



CENSUS · BA-D-2026-02

# International-readiness of university homepages: a raw-HTML census of hreflang, canonical, and language signals

---

University homepages compete for internationally mobile students who increasingly research destinations through AI assistants, yet the international-targeting layer that tells a crawler which language and regional variants of a page exist is frequently absent at the front door.

|          |                                    |
|----------|------------------------------------|
| DOCUMENT | BA-D-2026-02                       |
| SERIES   | Census                             |
| DATE     | 2026                               |
| SUBJECTS | Crawlers & access; Market behavior |
| LICENSE  | CC-BY-4.0                          |
| AUTHOR   | Barkhausen AI                      |

## ABSTRACT

University homepages compete for internationally mobile students who increasingly research destinations through AI assistants, yet the international-targeting layer that tells a crawler which language and regional variants of a page exist is frequently absent at the front door. This census measures that layer directly. Using the raw, unrendered HTML served to a non-JavaScript crawler, it examines the homepages of a frame of 500 universities — the 300 largest U.S. institutions by enrollment plus 200 international universities by traffic rank — of which 429 returned an analyzable page on 2026-07-09. In that HTML, 84.1% (361 of 429) declared no hreflang alternates at all. Among the 68 that did, quality was high: no invalid language codes and 89.7% carrying a self-reference. Canonical and language-tag hygiene showed the same adoption-not-correctness pattern. Findings describe raw HTML on a single day; sites change without notice, and results should be assumed perishable.

## IN BRIEF

When a prospective international student asks an AI assistant where to study, the assistant's answer is built from what a crawler could read on the institution's website — and most crawlers behind AI answers do not run JavaScript, so they see only the raw HTML the server sends. This census reads that raw HTML for a frame of 500 universities and asks a narrow question: does the homepage carry the standard signals that tell a machine which language and regional versions of the page exist, and which one is authoritative? For 429 homepages that returned a readable page on 2026-07-09, the answer is mostly no. More than four in five — 361 of 429 — carried no hreflang alternates at all, the tag family whose entire job is to declare international variants. The institutions that did use it used it well: no invalid language codes, and nearly nine in ten pointing a variant back at themselves. Canonical links and language attributes told the same story — widely, correctly used where present, simply absent elsewhere. The gap is adoption, not correctness. This is a homepage snapshot in server-delivered HTML on one day; it is not a verdict on any named institution, and the open dataset lets any reader inspect every row.

## KEY FINDINGS

1. In the raw HTML served to non-rendering crawlers on 2026-07-09, 84.1% of analyzed university homepages (361 of 429) declared no hreflang alternates — the international-targeting layer was absent at the homepage for more than four in five institutions, in a frame deliberately built to over-represent internationally active universities.
2. Among the 68 homepages that did declare hreflang alternates, the declarations were well-formed: 0 of 68 carried an invalid language code under a conservative structural check, and 61 of 68 (89.7%) included a self-reference — so the shortfall in international targeting is one of adoption, not correctness.
3. A canonical link was present in 304 of 429 homepages (70.9%); of those 304, 246 pointed to the homepage itself and 58 pointed to a different URL, and one homepage declared more than one canonical — a class of ambiguity a non-rendering crawler must resolve without operator guidance.
4. Language tagging was near-universal — an html lang attribute was present on 416 of 429 homepages (97.0%) and syntactically valid on 415 — while high-confidence language-content mismatches were rare at 6 of 429 (1.4%).
5. Six of 429 analyzed homepages (1.4%) carried a noindex directive in their raw HTML or response headers, instructing compliant crawlers not to index the front page — a self-removal defect reported here as a count only; the open dataset carries the affected rows.

## CONTENTS

|   |                                                             |   |
|---|-------------------------------------------------------------|---|
| 1 | Summary                                                     | 4 |
| 2 | Scope and method                                            | 4 |
| 3 | Results — international targeting                           | 5 |
| 4 | Results — canonical and language hygiene                    | 6 |
| 5 | What this means for institutions recruiting internationally | 7 |
| 6 | Reproducibility appendix                                    | 8 |
| 7 | Limitations                                                 | 8 |

## 1 Summary

Prospective international students increasingly research destinations through AI assistants rather than through search engines or printed prospectuses. The companion report BA-R-2026-01 reviews that evidence: it documents a steep, replicated rise in the United States and rising research-stage adoption among Chinese and other international students in the share of prospective students who consult AI tools. Whether an institution's international content reaches those assistants depends on plumbing most readers never see — the hreflang, canonical, and language signals that tell a machine which language and regional variants of a page exist, and which one is authoritative. In the maturity model of BA-C-1, these are hygiene items at the base of the ladder — search hygiene (BL-0), where canonicalization and the language declaration sit, through rendering and crawler access (BL-1) and structured-data consistency (BL-2); an entity that fails them is not reliably legible regardless of what sits above.

This census measures those signals directly, in the raw HTML that a non-rendering crawler receives. It reads the homepages of a frame of 500 universities — the 300 largest U.S. institutions by enrollment and 200 international universities by traffic rank — of which 429 returned an analyzable page on 2026-07-09. The headline is an absence: 84.1% of those homepages (361 of 429) declared no hreflang alternates at all. For a set of institutions selected to over-represent international activity, the tag family whose sole purpose is to declare international variants is simply missing at the front door. Where the international-targeting and hygiene signals were present, they were mostly used correctly — no invalid language codes among the adopters, near-universal language attributes, and a clean self-reference in the large majority of hreflang declarations. The gap this census finds is one of adoption, not correctness. The measurement is a single-day snapshot of server-delivered HTML; sites change without notice, and results should be assumed perishable.

## 2 Scope and method

### The frame

The population is a frame, not a census of all universities, and its construction is documented so a reader can re-derive it. The U.S. portion (300 rows) is the largest U.S. institutions by latest reported enrollment, drawn from the U.S. Department of Education's College Scorecard [1] and ordered by enrollment descending; this ordering surfaces large online and for-profit institutions first. The international portion (200 rows) is non-U.S. universities from the Hipolabs university-domains-list [2], restricted to domains present in the Tranco research ranking (list ID 46XZX, a manipulation-hardened traffic-rank aggregate) [3] and ordered by traffic rank. The two orderings differ — enrollment for the U.S. set, traffic rank for the international set — so the frame is a deliberate U.S.-plus-international mix, not a uniform global sample. The mix skews toward larger and higher-traffic institutions within each source, which is appropriate for a legibility census (crawlers reach trafficked sites) but is not a random draw. Registrable domains were computed with the Public Suffix List; the frame is disjoint from the other sector frames collected in the same window.

Of the 500 domains, 429 returned an analyzable homepage — an HTTP 200 with a non-empty body — and are the denominator for every readiness figure below. The remaining 71 were not analyzed: 33 returned a non-200 status (all 33 were 403, a bot-challenge or access-denied response at the homepage), and 38 failed at the transport layer (DNS, TLS, timeout, or other connection failure). These non-analyzed domains are reported separately and excluded from the readiness denominators rather than counted as “missing” signals.

## Raw-HTML-only measurement, and why it is the relevant lens

Every signal in this census is extracted from the raw, unrendered HTML the server returned — the bytes a client receives before any JavaScript runs. This is a deliberate choice, and it is the choice that makes the measurement relevant to AI answer engines. BA-C-1 records the public finding, from analyses of AI crawler behavior, that AI crawlers largely do not execute JavaScript; content that exists only after client-side rendering is, in practice, not seen by them. A non-rendering crawler sees exactly what this census reads. The consequence is that an hreflang, canonical, or language tag injected by client-side JavaScript is invisible here by design. A homepage that declares its international variants only after rendering will read as “no hreflang” in this measurement — which is precisely how it reads to a non-rendering crawler. The direction of this bias is therefore known and one-sided: for engines that do not render, the raw-HTML count is the operative count; for any consumer that does render, these figures undercount the signals present. This measurement targets the former.

## Operationalizations

The measurement definitions are fixed in advance and applied uniformly; the load-bearing ones are as follows.

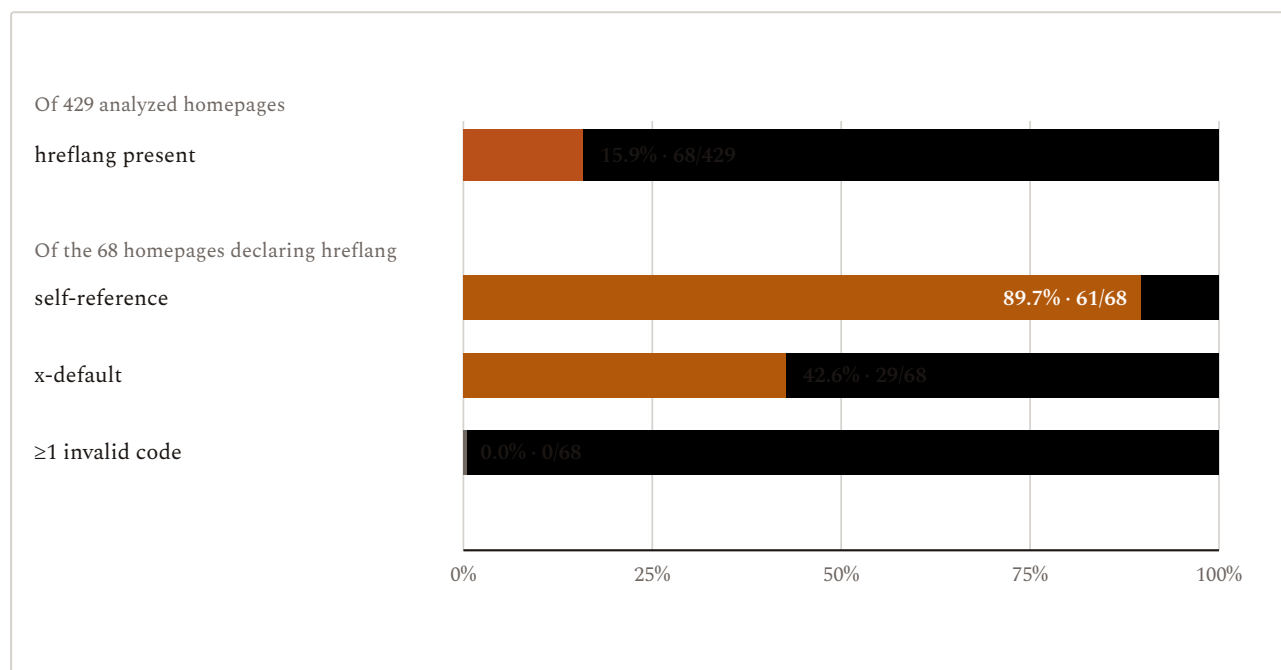
- **hreflang.** A counted alternate is a `<link>` whose `rel` token set includes `alternate` and which carries an hreflang attribute. Language-code validity is screened with a deliberately conservative BCP-47-lite check — the regular expression `^[a-zA-Z]{2,3}(-[a-zA-Z]{2}|-[0-9]{3})?$` plus the literal `x-default`. This lite check correctly rejects malformed tags such as `en_US` (underscore), but it also **rejects otherwise valid tags that carry a script subtag** — `zh-Hant`, `zh-Hans`, `sr-Latn` — and extended or variant tags. Reported invalid-code counts are therefore an upper bound on genuine errors and should be read as “invalid under the lite check,” not “invalid under full BCP 47” [5]. A **self-reference** is an hreflang href that, resolved against the final fetched URL and normalized (lowercase scheme and host, default ports dropped, empty path treated as `/`, a single trailing slash stripped on non-root paths, query preserved, scheme not unified), equals that final URL.
- **canonical.** `canonical_present` marks a `<link rel="canonical">`; `canonical_self_pointing` marks a first canonical whose href, resolved and normalized as above, equals the final URL; `canonical_multiple` marks the defect of more than one canonical link. A canonical that resolves to a different host or path is present but not self-pointing.
- **language.** The `<html lang>` attribute is captured and screened with the same lite check. Visible text is tag-stripped and whitespace-collapsed; below 200 characters, language detection is skipped (`insufficient_text`). Detection uses `langdetect` with a fixed seed. A language-content **mismatch** is flagged only when the text is sufficient, an `html lang` is present, detection probability is at least 0.9, and the `html-lang` primary subtag differs from the detected primary subtag — a conservative rule that treats a false accusation of mislabeling as worse than a miss, so borderline detections never flag.
- **noindex** is set when a `<meta name="robots">` content value contains `noindex`, or an X-Robots-Tag response header does. **site\_moved** marks a final served host that differs from the listed domain; it treats an apex-to-`www` change or any subdomain change as “moved,” so it is an identity note about which host answered, not a defect.

## 3 Results — international targeting

Across the 429 analyzed homepages, hreflang adoption is the low signal and the headline finding. Only 68 homepages (15.9%) declared any hreflang alternate in their raw HTML; the complementary 361 (84.1%),

computed as  $429 - 68$ ) declared none. In a frame built to over-represent internationally active universities, the layer whose explicit purpose is to enumerate language and regional variants is absent from more than four in five front pages.

Where hreflang was present, it was used well. Among the 68 adopters, 0 (0.0%) carried a language code that failed even the conservative lite check, and 61 (89.7%) included a self-reference — the recommended practice of having the set of alternates list the page itself. An x-default alternate, which names a fallback for unmatched locales, appeared on 29 of the 68 (42.6%). The contrast between the two levels is the substantive result: adoption is the binding constraint, not correctness. Institutions that reached for the tag largely configured it in a well-formed way; the large majority never reached for it at all.



**Figure 1.** International targeting in raw homepage HTML. Top: share of the 429 analyzed homepages declaring any hreflang alternate. Bottom: among the 68 homepages that declared hreflang, the share carrying a self-reference, an x-default, and at least one invalid language code (conservative lite check). The two panels use different denominators, each stated on the panel.

BA-D-2026-02, dataset university-intl-readiness-2026 · n = 429 analyzed (68 hreflang adopters) · 2026-07-09

## 4 Results — canonical and language hygiene

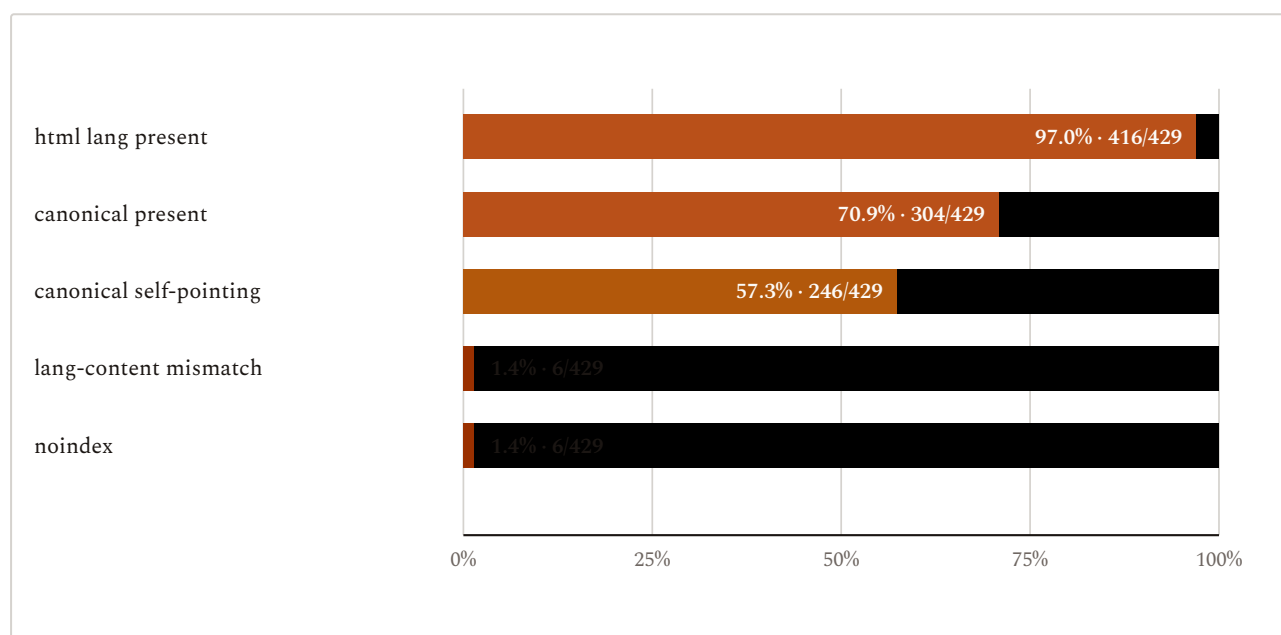
The remaining head-section signals govern which URL a crawler treats as authoritative and in which language it reads the page. They repeat the pattern of the hreflang result: broad presence, mostly correct where present, with a small residue of defects that the open dataset lets a reader inspect row by row.

A canonical link was present in 304 of 429 homepages (70.9%). Of those 304, 246 pointed to the homepage itself and 58 (computed as  $304 - 246$ ) pointed to a different URL. The 58 non-self-pointing canonicals are a mixed class: most point to a different path, or to a different host within the same registrable domain, and a few are cross-domain — a canonical that hands authority to another domain entirely. One of the 429 homepages declared more than one canonical link (0.2%), an ambiguity a crawler must resolve on its own. These are reported as counts and classes only; the dataset carries the specific rows for any reader who wants to see which pattern a given homepage exhibits.

Language tagging was near-universal. An `<html lang>` attribute was present on 416 of 429 homepages (97.0%) and syntactically valid under the lite check on 415. High-confidence language-content mismatches — cases where the detected language of the visible text disagreed with the declared lang, under the

conservative  $p \geq 0.9$  rule — were rare, at 6 of 429 (1.4%). The conservative rule limits false flags without eliminating them — closely related language pairs can trip a primary-subtag comparison — so the six are high-confidence candidates for review, not a settled count of mislabels, and genuine mislabels below the confidence threshold go uncounted. Six of the 429 homepages (1.4%) carried a noindex directive in their raw HTML or response headers, instructing a compliant crawler not to index the front page at all — a self-removal defect on the institution’s most valuable page. Consistent with this census’s reporting rule, the six are reported as a count; the affected domains are rows in the open dataset, not names in this prose.

One high-frequency signal is an identity note rather than a defect: 318 of 429 homepages (74.1%) were served from a host that differs from the listed domain (`site_moved`). Because this flag treats an apex-to-`www` redirect or any subdomain change as “moved,” its high rate mostly reflects the ordinary `example.edu` → `www.example.edu` convention and should not be read as institutions relocating.



**Figure 2.** Canonical and language hygiene across the 429 analyzed homepages, in raw HTML. Bars show the share of the 429 for each signal. `html lang present` was valid under the conservative lite check on 415 of the 416. Language-content mismatch and `noindex` are the two defect classes; both stand at 6 of 429.

BA-D-2026-02, dataset `university-intl-readiness-2026` ·  $n = 429$  analyzed · 2026-07-09

## 5 What this means for institutions recruiting internationally

Absent hreflang does not mean an institution is invisible to AI answer engines, and this census makes no such claim. It means the international-targeting signal that would tell an engine, unambiguously, which language and regional variants of a page exist — and which to serve to which audience — is not being supplied at the homepage in the HTML a non-rendering crawler reads. In its absence an engine must infer language and regional targeting from other cues: the visible text, the `lang` attribute, the URL structure, links to other-language versions. Inference can succeed; it is simply less certain than a declaration, and it puts the outcome in the engine’s hands rather than the institution’s. For a set of institutions selected to over-represent international activity, the finding is that the explicit layer is mostly not present, and the implicit layer is what remains.

Read against BA-R-2026-01, which documents a rising share of prospective students — including in China and other large source markets — consulting AI assistants at the research stage, the two findings sit adjacent: the audience is arriving through machine-mediated answers, and the machine-readable international-targeting layer at the front door is, for most of this frame, absent rather than merely

imperfect. This census describes that state; it does not prescribe a response, evaluate any institution, or recommend any practice. What it establishes is narrow and, within its bounds, well supported: for the homepages measured, in the HTML that non-rendering crawlers see, international targeting is an adoption gap, and the hygiene signals around it are used correctly wherever they are used at all.

## 6 Reproducibility appendix

The frame, the fetch, and the parse are specified to permit replication.

**Frame.** 500 registrable domains: 300 U.S. institutions ordered by latest reported enrollment from the College Scorecard institution-level data [1], and 200 non-U.S. universities from the Hipolabs university-domains-list [2] restricted to and ordered by the Tranco research ranking, list ID 46XZX [3]. Registrable domains were computed with the Public Suffix List; the frame was de-duplicated and made disjoint from the parallel sector frames collected in the same window. The two orderings (enrollment for the U.S. set, traffic rank for the international set) are recorded per row.

**Fetch.** Each homepage was retrieved with GET `https://<domain>/`, following up to five redirects, with a 20-second timeout, over HTTP/1.1, using the user-agent string Mozilla/5.0 (compatible; research-fetch). On a connect-level failure over https, the fetch retried once over http. The exact response bytes and headers, including the full redirect chain, were stored so every derived field can be recomputed offline without re-fetching. Analysis ran only for an HTTP 200 with a non-empty body; a header-level noindex was recorded even on non-200 responses.

**Parse.** Signals were extracted from the raw, unrendered HTML with a tolerant streaming parser that reads `<link>`, `<meta>`, `<title>`, and the `<html lang>` attribute; script, style, template, and head content are excluded from visible text. hreflang alternates, canonical links, the lang attribute, and noindex were operationalized exactly as in the method section above. Language detection used a fixed random seed for determinism, was skipped below 200 characters of visible text, and flagged a mismatch only at detection probability  $\geq 0.9$  with a differing primary subtag. URL comparison for self-reference and canonical self-pointing used the normalization stated above (lowercase scheme and host, default ports dropped, empty path treated as /, a single trailing slash stripped on non-root paths, query preserved, scheme not unified). The collection window was a single pass on 2026-07-09. The full dataset — one row per domain, every column documented — is published alongside this census.

## 7 Limitations

**Homepage only.** This census reads the homepage of each domain and nothing deeper. International targeting frequently lives on inner pages — language-specific landing pages, country microsites, program pages — that a homepage may or may not link with hreflang. A homepage that declares no hreflang can still operate a fully internationalized site below the front page. The direction of this bias is therefore an undercount: the homepage-level hreflang-adoption figure is a lower bound on whether an institution uses hreflang anywhere.

**Raw HTML only.** All signals are read from the server-delivered HTML before any JavaScript executes. hreflang, canonical, or language tags injected by client-side rendering are invisible here by design. This is the correct lens for the non-rendering crawlers that BA-C-1 documents as the norm among AI crawlers, but it means the figures undercount signals present for any consumer that does render. The direction of the bias is known and one-sided.

**Frame, not population.** The 500 domains are a documented frame, not a random sample of the world's universities. The U.S. portion is ordered by enrollment (surfacing large online and for-profit institutions) and the international portion by traffic rank; both skew toward larger and higher-traffic institutions. Findings describe this frame and do not generalize to all universities.

**Single day.** These results describe the homepages as served during the collection window (2026-07-09); sites change without notice, and results should be assumed perishable. Any figure here should be checked against a current fetch before it is relied upon.

**Conservative language checks.** The BCP-47-lite validity screen rejects otherwise valid tags that carry a script subtag (zh-Hant, zh-Hans, sr-Latn) or an extended or variant tag [5]; reported invalid-code counts are an upper bound on genuine errors, and validity should be read as "valid under the lite check." The language-content mismatch flag is deliberately conservative (flagging only at detection probability  $\geq 0.9$  with a differing primary subtag), so its 6-of-429 count is a set of high-confidence candidates rather than a settled count — closely related language pairs can produce false flags, genuine mislabels below the threshold go uncounted, and language detection is itself weaker on very short, code-heavy, or mixed-script text.

**Non-analyzed domains.** 71 of the 500 domains returned no analyzable homepage — 33 with a non-200 status (all 33 were 403 bot-challenge or access-denied responses) and 38 with a transport-level failure (DNS, TLS, timeout, or other connection failure). They are excluded from the readiness denominators and reported separately; a 403 at the homepage is itself a legibility fact for a crawler, but it is not evidence about the hreflang or canonical signals the analyzed set measures.

Within these bounds, the census's conclusion is narrow and well supported: in the raw HTML that non-rendering crawlers read, the international-targeting layer is absent at most homepages in this frame, the hygiene signals around it are used correctly where present, and the gap is one of adoption rather than correctness.

## REFERENCES

1. U.S. Department of Education, College Scorecard. *College Scorecard data (institution-level open data and API)* (2026). <https://collegescorecard.ed.gov/data> Accessed 2026-07-09. [archived]
2. Hipolabs. *university-domains-list (world universities and domains)* (2026). <https://github.com/Hipo/university-domains-list> Accessed 2026-07-09. [archived]
3. Tranco (Le Pochat et al., NDSS 2019). *A research-oriented top sites ranking hardened against manipulation (list ID 46XZX)* (2026). <https://tranco-list.eu/> Accessed 2026-07-09. [archived]
4. Google Search Central. *Tell Google about localized versions of your page (hreflang)* (2026). <https://developers.google.com/search/docs/specialty/international/localized-versions> Accessed 2026-07-09. [archived]
5. IETF (A. Phillips, M. Davis, eds.). *Tags for Identifying Languages (BCP 47 / RFC 5646)* (2009). <https://www.rfc-editor.org/rfc/rfc5646.html> Accessed 2026-07-09. [archived]

---

## HOW TO CITE

Barkhausen AI (2026). International-readiness of university homepages: a raw-HTML census of hreflang, canonical, and language signals. <https://barkhausen.ai/research/university-intl-readiness-census-2026/>

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

