INDUSTRY WHITE PAPER · JUNE 2026

The Web Data Demand Map 2026

What **37.9 billion** delivered records reveal about where the world actually pulls its data: which sites, which use cases, and which languages.

By the Octoparse Team

Octoparse

Contents

Foreword: Why We're Sharing This

At Octoparse, we sit on an unusual vantage point. Every day, people use our platform to pull data from thousands of websites, and over the past year that added up to 37.9 billion records across the 1,281 most-requested sites. Most of that activity is invisible to the people doing it. Each team sees only its own scrapes, never the pattern across everyone.

We do see the pattern, and we think it is too useful to keep to ourselves.

This report is our attempt to give something back to the community we serve: data collectors, analysts, founders, and anyone whose work depends on the open web. It is not a sales brochure. It is the honest answer to questions we get asked constantly. What are people actually scraping? Which use cases dominate? How hard is each site to extract, really? Where does demand outrun the tools? We answer not with opinion, but with what we observe at scale.

One note on honesty up front: this is our data, from our users. It is a large lens, but a single one (see Methodology). We have tried to report what we see plainly, including the parts that complicate our own story. If it helps you scope a project, set realistic expectations, or find a source you had not considered, it has done its job.

— The Octoparse Team

Executive Summary

Web data has quietly become foundational infrastructure. Roughly 65% of enterprises already used web scraping to feed AI and machine-learning projects in 2024 (Mordor Intelligence), training-related crawling made up the majority of AI crawler traffic in 2025 (Cloudflare Radar), and web-scraped datasets are now the single largest category of "alternative data" bought by financial firms, at about 15% of all alt-data spend (Neudata). Yet for all the talk about the supply of web data (proxies, parsers, anti-bot arms races), remarkably little has been published about the shape of demand: of all the websites on earth, which ones do people actually need data from, for what, and in which languages?

We built this report from our own delivery data: trailing-twelve-month telemetry across the 1,281 most-scraped target sites, covering 37.9 billion delivered records at an aggregate 92.7% success rate[1], with every site tagged by industry, use case, language, real-world extraction success rate, and whether a ready-made template exists.

Six findings stand out:

- **Demand is concentrated, but not winner-take-all.** The top 10 target sites account for roughly half of all delivered records, and the top 50 for about 82.5%, a heavy head over a long, genuinely active tail.
- **The job to be done is research, not theft.** The dominant use cases are Market Research, Lead Generation, Price Monitoring, Review Analysis, and E-commerce Intelligence: commercial decision-support, not data exfiltration.

[1]Aggregate, record-weighted across all delivered records. The per-language averages (Figure 7) and per-use-case averages (Finding 6) are per-site means within their groups and span only the seven reference languages, so they read higher; the two lenses are not directly comparable. See Section 10.

- **E-commerce is the gravitational center.** Retail has about 2.6 times as many distinct high-value target sites as the next industry, followed by recruitment, real estate, and local services, the listing-heavy verticals where the long tail lives.
- **Most demand has no ready-made template.** Across 1,281 high-value sites, 76% have no off-the-shelf template, concrete evidence that pre-built catalogs cannot keep pace with the demand surface. This is exactly where a custom workflow builder earns its keep: when no template exists, you build your own extraction for any site rather than wait for a catalog to catch up.
- **The web is multilingual, and non-English is often more reliable.** 78% of high-value target sites are served in only one language, several of the single largest sources by volume are non-English (Japanese travel and e-commerce, Korean recruitment), and counter-intuitively, those non-English sources extract more successfully than English ones (Japanese about 98%, Korean about 97%, versus Spanish about 93%). The non-English web is underserved by tooling but not harder to collect.
- **The web is more scrapable than its reputation, but unevenly.** Half of these top sites extract at 99% success or better, yet about 11% sit below 80%, a real “hard tail” where anti-bot defenses are winning. Difficulty is predictable: travel pricing and broad market research are the toughest jobs; government data, recruitment, and POI/geo are the most dependable.

The strategic implication: the web-data industry has optimized heavily for the head (a dozen giant platforms) and for English-speaking, technical users. The growth frontier is the long tail, the non-English web, and the non-technical operator. For anyone collecting web data, the practical takeaway is simpler: demand is more diverse, more multilingual, and more reachable than the conventional wisdom suggests, if your tooling can reach past the template library.

1. Why a Demand Map, and Why Now

For two decades, web scraping was a niche engineering task. In 2026 it is a market with real money and real strategic weight behind it. The web-scraping market was valued at about $1.03 billion in 2025 and is projected to reach $2.23 billion by 2031, a 13.78% CAGR (Mordor Intelligence). The AI-driven slice is growing far faster: the AI-driven web scraping market is forecast to add $3.15 billion between 2024 and 2029 at a 39.4% CAGR (Research and Markets), well ahead of the category as a whole.

Two forces are pulling demand upward at once. The first is AI: generative models are hungry for current, structured, real-world data, and training-related crawling made up the majority of AI crawler traffic in 2025 (Cloudflare Radar). The second is competitive intelligence: web-scraped datasets are the largest single category of alternative data bought by investment managers, at roughly 15% of spend, and alt-data spending continued double-digit annual growth through 2024 (Neudata).

But almost all published analysis looks at the supply side: how to get past anti-bot systems, which proxy network to use, which API is fastest. The demand side is largely undocumented. If you wanted to know “across everyone who scrapes the web, where does the demand actually point?” there has been no good public answer.

This paper is an attempt at that answer, built not on survey opinion but on observed behavior: tens of billions of records actually requested and delivered over a full year.

2. Methodology and Data Sources

This analysis draws on revealed demand: what users actually requested and received, not what they said they wanted in a survey.

The delivery dataset. Aggregated, anonymized delivery data for the 1,281 most-scraped target sites over a trailing twelve-month window. Important: this is not the full universe of sites users extract from. Users scrape a far longer tail of destinations. It is a ranked shortlist of the highest-volume, highest-value targets, selected to map where demand concentrates. For each site we have: site type, industry classification, the languages it is delivered in, whether a ready-made extraction template exists, total records delivered, distinct users served, and an extraction success rate. In total this covers 37.9 billion delivered records served to roughly 82,000 cumulative distinct users at a mean 92.7% success rate.

A note on the 92.7% figure. The aggregate 92.7% is computed across all delivered records (record-weighted). The per-language averages in Finding 5 (Figure 7) and the per-use-case averages in Finding 6 are computed per site within their respective groups, and the language averages span only the seven reference languages. Because high-volume sites and out-of-scope languages are weighted differently, the record-weighted aggregate can sit slightly below the per-group means. The two numbers answer different questions and are not directly comparable.

Each site is also tagged with one or more use-case labels (for example, Market Research, Price Monitoring), which lets demand be sliced by purpose, by industry, and by language.

Each site additionally carries a priority tier (S / A / B / C) reflecting a composite of demand, success rate, and scenario value. The distribution (128 S-tier, 383 A-tier, 655 B-tier, 115 C-tier) is itself a useful signal: a small set of premier sites, a broad productive middle, and a thin low-priority fringe.

The scenario × language reference set. A companion ranking dataset orders the best target domains for 19 distinct use cases across 7 languages (English, German, Spanish, French, Japanese, Korean, Italian). This is the basis for the practical reference matrix in the Appendix.

A full discussion of caveats appears in Section 10. The short version: this is a large but specific lens, one platform's delivery footprint, not a census of all human scraping activity. Read the directional patterns, not the third decimal place.

3. Finding 1: A Heavy Head Over a Long, Active Tail

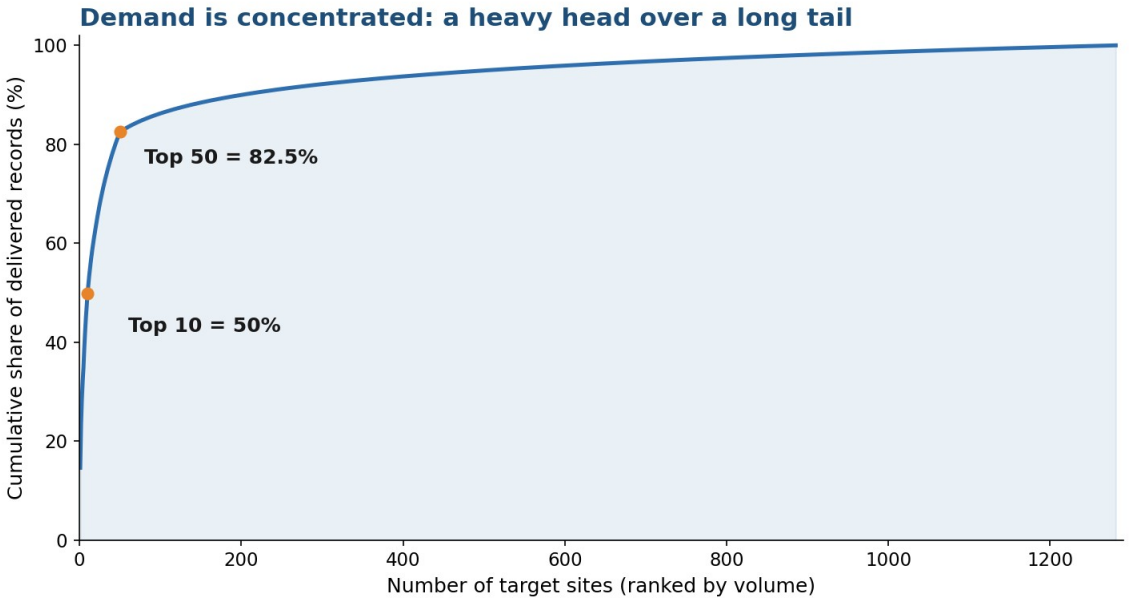

Figure 1. Cumulative share of delivered records by ranked target site.

Web-data demand is concentrated, but it is not winner-take-all. Measured by delivered records, the distribution has a heavy head and a long, genuinely productive tail:

Concentration metric	Share of all delivered records
Top 5 sites	~34.8%
Top 10 sites	~49.8%
Top 20 sites	~63.1%
Top 50 sites	~82.5%

Ten sites deliver about half of all records; fifty sites deliver more than four-fifths. But the remaining roughly 1,230 sites still account for about one record in six, a tail that is thin per-site yet large in aggregate, and broad in the range of industries and languages it spans.

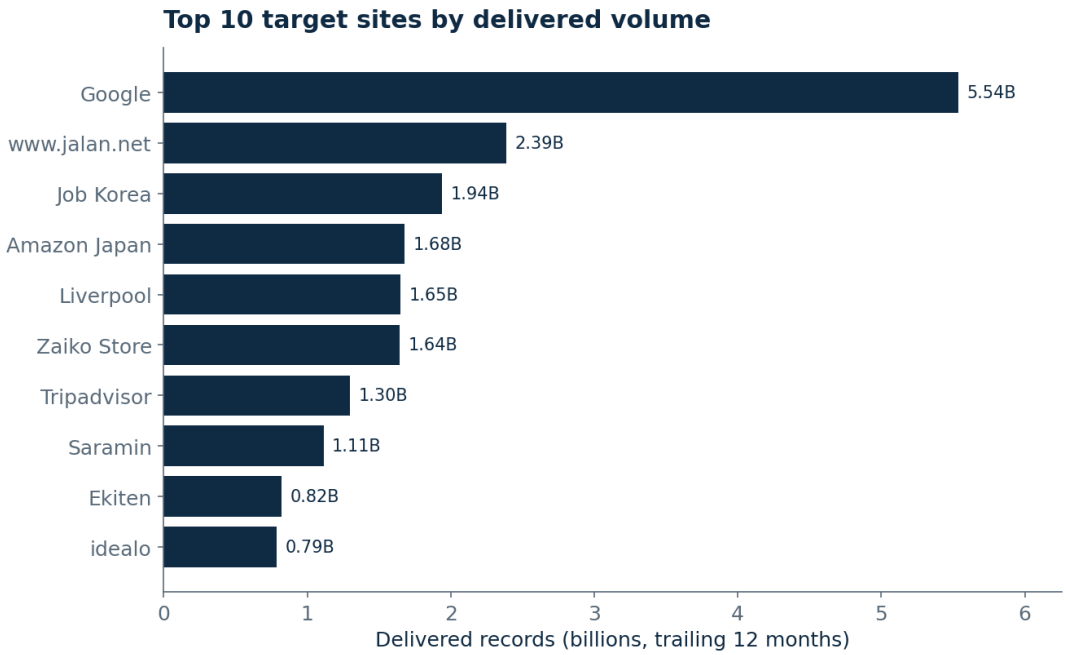

Figure 2. The ten highest-volume target sites, including single-customer power-user sources.

The head itself is revealing. The single largest source by delivered volume is Google (about 5.5 billion records, 12,000+ distinct users): search and maps data underpinning SEO, local-business, and market-research workflows for an enormous base. But right behind it sit sources that look nothing like the usual "top platforms" list. The table below lists the broad-demand head[2]:

Site	Industry	Delivered records (≈)	Success rate
Google	Search engine	5.54B	96.1%
jalan.net	Travel / Hotels (JP)	2.39B	99%
Job Korea	Recruitment (KR)	1.94B	100%
Amazon Japan	E-commerce (JP)	1.68B	99.3%
Liverpool	E-commerce (MX)	1.65B	100%
Tripadvisor	Travel / Hotels	1.30B	100%
Saramin	Recruitment (KR)	1.11B	99%
idealo	E-commerce (price comparison)	0.79B	100%
Booking.com	Travel / Hotels	0.69B	98.9%
Tabelog	Food / Dining (JP)	0.61B	100%

Why this matters. Concentration explains the competitive structure of the industry. Because a small head drives most volume, every vendor builds dedicated, hardened scrapers for those sites first, which is why the head is over-served and the tail under-served. But notice what the head actually contains: Japanese travel, Korean recruitment, Mexican and German e-commerce. The biggest

[2]Figure 2 plots the literal ten highest-volume sites, which include two single-customer power-user sources (Zaiko Store, 1.64B; Ekiten, 0.82B). The table here lists the broad-demand head and holds those two back; both are discussed under "power-user niches" below.

demand is not only the famous US platforms. It is deeply regional and multilingual, a theme that recurs throughout this report.

Two kinds of high-volume site: broad demand vs. power-user niches

Look closer and the head splits into two very different shapes. Some sites are scraped by many people: Google (12,000+ distinct users), YouTube, Amazon, X, TikTok, LinkedIn, and Booking, broad mainstream demand where thousands of teams want the same kind of data. Others post enormous volume from a handful of heavy users: Zaiko Store (1.64B records from just 2 users), Ekiten (0.82B), Namesdir (521M from 1), zalando.it (375M from 7), and Daangn (352M from 4). These are power-user niches, usually e-commerce or property, where one or two operators run industrial-scale, recurring jobs against a single source. Figure 2 shows the literal top 10 by delivered volume (which includes Zaiko Store and Ekiten); the table above lists the broad-demand head, holding back these single-customer sites so they can be discussed here.

For a data collector, the distinction is a map of opportunity. The broad-demand sites are crowded and commoditized; the power-user niches show that a single well-chosen source, scraped deeply and repeatedly, can be worth as much as a famous platform. The volume is not only where the crowd is.

4. Finding 2: The Use-Case Hierarchy

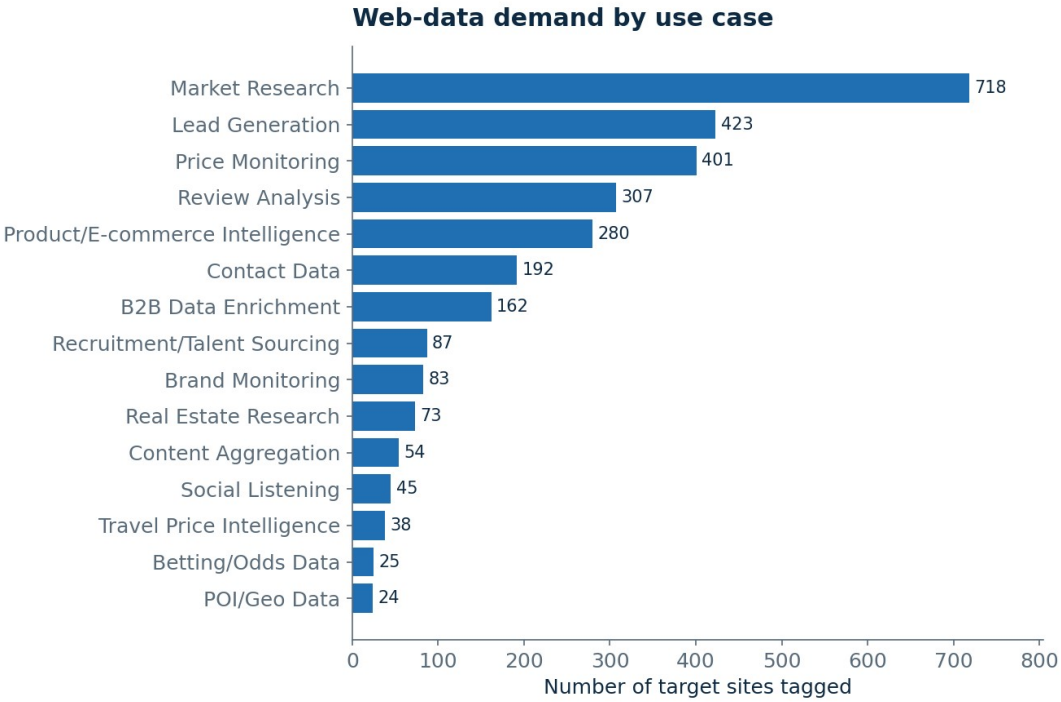

Figure 3. Relative web-data demand by use case (count of tagged target sites).

When every target site is tagged by the purpose the data serves, a clear hierarchy of intent emerges. This is the most decision-relevant view in the dataset, because it describes why anyone scrapes at all.

Rank	Use case	Relative demand (site-tag frequency)
1	Market Research	718
2	Lead Generation	423
3	Price Monitoring	401
4	Review Analysis	307
5	Product / E-commerce Intelligence	280
6	Contact Data	192
7	B2B Data Enrichment	162
8	Recruitment / Talent Sourcing	87
9	Brand Monitoring	83
10	Real Estate Research	73
11	Content Aggregation	54
12	Social Listening	45
13	Travel Price Intelligence	38
14	Betting / Odds Data	25

Rank	Use case	Relative demand (site-tag frequency)
15	POI / Geo Data	24
16	Financial / Investment Research	21

Three things stand out.

First, this is a decision-support market, not a data-theft market. The top five purposes (market research, lead generation, price monitoring, review analysis, e-commerce intelligence) are all about understanding a market to make a commercial decision. The popular framing of scraping as adversarial “data exfiltration” badly misdescribes what the demand actually is: businesses trying to see their own competitive landscape clearly.

Second, sales and marketing quietly rival research. Combine Lead Generation, Contact Data, and B2B Data Enrichment and you get a sales-intelligence cluster that, in aggregate, rivals market research as the largest reason people pull web data. This is the engine behind the directory, company-data, and professional-profile demand visible across the dataset.

Third, commerce is everywhere. Price Monitoring, Product/E-commerce Intelligence, Review Analysis, and Travel Price Intelligence together form a commerce-analytics supercluster. Retail and travel pricing are dynamic, high-frequency, and competitively sensitive, exactly the conditions that make recurring, scheduled extraction valuable rather than one-off pulls.

5. Finding 3: The Industry Mix

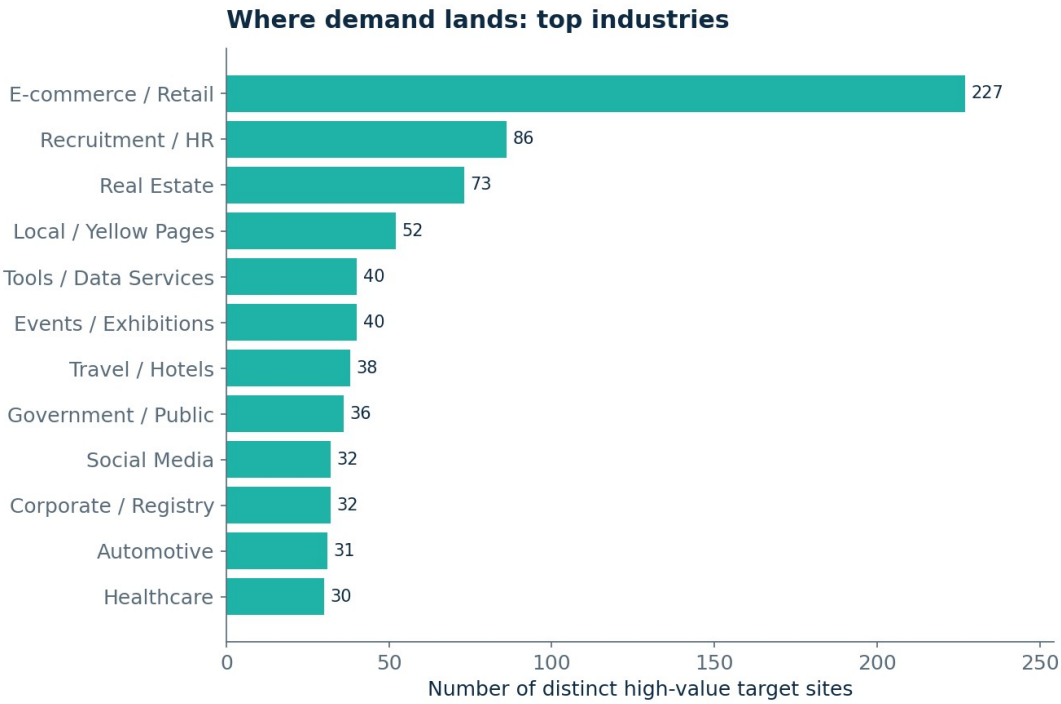

Figure 4. Leading industries by number of distinct high-value target sites (excludes "Other").

Classifying target sites by the industry they belong to shows where the demand physically lands. Excluding a large, deliberately broad "Other" bucket (a reminder of how diffuse the long tail is), the leading industries are:

Industry	Target sites in dataset
E-commerce / Retail	227
Recruitment / HR	86
Real Estate	73
Local Services / Yellow Pages	52
Tools / Data Services	40
Events / Exhibitions	40
Travel / Hotels	38
Government / Public Data	36
Social Media / Communities	32
Corporate / Business Registry Data	32
Automotive	31
Healthcare	30

E-commerce is the gravitational center: it has about 2.6 times as many distinct high-value target sites (227) as the next industry (recruitment, 86). This aligns perfectly with the use-case hierarchy:

price monitoring and product intelligence need many retail sources, not just a few giant ones. (See our e-commerce extraction templates for the most-requested retail sources.)

The more interesting signal is in the middle of the table. Recruitment, Real Estate, Local Services, Events, and Corporate/Registry data are all substantial, structurally similar (large catalogs of semi-structured listings), and far less “platform-shaped” than social media. These are the industries where the long tail lives, where demand is real but no single dominant site exists, and where ready-made templates are least likely to already cover the exact site a given user needs.

6. Finding 4: The Template Gap

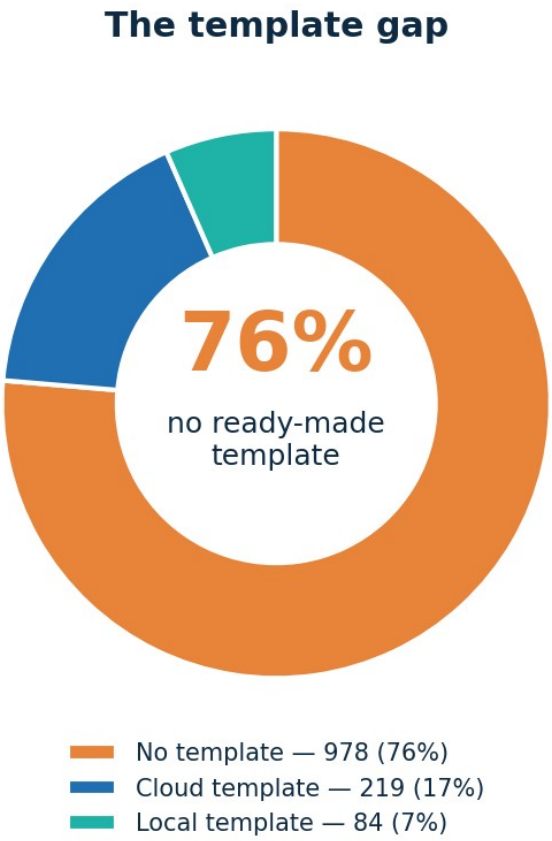

Figure 5. Template coverage across 1,281 high-value sites.

This is the finding with the sharpest strategic edge, and it comes straight from the delivery data. Every one of the 1,281 high-value target sites is flagged for whether a ready-made extraction template exists. The result:

Template status	Sites	Share
No ready-made template	978	76.3%
Cloud template	219	17.1%
Local template	84	6.6%

More than three out of four high-value, actively-requested sites have no off-the-shelf template. These are not obscure pages nobody wants. By construction, every site in this dataset is one people are actually pulling data from. Yet for the clear majority, the data was delivered without a pre-built template to lean on.

The interpretation is not "the catalog is too small." It is that the demand surface is structurally larger and faster-moving than any template library can cover. New sites appear, layouts change, and regional long-tail demand fragments across thousands of destinations. A curated catalog will always trail real demand, not because anyone is failing to maintain it, but because the target is unbounded.

This has a direct consequence for buyers and builders alike:

- **For buyers,** it means a tool chosen primarily on the size of its template library will, for any non-trivial tail use case, frequently not have what you need pre-built. Catalog breadth is a weaker differentiator than it appears.
- **For builders and vendors,** it means the durable advantage is shifting away from catalog size toward generality: the ability to extract from an arbitrary site on demand, whether or not a template for it has ever existed.

This is precisely the gap Octoparse's custom workflow builder is built to close. When no pre-built template exists (the 76% case), users can configure their own extraction visually, pointing and clicking through pagination, detail pages, and field selection on any site, rather than waiting for the catalog to add it. In other words, the template library handles the head, and the workflow builder absorbs the long tail the catalog can never fully reach.

That shift, from "do we have a template for this site?" to "can the tool handle any site you point it at?", is the central product story of the next few years.

7. Finding 5: The Multilingual Blind Spot (and the Reliability Surprise)

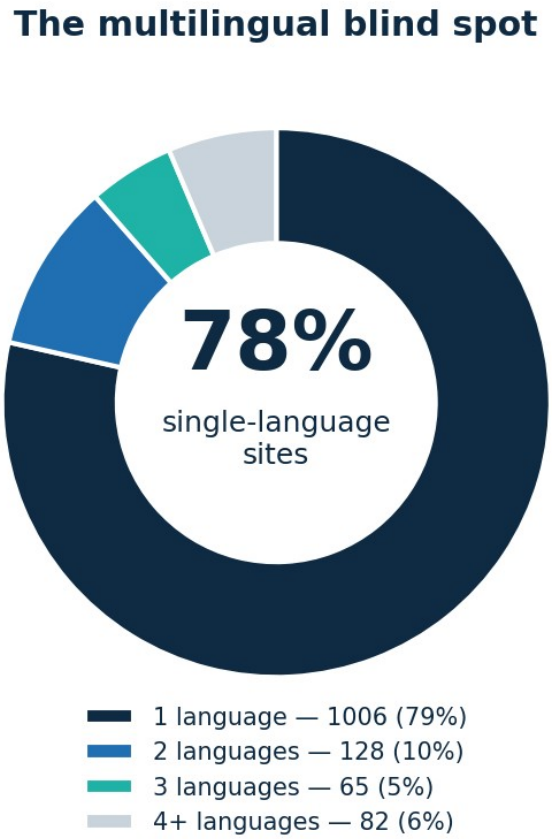

Figure 6. Number of languages each high-value target site is served in.

Web-data tooling is overwhelmingly built and marketed in English, but demand is not. The language distribution of high-value target sites is stark:

Languages a site is served in	Share of target sites
1 language only	~78.5% (1,006 of 1,281)
2 languages	~10.0%
3 languages	~5.1%
4–7 languages	~6.4%

Most high-value targets live in a single-language world. And critically, as Finding 1 already hinted, some of the highest-volume individual sites among these top targets are non-English. Among the very top sources by delivered records:

- jalan.net (Japanese travel): about 2.39 billion records
- Job Korea and Saramin (Korean recruitment): about 1.94B and 1.11B records
- Amazon Japan: about 1.68 billion records

- Tabelog (Japanese restaurant reviews), SUUMO (Japanese real estate), Ekiten (Japanese local listings): each in the hundreds of millions

The scenario-by-language reference work (see Appendix) spans seven languages (English, German, Spanish, French, Japanese, Korean, and Italian) across 19 distinct use cases. The non-English segments are not an afterthought; for travel, recruitment, local services, and regional e-commerce, they are frequently where the volume is.

Why this is a blind spot. A tool whose value rests on a curated, mostly-English template library is structurally weakest exactly where some of the largest demand sits: Japanese, Korean, and continental-European long-tail sites that no English-first catalog prioritizes. The multilingual web is both large and under-tooled, a textbook underserved market. Notably, the delivery data behind this report already spans all seven languages, including the high-volume Japanese and Korean sources above. We cover this segment rather than treating it as an edge case, which is part of why these non-English sites rank so highly by delivered volume.

The reliability surprise: non-English is not harder, it is easier

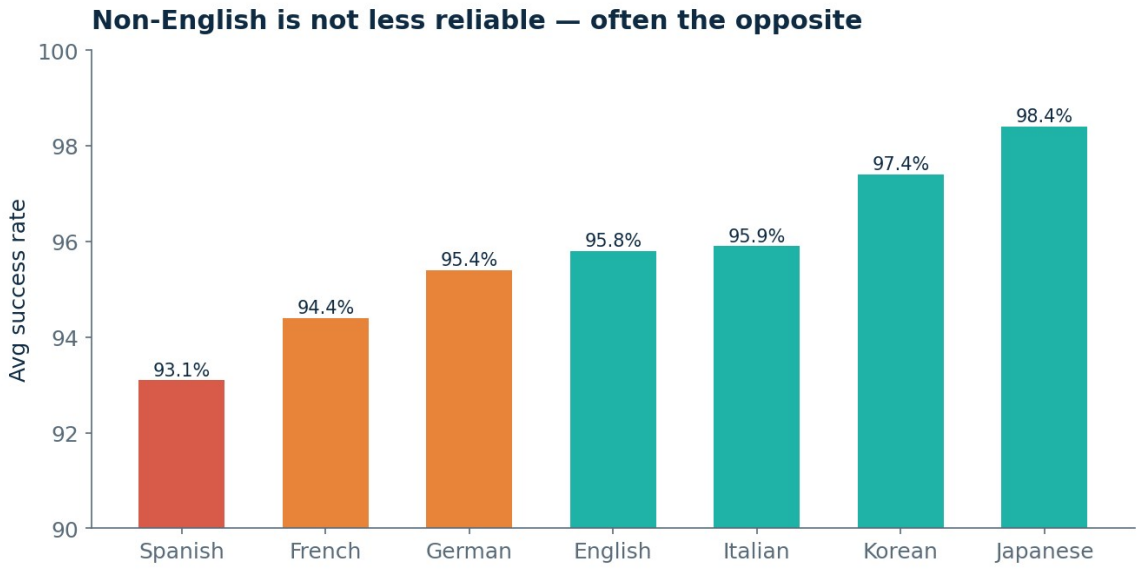

Figure 7. Average extraction success rate by language; non-English leads.

Here is the part that surprised even us. The common assumption is that non-English, non-Latin-script sites are harder to extract. In our data, the opposite holds. Average extraction success rate by language runs:

Language	Avg success rate
Japanese	98.4%
Korean	97.4%
Italian	95.9%
English	95.8%
German	95.4%
French	94.4%
Spanish	93.1%

Japanese and Korean sources are the most reliable to collect in the entire dataset; Spanish and French sit at the bottom. The likely reason is structural, not linguistic: many top Japanese and Korean targets (jalan.net, SUUMO, Job Korea) are stable, well-structured listing sites, while the Spanish- and French-language long tail is more fragmented across many smaller, messier pages.

What this means for you. If your roadmap includes Japanese or Korean market data, do not price in a “hard language” penalty. By our numbers those sources are more dependable than the English baseline. The real difficulty signal is the type of site and its anti-bot posture (Finding 6), not the language it is written in.

8. Finding 6: How Hard Is the Web to Scrape, Really?

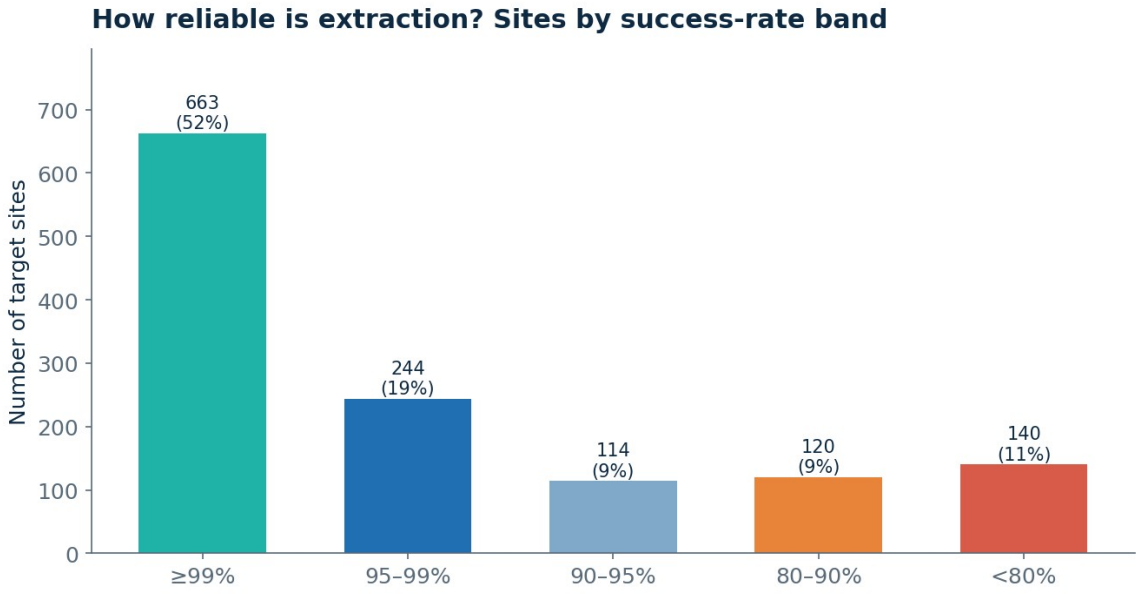

Figure 8. Top sites grouped by extraction success-rate band.

Most of this report is about demand. This finding is about difficulty, and it is the one we are most often asked about directly: “We want data from site X. Is it actually going to work?” Because we observe 37.9 billion real extraction attempts, we can answer that with evidence instead of a shrug.

The headline: the web is more scrapable than its reputation, but unevenly. Across the 1,281 top sites, the median site extracts at 99% success, and 52% clear 99% or better. But the average hides a real hard tail: about 11% of sites (140 of them) sit below 80% success, where anti-bot defenses, heavy JavaScript, or aggressive rate-limiting are actively winning.

Success-rate band	Share of top sites
99% or higher	52%
95–99%	19%
90–95%	9%
80–90%	9%
Below 80%	11%

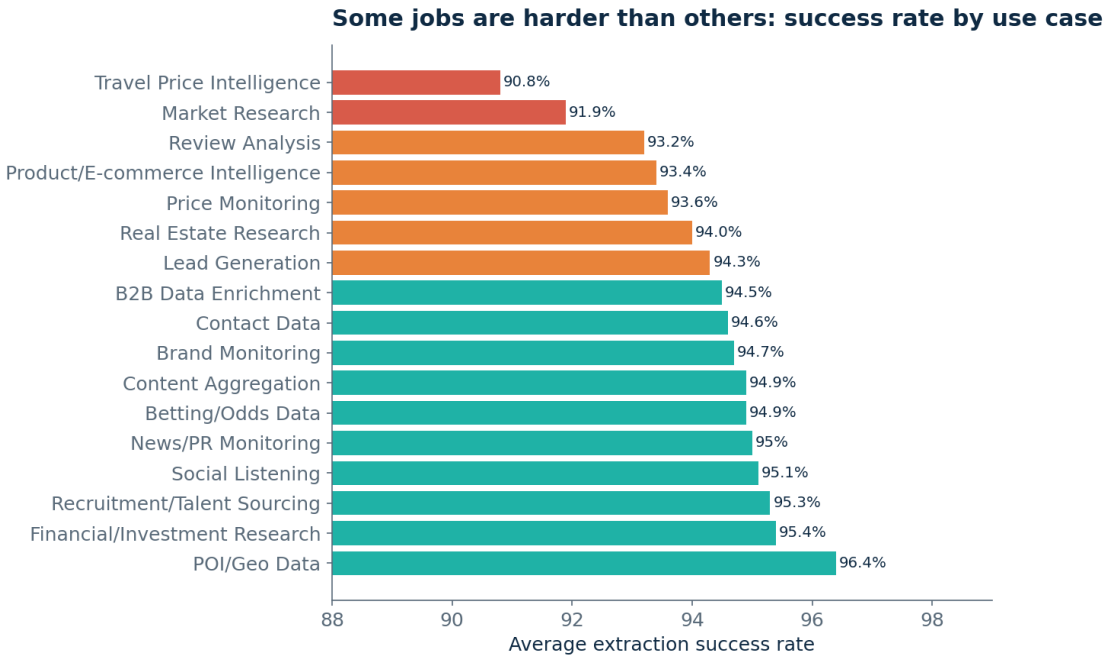

Figure 9. Average extraction success rate by use case (hardest at top).

Difficulty is predictable by use case. The hardest jobs are not random; they cluster where pages are most dynamic and most defended:

- **Hardest:** Travel Price Intelligence (about 91%) and broad Market Research (about 92%): fast-changing prices, heavy interstitials, and sprawling source lists.
- **Middle:** Review Analysis, E-commerce Intelligence, and Price Monitoring (about 93–94%): high volume, frequent layout changes.
- **Most dependable:** POI/Geo Data (about 96%), Financial/Investment Research, Recruitment, and Social Listening (about 95%): comparatively stable, structured sources.

By industry, the pattern echoes: travel/hotels is the toughest vertical (about 91%), while government/public data and tooling/data services are the most reliable (about 96–97%).

A note on templates, handled honestly. One counter-intuitive cut deserves a straight explanation. Sites that have a ready-made template show a slightly lower average success rate (about 91%) than sites with none (about 93%). That looks backwards until you see why: templates get built first for the hardest, highest-demand, most-defended sites (the Googles, the social platforms, the big marketplaces) precisely because those are where users need the most help. So the template population is skewed toward difficult targets. The honest reading is not “templates hurt”; it is “templates exist because a site is hard, and they lift those hard sites much closer to the easy ones than they would otherwise be.”

What this means for you. Use difficulty to set expectations and sequence work. If you are scraping government records, job boards, or geo/POI data, plan for near-total reliability. If you are after travel pricing or running broad market-research sweeps across many sources, budget for retries, monitoring, and a harder maintenance curve, and treat the sub-80% tail as projects that need a resilient, self-correcting approach, not a fragile one-off script.

9. What This Means, and What We'd Tell a Fellow Data Team

Pulling the six findings together yields a coherent picture of where the web-data market is and where it is going. Here is how we read it, and what we would say to a fellow data team scoping their next project.

The head is solved and commoditized. A small set of sites drives roughly half of all volume, and they are served by dedicated, hardened scrapers. Winning here is about reliability, anti-bot resilience, and price, a scale game that favors large infrastructure players. There is little white space left at the head.

The tail and the non-English web are the frontier. The template gap and the multilingual blind spot point to the same conclusion: most distinct demand (by site count, by industry breadth, by language) lives outside the head, and it is poorly served by template-library tools precisely because a catalog can never be comprehensive enough, fast enough, or multilingual enough to cover it. Three-quarters of high-value sites already arrive with no template; that share only grows as you move further down the tail.

This rewards a different kind of product. The industry is bifurcating:

- **Infrastructure / API players** (the Bright Data and Oxylabs tier) win the head and the industrial-scale, high-frequency pipelines: proxies, ready datasets, compliance documentation.
- **Template ecosystems** (such as Octoparse's library) win the "someone already built this" middle, and are genuinely strong wherever a maintained template exists for your exact target.
- **AI-native, plain-language extractors** are emerging to win the tail and the non-technical operator: the long list of one-off, regional, or never-before-scraped sites where no template exists and the user cannot or will not write code.

The compliance overlay is tightening. None of this happens in a legal vacuum. Europe's AI Act (Regulation (EU) 2024/1689), the European Data Protection Board's restrictive stance on scraping personal data for AI training (EDPB Opinion 28/2024), and the US Department of Justice's 2025 rule restricting bulk sensitive-data transfers (DOJ Data Security Program) point directionally toward a clear expectation: data collection should be auditable, transparent, and minimally invasive, with a focus on public data and a clear purpose. This is a general read of the regulatory direction, not legal advice; teams should review the primary texts and their own counsel for specifics. The vendors that thrive will be the ones whose tooling makes responsible, well-scoped, public-data extraction the path of least resistance.

The bottom line for decision-makers. If you buy web data, choose your tool by the shape of your demand, not by headline template counts. Head-platform volume, tail breadth, and language coverage are three different problems with three different best answers. If you build web-data products, the durable moat is moving from how big your catalog is to how general, accessible, and compliant your extraction is. The demand is no longer a secret; this map shows it lives in the long tail, the non-English web, and the hands of non-engineers.

What we'd tell a fellow data team. Three things, plainly. First, do not over-index on the famous platforms; the long tail and the power-user niches are where differentiated data lives, and they are less crowded. Second, set expectations from difficulty, not vibes: government, recruitment, and geo data are near-certain; travel pricing and broad market sweeps need real engineering and monitoring.

Third, do not write off the non-English web; it is underserved by tooling yet often more reliable to collect. Where we can help is the part this map keeps pointing at: the 76% of demand with no template, in any language, where you need to build your own extraction without writing code or waiting for a catalog. That is the problem Octoparse exists to solve, and it is why we are comfortable publishing the map instead of keeping it.

10. Methodology Notes & Limitations

In the interest of intellectual honesty, the constraints of this analysis:

- **One lens, not a census.** The delivery dataset is large but specific: it reflects one platform's user base and the sites it supports. It does not capture private in-house scraping, custom enterprise pipelines, or activity on tools we did not measure. Patterns are directional.
- **Records measure output volume, not value.** Delivered-record counts indicate how much data flowed, not its commercial worth. A few high-volume sites can dominate raw counts without being the most valuable use cases.
- **Use-case and industry tags are interpretive.** Each site carries human/heuristic labels; a site can legitimately serve several purposes. Treat the hierarchy as relative ordering, not precise share.
- **Template status is a point-in-time flag.** "No template" means none existed for that site during the window; it does not imply the site is impossible to extract, only that it was served without a pre-built one.
- **Two success-rate lenses.** The headline 92.7% is record-weighted across all deliveries; the per-language and per-use-case figures are per-site averages within their groups and cover only the reference languages. Compare like with like.
- **External market figures vary widely.** Published market-size and growth estimates for web scraping and alternative data differ by an order of magnitude across research firms depending on scope and definition. We cite them as range and direction, not as settled fact, and source them to the originating research firm below.
- **Point-in-time snapshot.** The dataset is a snapshot from the trailing year as of mid-2026. The head is stable; the tail churns.

11. Appendix: Scenario × Language Reference Matrix

This is the practical companion to the trends above: a “where to get this data” lookup covering 19 use cases × 7 languages. For each scenario and language, the table lists the top 3 recommended source domains, ranked by a composite recommendation score (0–100) that blends delivery success rate, demand, and scenario fit. A dash (—) means the dataset had no qualifying source for that scenario-language pair. Languages covered: English, German, Spanish, French, Japanese, Korean, Italian.

How to read it: scan down a scenario to your target language, and the listed domains are the highest-confidence sources to start from. Note how the picks shift by language; the same scenario often points to entirely different sites in Japanese or Korean than in English, which is the multilingual pattern from Finding 5 made concrete.

What “language” means here: the language column refers to the scenario/market language a user is targeting, not necessarily the language the site is published in. A high-value source can appear under a market whose users frequently request it even when the site itself is in another language. This is why, for example, Japanese sources such as townwork.net or tabelog.com appear under English: English-language users request them often enough to rank. Read each row as “the best sources for users working in this language,” not “sites written in this language.”

POI/Geo Data

Language	Top 3 sources (recommendation score)
English	tabelog.com (100) · www.ubereats.com (98.9) · maps.app.goo.gl (95.5)
German	maps.app.goo.gl (95.5) · www.falstaff.com (69.3)
Spanish	maps.app.goo.gl (95.5) · maps.google.com (84.6)
French	www.ubereats.com (98.9) · maps.app.goo.gl (95.5) · maps.google.com (84.6)
Japanese	tabelog.com (100) · www.ubereats.com (98.9) · www.hotpepper.jp (95.7)
Korean	maps.app.goo.gl (95.5) · map.kakao.com (92.2) · map.naver.com (92.2)
Italian	maps.app.goo.gl (95.5) · researchmap.jp (60.7)

Recruitment/Talent Sourcing

Language	Top 3 sources (recommendation score)
English	townwork.net (100) · www.linkedin.com (100) · 弁護士ドットコム (99.2)
German	www.linkedin.com (100) · www.upwork.com (90.4) · de.indeed.com (88.8)
Spanish	www.linkedin.com (100) · www.upwork.com (90.4) · www.seek.com.au (87.6)
French	www.linkedin.com (100) · www.upwork.com (90.4) · www.welcometothejungle.com (88.8)
Japanese	townwork.net (100) · doda.jp (99.6) · 弁護士ドットコム (99.2)
Korean	www.jobkorea.co.kr (84.4) · www.saramin.co.kr (82.1) · www.alba.co.kr (76.8)
Italian	www.linkedin.com (100) · www.upwork.com (90.4)

Price Monitoring

Language	Top 3 sources (recommendation score)
English	www.ebay.com (100) · www.jalan.net (100) · jp.mercari.com (100)
German	www.ebay.com (100) · www.amazon.com (100) · www.booking.com (100)
Spanish	www.amazon.com (100) · www.booking.com (100) · www.tripadvisor.com (94.4)
French	www.amazon.com (100) · www.booking.com (100) · www.amazon.co.uk (96.6)
Japanese	www.jalan.net (100) · beauty.hotpepper.jp (100) · jp.mercari.com (100)
Korean	www.amazon.co.jp (100) · www.amazon.com (100) · www.booking.com (100)
Italian	www.booking.com (100) · shopping.google.com (92.1) · www.amazon.it (87.9)

Product/E-commerce Intelligence

Language	Top 3 sources (recommendation score)
English	www.ebay.com (100) · jp.mercari.com (100) · www.amazon.co.jp (100)
German	www.ebay.com (100) · www.amazon.com (100) · www.amazon.de (98.5)
Spanish	www.amazon.com (100) · shopping.google.com (92.1) · www.ikea.com (90.6)
French	www.amazon.com (100) · www.amazon.co.uk (96.6) · www.amazon.fr (92.4)
Japanese	www.ebay.com (100) · jp.mercari.com (100) · beauty.hotpepper.jp (100)
Korean	www.amazon.co.jp (100) · www.amazon.com (100) · www.coupang.com (92.5)
Italian	shopping.google.com (92.1) · www.amazon.it (87.9) · www.subito.it (85.8)

Review Analysis

Language	Top 3 sources (recommendation score)
English	www.jalan.net (100) · www.ebay.com (100) · jp.mercari.com (100)
German	www.ebay.com (100) · www.amazon.com (100) · www.booking.com (100)
Spanish	www.amazon.com (100) · www.booking.com (100) · www.tripadvisor.com (94.4)
French	www.amazon.com (100) · www.booking.com (100) · www.ubereats.com (98.9)
Japanese	www.jalan.net (100) · tabelog.com (100) · jp.mercari.com (100)
Korean	www.amazon.co.jp (100) · www.amazon.com (100) · www.booking.com (100)
Italian	www.booking.com (100) · shopping.google.com (92.1) · www.amazon.it (87.9)

News/PR Monitoring

Language	Top 3 sources (recommendation score)
English	news.google.com (91.1) · www.bbc.com (80.8) · www.detik.com (80.6)
German	www.rnd.de (70.1) · www.tribunnews.com (70) · news.naver.com (61.8)
Spanish	www.vogue.com (71.9) · news.naver.com (61.8)
French	lespepitestech.com (72.5) · www.computerworld.com (71.8) · industrie.usinenouvelle.com

Language	Top 3 sources (recommendation score)
	(71.3)
Japanese	prtimes.jp (96.1) · news.yahoo.co.jp (87.9) · www.nikkei.com (78.5)
Korean	—
Italian	www.bbc.com (80.8) · www.ansa.it (66) · news.naver.com (61.8)

Social Listening

Language	Top 3 sources (recommendation score)
English	www.tiktok.com (100) · x.com (100) · www.youtube.com (100)
German	x.com (100) · www.tiktok.com (100) · www.youtube.com (100)
Spanish	www.tiktok.com (100) · x.com (100) · www.youtube.com (100)
French	www.tiktok.com (100) · x.com (100) · www.youtube.com (100)
Japanese	www.tiktok.com (100) · x.com (100) · www.youtube.com (100)
Korean	www.tiktok.com (100) · x.com (100) · www.youtube.com (100)
Italian	www.tiktok.com (100) · x.com (100) · www.youtube.com (100)

Market Research

Language	Top 3 sources (recommendation score)
English	www.google.com (98.2) · tabelog.com (91.8) · townwork.net (88.4)
German	www.google.com (98.2) · www.koheimusic.com (84.4) · maps.app.goo.gl (83.5)
Spanish	www.google.com (98.2) · scholar.google.com (85) · www.bing.com (84.7)
French	www.google.com (98.2) · www.ubereats.com (86.9) · scholar.google.com (85)
Japanese	www.google.com (98.2) · tabelog.com (91.8) · townwork.net (88.4)
Korean	www.google.com (98.2) · scholar.google.com (85) · maps.app.goo.gl (83.5)
Italian	www.google.com (98.2) · scholar.google.com (85) · www.bing.com (84.7)

Contact Data

Language	Top 3 sources (recommendation score)
English	www.yellowpages.com (99.9) · www.gelbeseiten.de (99.6) · baseconnect.in (99.2)
German	www.gelbeseiten.de (99.6) · www.wlw.de (98.2) · tel.search.ch (95.7)
Spanish	hipages.com.au (92.6) · www.paginasamarillas.es (89.6) · www.mwcbarcelona.com (88.8)
French	www.pagesjaunes.fr (95.9) · tel.search.ch (95.7) · www.local.ch (94.6)
Japanese	baseconnect.in (99.2) · www.ekiten.jp (90.6) · itp.ne.jp (88)
Korean	—
Italian	www.paginegialle.it (94.7) · www.local.ch (94.6) · www.paginebianche.it (84.9)

SEO/SERP Tracking

Language	Top 3 sources (recommendation score)
English	www.google.com (100) · www.bing.com (96.7) · search.yahoo.com (84.1)
German	www.google.com (100) · www.google.de (87.1) · start.duckduckgo.com (79)
Spanish	www.google.com (100) · www.bing.com (96.7) · search.yahoo.com (84.1)
French	www.google.com (100) · www.bing.com (96.7) · duckduckgo.com (86.2)
Japanese	www.google.com (100) · www.bing.com (96.7) · search.yahoo.co.jp (95.8)
Korean	www.google.com (100) · search.naver.com (80.4) · www.naver.com (78.3)
Italian	www.google.com (100) · www.bing.com (96.7) · google.com (82.2)

Content Aggregation

Language	Top 3 sources (recommendation score)
English	www.youtube.com (100) · scholar.google.com (97) · www.koheimusic.com (96.4)
German	www.youtube.com (100) · www.koheimusic.com (96.4) · patents.google.com (90.9)
Spanish	www.youtube.com (100) · scholar.google.com (97) · www.koheimusic.com (96.4)
French	www.youtube.com (100) · scholar.google.com (97) · www.koheimusic.com (96.4)
Japanese	www.youtube.com (100) · scholar.google.com (97) · prtimes.jp (96.1)
Korean	www.youtube.com (100) · scholar.google.com (97) · youtu.be (91.1)
Italian	www.youtube.com (100) · scholar.google.com (97) · www.koheimusic.com (96.4)

Financial/Investment Research

Language	Top 3 sources (recommendation score)
English	coinmarketcap.com (91.1) · brokersnapshot.com (87.8) · finance.yahoo.com (82.2)
German	www.dvag.de (83.5) · www.investing.com (75.2) · stock.finance.sina.com.cn (62.8)
Spanish	www.coldwellbanker.com (81.4) · www.investing.com (75.2) · global.morningstar.com (66.8)
French	emma.msrb.org (55.4)
Japanese	finance.yahoo.co.jp (80.8) · kabutan.jp (80.3)
Korean	—
Italian	www.centroquote.it (67.7) · it.finance.yahoo.com (66) · www.mondialbroker.com (62.4)

Travel Price Intelligence

Language	Top 3 sources (recommendation score)
English	www.jalan.net (100) · www.booking.com (100) · www.agoda.com (96)
German	www.booking.com (100) · flights.booking.com (74.9) · booking.nautaliaviajes.com (67.9)
Spanish	www.booking.com (100) · www.tripadvisor.com (94.4) · www.tripadvisor.co.uk (92.1)

Language	Top 3 sources (recommendation score)
French	www.booking.com (100) · www.tripadvisor.co.uk (92.1) · www.tripadvisor.fr (73.4)
Japanese	www.jalan.net (100) · www.booking.com (100) · travel.rakuten.co.jp (80.4)
Korean	www.booking.com (100) · www.agoda.com (96) · www.yeogi.com (73.3)
Italian	www.booking.com (100) · www.tripadvisor.it (67.8) · hotels.ctrip.com (53.4)

Lead Generation

Language	Top 3 sources (recommendation score)
English	www.linkedin.com (100) · townwork.net (100) · www.yellowpages.com (99.9)
German	www.linkedin.com (100) · www.gelbeseiten.de (99.6) · www.google.com (98.2)
Spanish	www.linkedin.com (100) · www.google.com (98.2) · maps.app.goo.gl (95.5)
French	www.linkedin.com (100) · www.google.com (98.2) · www.pagesjaunes.fr (95.9)
Japanese	townwork.net (100) · doda.jp (99.6) · baseconnect.in (99.2)
Korean	www.google.com (98.2) · maps.app.goo.gl (95.5) · map.kakao.com (92.2)
Italian	www.linkedin.com (100) · www.google.com (98.2) · maps.app.goo.gl (95.5)

Real Estate Research

Language	Top 3 sources (recommendation score)
English	www.zillow.com (92.6) · www.homes.com (91.5) · www.realtor.com (90.1)
German	www.homes.com (91.5) · www.immowelt.de (90.1) · www.immobilienscout24.de (84.7)
Spanish	www.homes.com (91.5) · www.idealista.com (89.4) · www.rightmove.co.uk (81.3)
French	www.seloger.com (87.3) · www.safti.fr (82.1) · www.rightmove.co.uk (81.3)
Japanese	suumo.jp (100) · www.athome.co.jp (86.5) · www.homes.co.jp (84.8)
Korean	new.land.naver.com (86.8) · m.land.naver.com (70.1) · fin.land.naver.com (68.2)
Italian	www.immobiliare.it (93.5) · www.idealista.it (79.9) · www.airbnb.it (68.9)

Brand Monitoring

Language	Top 3 sources (recommendation score)
English	x.com (100) · www.tiktok.com (100) · www.trustpilot.com (97.7)
German	www.tiktok.com (100) · x.com (100) · old.reddit.com (97.4)
Spanish	x.com (100) · www.tiktok.com (100) · www.youtube.com (92.5)
French	x.com (100) · www.tiktok.com (100) · www.trustpilot.com (97.7)
Japanese	x.com (100) · www.tiktok.com (100) · www.youtube.com (92.5)
Korean	www.tiktok.com (100) · x.com (100) · old.reddit.com (97.4)
Italian	x.com (100) · www.tiktok.com (100) · old.reddit.com (97.4)

B2B Data Enrichment

Language	Top 3 sources (recommendation score)
English	www.yellowpages.com (99.9) · www.gelbeseiten.de (99.6) · baseconnect.in (99.2)
German	www.gelbeseiten.de (99.6) · www.wlw.de (98.2) · tel.search.ch (95.7)
Spanish	hipages.com.au (92.6) · www.linkedin.com (89.8) · www.paginasamarillas.es (89.6)
French	www.pagesjaunes.fr (95.9) · tel.search.ch (95.7) · www.local.ch (94.6)
Japanese	baseconnect.in (99.2) · www.ekiten.jp (90.6) · itp.ne.jp (88)
Korean	—
Italian	www.paginegialle.it (94.7) · www.local.ch (94.6) · find-and-update.company-information.service.gov.uk (91.6)

Betting/Odds Data

Language	Top 3 sources (recommendation score)
English	www.oddsportal.com (90.4) · racing.hkjc.com (84.4) · www.flashscore.com (80.6)
German	www.oddsportal.com (90.4) · www.transfermarkt.co.uk (77.2) · www.forebet.com (75.6)
Spanish	www.oddsportal.com (90.4) · racing.hkjc.com (84.4) · www.flashscore.com (80.6)
French	www.oddsportal.com (90.4) · racing.hkjc.com (84.4) · www.flashscore.com (80.6)
Japanese	—
Korean	—
Italian	www.transfermarkt.co.uk (77.2) · www.forebet.com (75.6) · www.snai.it (70.7)

Academic/Research Data

Language	Top 3 sources (recommendation score)
English	scholar.google.com (97) · patents.google.com (90.9) · www.sciencedirect.com (70.5)
German	patents.google.com (90.9) · www.sciencedirect.com (70.5) · patentscope.wipo.int (69.9)
Spanish	scholar.google.com (97) · www.coursera.org (74.9)
French	scholar.google.com (97) · www.education.gouv.fr (83.5) · fr.wikipedia.org (82.4)
Japanese	scholar.google.com (97) · scholar.google.co.jp (81.8)
Korean	scholar.google.com (97)
Italian	scholar.google.com (97) · patentscope.wipo.int (69.9)

Scores are relative recommendation ratings within this dataset, not absolute quality measures. Where fewer than three sources are shown, the dataset contained no further qualifying source for that pair.

Sources

External market context (each figure cited to the originating research firm; verified June 2026):

Claim	Primary source
Web-scraping market $1.03B (2025) → $2.23B (2031), 13.78% CAGR; ~65% of enterprises feeding AI/ML with web scraping in 2024	Web Scraping Market, Mordor Intelligence — mordorintelligence.com/industry-reports/web-scraping-market
AI-driven web scraping forecast: +$3.15B (2024–2029), 39.4% CAGR	AI-Driven Web Scraping Market, Research and Markets — researchandmarkets.com/reports/6115359
Training-related crawling as the majority of AI crawler traffic in 2025	Cloudflare Radar 2025 Year in Review — radar.cloudflare.com/year-in-review/2025
Web-scraped data the largest single category of alternative data (~15% of spend); continued double-digit annual growth	Neudata — neudata.co/education/how-big-is-the-alternative-data-market-for-investment-managers
EU AI Act (Regulation (EU) 2024/1689)	EUR-Lex — eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
EDPB stance on personal data and AI models	EDPB Opinion 28/2024 — edpb.europa.eu
US DOJ 2025 rule on bulk sensitive-data transfers	DOJ National Security Division, Data Security Program — justice.gov/nsd/data-security

Primary data: aggregated trailing-twelve-month delivery telemetry across 1,281 target sites, plus a companion 19-scenario × 7-language ranking set, compiled from internal platform records. All internal figures in this report (37.9B delivered records, 92.7% success rate, 76% no-template share, language and use-case distributions, top-site volumes) are computed directly from that workbook; because the underlying dataset is proprietary, these figures are not externally linkable.