WordCamp Europe
Kraków 2026

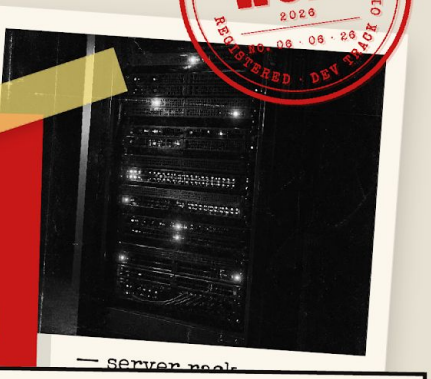
The AI-first WordPress site
CRAWLER TO CITATION

Alain Schlesser

// WCEU / 2026
// KRAKOW · JUN 06

SATURDAY · 2:45 PM CEST
DEVELOPMENT · TRACK 1

CRAWLER TO CITATION.



▸ ABSTRACT

The AI-first WordPress site. Make it readable, citable, and measurable by the machine layer that just appeared on top of the web.

▸ AUTHOR

ALAIN SCHLESSER Agentic architect & engineer · WP-CLI maintainer · Google Developer Expert
alainschlesser.com

// SECTION DIVIDER
// ACT ONE OF THREE

5 MIN
5 SLIDES

▸ PROBLEM
▸ 5 MIN

ACT I.

THE PROBLEM

AI is now a traffic source. **It sends visitors, and it takes content.**
The shift is settled — we're here to talk about what to do.

PICTURE THIS

Sometime in the last week, an AI system decided whether to recommend your site as the answer to someone's question.

YOU WEREN'T IN THE ROOM.

AND MOST OF THE TIME, THE ANSWER WAS NO.

— A.S.

// ACT I · THE PROBLEM
// FIGURE 03.1

THE AUDIENCE JUST DOUBLED.

THE SHIFT, IN ONE NUMBER

— PAGE 14 · FOLD C —

BOTS OUTNUMBER HUMANS

▸ READ THIS

▸ AI BOT TRAFFIC · YOY · OPEN WEB

+187%

Growth across the open web, Jan - Dec 2025. The direction is one-way. Don't argue with it — design for it.

— CLOUDFLARE RADAR · Q1 2026

Humans, same window **+3.1%**

// ACT I · THE PROBLEM
// FIGURE 04.1

QUALITY, NOT JUST VOLUME

A FULL REVERSAL.

BEFORE → AFTER
in 12 months.

▸ MARCH 2025

AI traffic converted -38 % vs non-AI. The skeptic position made sense.

-38%

▸ MARCH 2026

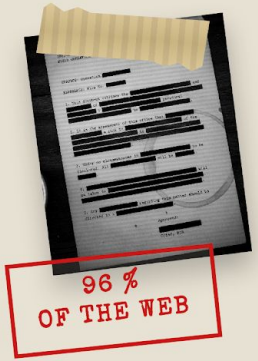
AI traffic converts +42 %, RPV +37 %, time on site +48 %, pages/visit +13 %

+42%

// ACT I · THE PROBLEM
// THE GAP

THREE WAYS TO BE INVISIBLE.

WHERE MOST WP SITES ARE



1

INVISIBLE

No robots.txt strategy. No schema. No measurement. The AI ecosystem flows past you in both directions.

2

VULNERABLE

Default-open to training crawlers. No governance. Content harvested by anyone. No idea who, or how much.

3

OLD-PARADIGM

Optimised hard for Google 2018. Nothing ready for the AI layer that just appeared on top of it.

// SECTION DIVIDER
// ACT TWO OF THREE

18 MIN
16 SLIDES · 5 LAYERS

▸ STACK
▸ 5 LAYERS

ACT II.

THE STACK

Five layers — **crawlers · robots · schema · content · measurement.**
Take notes on the ones you don't have yet.

// ACT II · STACK · LAYER 1
// CRAWLER AWARENESS

THE THREE TIERS

THREE BOTS. THREE STANCES.

→ tip:

agents = users.
treat them so.

▸ TIER 01

TRAINING

GPTBot · ClaudeBot · CCBot
Google-Extended · Bytespider
· Amazonbot

BLOCK OR RATE-LIMIT
UNLESS PAID

▸ TIER 02

SEARCH / CITATION

OAI-SearchBot ·
PerplexityBot
ChatGPT-User · AppleBot-
Extended · Bingbot

ALLOW.
THESE DRIVE CITATIONS.

▸ TIER 03

AGENT

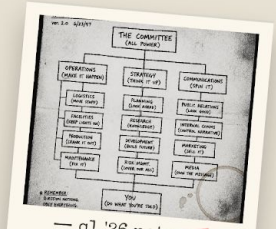
OpenAI Operator
Anthropic Computer Use
Perplexity Comet

ALLOW.
TREAT LIKE A USER.

// ACT II · STACK · LAYER 1
// VENDOR SHARE, Q1 2026

MULTI-VENDOR REALITY

89 % IS TRAINING.



— q1 '26 notes —

CLOUDFLARE Q1 2026

1 GOOGLEBOT
31.6% — the elephant. Still leads, no longer a monopoly.

2 META-EXTERNALAGENT
16.7% — quietly enormous. Mixed-purpose crawler.

3 GPTBOT
12.0% — OpenAI training. Clear opt-out.

4 CLAUDEBOT
11.7% — Anthropic. Distinct UAs for training vs search.

5 APPLEBOT-EXTENDED
5.8% — up 124% in Q1. Siri Intelligence is here.

6 THE REST
22.2% — Bytespider, Amazonbot, CCBot, PerplexityBot, dozens more.

// ACT II · STACK · LAYER 1
// WHO'S AT YOUR DOOR

TOOLS YOU ALREADY HAVE

YOU CAN SEE ALL OF THEM.

1

USER-AGENT

Every major vendor publishes theirs. UA-sniffing works — until it doesn't. Trivially spoofable.

2

IP / ASN

OpenAI, Anthropic, Apple publish ranges. UA + ASN = verified bot.

3

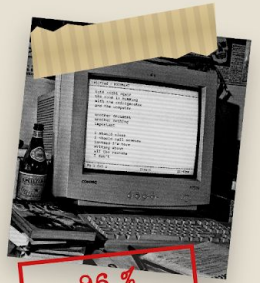
SERVER LOGS

One grep one-liner on your access log. Visibility today, no plugins.

// ACT II · STACK · LAYER 2
// ROBOTS.TXT STRATEGY

WHAT WP SHIPS

DEFAULT IS NOT A STRATEGY.



ROBOTS.TXT

WORDPRESS 6.5 DEFAULT

```
# wordpress default — same since 2008

User-agent: *
Disallow: /wp-admin/
Allow:    /wp-admin/admin-ajax.php

# zero ai-specific directives.
# every ai crawler gets exactly the same treatment.
# this is not a stance — it's the absence of one.
```

// ACT II · STACK · LAYER 2
// THE STRATEGIC CONFIG

THREE TIERS. THREE DECISIONS.

ROBOTS.TXT

THE 2026 CONFIG

```
# tier 1 — training (default: deny)
User-agent: GPTBot
User-agent: ClaudeBot
User-agent: CCBot
Disallow: /

# tier 2 — citation (default: allow)
User-agent: OAI-SearchBot
User-agent: PerplexityBot
Allow: /

# tier 3 — agent (allow, like users)
User-agent: ChatGPT-User
Allow: /
```

YOUR CONTENT. YOUR DECISION.

1

THE RETURN IS ZERO

Training crawlers take, don't give. No citation, no referral, no attribution. Bandwidth on the way in is your bill.

2

LEGAL DIRECTION

Publishers vs major AI labs in active litigation. Direction of travel: more control, not less. A clear opt-out today is cheap insurance.

3

TRACK RECORD

Some operators have a documented history of ignoring opt-outs. The compute they spend crawling you is bandwidth you pay for.

4

RE-AUDIT QUARTERLY

The line between training and search is shifting. Some vendors run unified crawlers. Not a permanent answer — a current one.

// ACT II · STACK · LAYER 2
// WHEN HONOR-SYSTEM FAILS

ENFORCEMENT LAYERS

WHEN HONOR SYSTEM FAILS.

1

WAF RULE

Cloudflare AI Crawl Control, Fastly's equivalent, Akamai's. One toggle in a dashboard.

2

RATE-LIMIT + IP BLOCK

Server-level. More granular. Useful when you want to allow a vendor sometimes, not always.

3

WEB BOT AUTH

HTTP message signatures. Cryptographic identity for well-behaved bots. Standards work in progress.

> Schema used to be an SEO trick. In 2026 it's something else.

TREAT IT LIKE AN API. THE REST OF THIS LAYER MAKES SENSE.

— A.S.

// ACT II · STACK · LAYER 3
// RECEIPTS

WHY WE KNOW THIS

THREE INDEPENDENT RECEIPTS.

1

PLATFORM

Fabrice Canel, Principal PM Bing, SMX Munich: "Bing & Copilot use schema to feed their LLMs." Not me inferring — a Microsoft PM on a public stage.

2

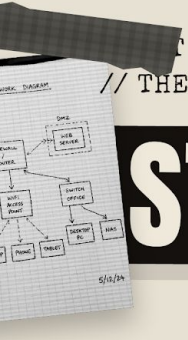
TRAINING TIME

LLM pre-training pulls from Common Crawl. Your structured data is in that snapshot. Every model inherits the structure. Effect is one-way and permanent.

3

PROTOCOL LAYER

NLWeb, MCP-based approaches, llms.txt — several proposals worth taking seriously. All of them speak schema. Disclosure: I led NLWeb work at Yoast.



STUB SCHEMA HURTS YOU.

DO IT PROPERLY

▸ CITATION RATE · MINIMAL / GENERIC SCHEMA

41.6%

Pages with no schema were cited 59.8% of the time. Rich attribute-complete schema hit 61.7%. Half-finished schema looks like low effort. AI systems notice.

— GROWTH MARSHAL · N=730

Either commit, or don't bother. **61.7%**

// ACT II · STACK · LAYER 3
// THE @ID GRAPH

GRAPH BEATS BLOBS

CONNECT EVERY @ID.

graph > blob.

every. single. time.

▸ BEFORE — isolated blobs

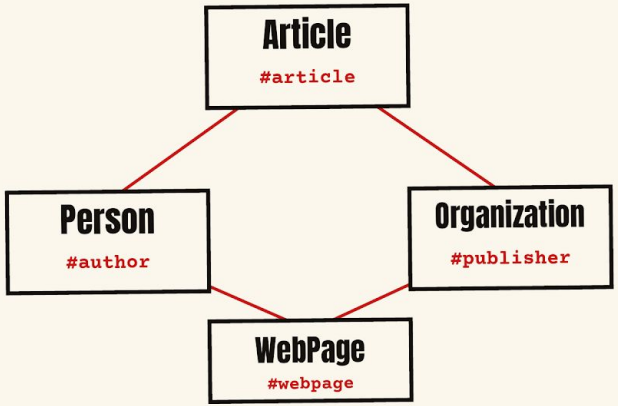
Article
no @id

Person
no @id

Organization
no @id

→ AI sees three loose objects.

▸ AFTER — interconnected @id graph



// ACT II · STACK · LAYER 4
// CONTENT PATTERNS

EVIDENCE-BASED

THREE EDITS.

+30 - 40 %.

// content edit pa…
cite. quote. count.
→ ~30–40 % lift

PRINCETON GEO-BENCH

1 CITE SOURCES

Pages that cite get cited. Add references. Hyperlink to primary sources. Position-adjusted lift: ~30-40 %.

2 QUOTE AUTHORITIES

Pull direct quotes into blockquotes. AI treats your page as quoting an authority. Same lift band.

3 USE STATISTICS

Replace "a lot" with the actual number. Specific figures get pulled when AI grounds its own claims.

FRESH BEATS FOSSIL.

1

FRESHNESS SIGNAL

Microsoft Canel: "Generative AIs value fresh content — as a reference check of LLM training." His channel:

INDEXNOW

. WP plugins exist. Cheap.

2

ENTITY DENSITY

Named entities + internal links improve AI comprehension. King (iPullRank): relevance, not authority, drives AI Overview placement.

// ACT II · STACK · LAYER 4
// AWKWARD TRUTH

BECOMING THE SUBSTRATE

WIKIPEDIA. REDDIT. YOU.

SEMRUSH
100M+ CITES

1 CHATGPT

Wikipedia 26-48 % of top-10 share. Reddit went 60 % → 10 % in six weeks after a Google parameter change.

2 GOOGLE AI MODE

Reddit, LinkedIn, YouTube growing. Medium, Quora declining.

3 PERPLEXITY

Reddit at ~40 %. WSJ, NYT, Bloomberg absent from top 20.

4 THE PLAY

You can't out-Reddit Reddit. Be the original-research source. The thing the forum thread links to.

REFERRERS TO WATCH.

1. **CHAT.OPENAI.COM**
   The biggest single source for most categories.

2. **PERPLEXITY.AI**
   Smaller but converts hard. Power users.

3. **GEMINI.GOOGLE.COM**
   Growing fast. Tied into the Google account flow.

4. **COPILOT.MICROSOFT.COM**
   Enterprise traffic. Edge default.

5. **X.COM (GROK)**
   Stripped sometimes. Show up as x.com.

6. **KAGI.COM**
   Small, paid, high-intent.

FOUR METRICS. NO MORE.

start small.

measure something. beats nothing.

1 VISITS BY AI REFERRER

Undercounted baseline. Trend over time is what matters, not absolute.

2 CONVERSION / ENGAGEMENT BY REFERRER

Pages, time, goal completions. Does the Adobe pattern hold for you?

3 CITATION SHARE

Pick ten queries. Rotate weekly. Log which sites get cited and track your share over time.

4 CRAWLER SHARE

Grep your access log weekly. Training-tier vs citation-tier bots. The bandwidth picture you don't see in GA.

> Schema is the foundation. Several proposals are competing for what gets built on top — and we don't yet know which one wins.

THE NEXT LAYER ISN'T ABOUT BEING READABLE. IT'S ABOUT BEING CALLABLE.

NLWEB · MCP-FOR-WEB · LLMS.TXT · AGENT-SKILL MANIFESTS. PICK WISELY. OR DON'T — THE PREREQUISITE WORK IS THE SAME.

— A.S.

// SECTION DIVIDER
// ACT THREE OF THREE

4 MIN
4 SLIDES

ACT III.

DO THE WORK

▸ ACTION
▸ TONIGHT

Two timelines. **Tonight, and the next thirty days.**
You can be ahead of 96 % of the web before you sleep.

// ACT III · CHECKLIST
// FIGURE 23.1

SAME-DAY ACTIONS

DO THIS TONIGHT.

▸ ACTION
▸ ITEMS

1 AUDIT YOUR ROBOTS.TXT

Explicit decision per tier: training, citation, agent.

2 VERIFY YOUR SCHEMA GRAPH

@id cross-references. Not isolated JSON-LD blobs.

3 ATTRIBUTE-COMPLETE SCHEMA

Organization. Person. Article. Stub schema underperforms none.

4 RUN A READINESS SCANNER

One scan today. Capture the baseline score.

5 EDIT FIVE CORNERSTONE PAGES

Citations + quotes + statistics. Princeton GEO patterns.

6 ADD LLMS.TXT + REFERRER TRACKING

Five minutes each. Low risk. Cheap signal.

// ACT III · CHECKLIST
// FIGURE 24.1

THE 30-DAY ARC

FOUR WEEKS. FOUR THEMES.

1

WEEK 01 — INSTRUMENT

Wire the dashboard. Baseline referrals, citations, crawler share. Measurement is slowest to start, so front-load it.

2

WEEK 02 — SCHEMA

Top 10 pages to full attribute-completeness. Don't add types — fill in what's there.

3

WEEK 03 — CONTENT

Top 5 cornerstone articles. Citations, quotations, statistics. Five well-done beats fifty half-done.

4

WEEK 04 — REVIEW & PUBLISH

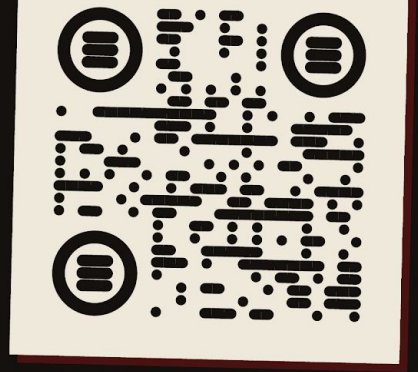
Iterate where the signal is. Publish your AI-readiness page on your site. Public commitment for round two.

// ACT III · TAKE IT HOME
// INTERACTIVE GUIDE

AUDIT YOUR OWN SITE

RUN THE CHECKLIST.

▸ SCAN TO START



tiki.tf/wceu2026

▸ INTERACTIVE GUIDE

Walk every **layer**
on your **own site**,
one tick-box at a time.

Crawlers. Robots. Schema.
Content. Measurement.

– start tonight.

// THE END
// THANK YOU + Q&A

FIN.
OVER TO YOU

THANK YOU.

QUESTIONS?

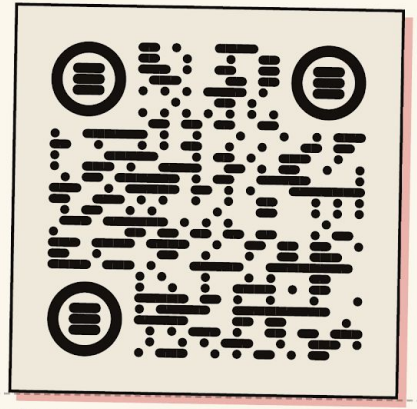
▸ YOUR TURN

Ask me **anything**.

Pick a **layer**,
share what's **working**,
or push back on a **claim**.

– let's talk.

▸ SCAN TO CONNECT



@schlessera · alainschlesser.com

WordCamp Europe
Kraków 2026

THANK YOU!