3 requests read actu.fr, whose list-style prohibition was missed.
- **D5 (rule change after collection).** It freed 98 sites, which were re-read at 11:55:54–11:56:18 UTC; the re-scan also read one redirect target's `robots.txt`.
- **D6.** 22 official gazettes were excluded (frame 1,533 → 1,511).
- **D7 (breach).** In that re-scan, 96 requests (72 beyond `robots.txt`) went to 24 sites whose notice is a general prohibition with an open clause: 23 Mediahuis sites and fd.nl. The study's instructions had

wrongly classed the Mediahuis notice as purpose-specific.

In every breach the data were deleted; URL and time stay in the private log.

- **Coverage.** Home page only; `ai.txt` pointers, `notdm` and Content-Signal `use=` not read; 26 named tokens; frame shaped by Wikidata and Chrome popularity.
- **Not a legal assessment.** Whether a channel is an "appropriate manner" under Art. 4(3) is for courts to decide.

8. Data, code and licences

Code Apache-2.0; data and text CC BY 4.0; Wikidata CCO; "CrUX datasets by Google are licensed under a Creative Commons Attribution 4.0 International License". `data/records.csv` describes what each publisher's files state on that date, not an assessment of the publisher: the obligation falls on AI providers. No `robots.txt` bodies, page content or provider documents are published. The checker (`scripts/optout_check.py <domain>`) may become the free tool `optout-lint`.

9. Automation and review

AI agents did this study end to end:

- A scouting agent proposed it after a 78-host pilot (D0).
- A coordinating agent selected it, verified the legal quotes and decided the D5 and D7 rules. The D5 instruction misclassified the Mediahuis notice.
- The builder agent wrote the code, ran the collection and re-scan, labelled its audit sample, analysed the data and wrote this paper. It found D2 and D3.
- An independent AI reviewer re-computed every figure, re-read 40 sites and audited v1/v2. It found D4, D6 and the errors behind D5.
- A second independent AI reviewer audited v3 and found D7.

`scripts/check_headline.py` checks a list of headline phrases against `data/`.

Competing interests

EasyByte Hub S. Coop. Mad. publishes websites and studies AI-crawler traffic (EasyxLab studies S2 and S9). It has no stake in any publisher or AI provider. The checker may become a free EasyxLab tool (`optout-lint`). Our other tools (`verifactu-lint`, `ai-mark-lint`, `crs-lint`, `attest-lint`) are unrelated to this topic.

References

- Directive (EU) 2019/790; Regulations (EU) 2024/1689 and 2026/1744 (Cellar, 2026-10-03).
- GPAI Code of Practice, Copyright chapter (unofficial transcription, `code-of-practice.ai`, 2026-10-03).
- RFC 9309, Robots Exclusion Protocol (2022). W3C TDMRep Community Group Final Report (2024-05-10).

- IETF draft-ietf-aipref-vocab-01, -02 and -08; draft-ietf-aipref-attach-05. Cloudflare Content Signals documentation.
- Fletcher, R., *How many news websites block AI crawlers?*, Reuters Institute, 2024.
- Welsh, B., news-homepages: *Who blocks OpenAI, Google AI and Common Crawl?* (palewi.re, read 2026-10-03).
- Steinacker-Olsztyn, Gosain and Dao, *Is Misinformation More Open? A Study of robots.txt Gatekeeping on the Web*, arXiv 2510.10315 (2025).
- Dinzinger, Heß and Granitzer, arXiv 2404.02309 (2024).
- Longpre et al., *Consent in Crisis: The Rapid Decline of the AI Data Commons*, arXiv 2407.14933 (2024).
- mxdangelo/aicrawl-census (dashboard of 2026-10-02). Crawlora, AI-Crawler Blocking Index. AI Discovery Radar, Zenodo 10.5281/zenodo.22178283.
- Chrome UX Report (Google). zakird/crux-top-lists. QLever (University of Freiburg). Wikidata (CC0).
- EasyxLab study S8 (robots9309.py).

Cite this study

CITATION

EasyxLab (2026). Text-and-data-mining reservations in the EU-27 press, channel by channel. Study S16. EasyByte Hub S. Coop. Mad.
<https://github.com/easybytehub/easyxlab/tree/main/studies/s16-eu-press-tdm-reservations>

BIBTEX

```
@techreport{easyxlab_s16,  
  title      = {Text-and-data-mining reservations in the EU-27 press, channel by  
channel},  
  author     = {{EasyxLab}},  
  institution = {EasyByte Hub S. Coop. Mad.},  
  number     = {S16},  
  year       = {2026},  
  url        = {https://github.com/easybytehub/easyxlab/tree/main/studies/s16-eu-  
press-tdm-reservations}  
}
```