

The Crawl Gap: Where Endangered-Language Web Content Is Lost Before It Reaches a Corpus

Sharvin Goyal

Independent Researcher

sharvingoyal09@gmail.com

Abstract

Web-crawled corpora now claim that they have coverage of more than a thousand languages, but their subsets for endangered languages remain very small. Prior audits examine what is inside these corpora. We ask a question that complements this idea: what happens to the endangered-language web content that does exist? We curate and verify an inventory of 40 pages from active institutional websites (tribal language programs, community dictionaries, and learning portals) for 16 endangered North American languages from 7 families, and we attribute losses along the full pipeline: (S1) whether Common Crawl ever crawled the page, (S2) whether language identification (LID) recognizes the captured text, (S3) whether a standard quality-filter cascade can help retain it, and (S4) whether any of it reaches a released corpus (FineWeb-2, GlotCC). Most of the loss happens before quality filtering is even reached. These websites present to crawlers almost entirely in English, much of the actual language text is served by client-rendered dictionary applications that expose no text to crawlers, 6 of 16 languages have no label in the standard LID model, and none of the 36 inventory domains contributes a single document to any released per-language subset we examined. We release the inventory, the pipeline, and the per-stage attribution to support both crawl-side and community-side remedies.

1 Introduction

Most accounts of low-resource data scarcity focus on the corpus side. Quality audits show that web-mined subsets for low-resource languages are noisy or mislabeled (Kreutzer et al., 2022), and better mining recovers more (Dent et al., 2025). This paper examines the pipeline from the other end. Endangered languages with active revitalization movements have a real web presence, including language departments of sovereign nations, community dictionaries with audio, and lesson portals.

If the released corpora for these languages are tiny, the loss must occur somewhere between that web presence and the corpus. This paper locates where.

We start from an independently curated, script-verified inventory (§3) of 40 pages across 16 endangered North American languages and trace each page through four loss stages (§4): Common Crawl capture (S1), language identification of the captured payload (S2), a FineWeb-2-style quality-filter cascade (S3), and presence in released corpora (S4). All stages are measured with public artifacts, namely the Common Crawl capture index (CDX), the underlying Web ARChive (WARC) records, GlotLID v3 (Kargaran et al., 2023), and the released FineWeb-2 (Penedo et al., 2025) and GlotCC (Kargaran et al., 2024) subsets. All code, the inventory with per-URL provenance, and the stage-level results are open source.¹

2 Related Work

Corpus-side audits established that low-resource subsets of web corpora are often erroneous (Kreutzer et al., 2022). LID for low-resource languages has improved in coverage (Kargaran et al., 2023; Burchell et al., 2023) and in precision on closely related varieties (Fedorova et al., 2026), and rare-language extraction from Common Crawl has been reframed as targeted mining (Dent et al., 2025). Quality-filter cascades derive from C4 (Raffel et al., 2020) and Gopher (Rae et al., 2021). Our work complements this line of research by attributing where existing endangered-language content is lost before any mining begins. Our framing follows calls for community-centred data practice (Bird, 2020; Carroll et al., 2020).

3 The Inventory (S0)

Curation. We selected 16 endangered North American languages spanning 7 families and,

¹<https://github.com/Sharvin-coder/crawl-gap>

through documented web searches, collected 40 candidate URLs on 36 domains. The sites include official language programs of sovereign nations, community and university dictionaries, and learning portals. Every entry records its provenance, and no URL was invented or taken from any corpus. The inventory is a census of the verified pages we could locate, not a sample from a larger population, so we make no statistical generalization; the value of tracing it lies in attributing each individual loss.

Verification. A scripted pass fetched each page politely, extracted the server-visible text, and ran GlotLID v3. 37 of 40 pages are live. Three findings emerge at this stage. First, the document-level top-1 label of every live page is English. Second, the major dictionaries (FirstVoices, Cherokee Dictionary, itwêwina, the Blackfoot atlas) are client-rendered applications whose server-side HTML contains almost no text, so their content is invisible to any crawler. Much of what these sites publish is also not HTML at all: the pages we fetched link 299 distinct PDF files across 14 of 33 live domains, content a text-extraction pipeline does not see even when the linking page is crawled. Third, after following content-related internal links one hop deep and scoring the best page per site, only 11 of 34 language and site pairs expose any machine-detectable running text in the target language.

Accounting. One set of counts is used throughout. The curated inventory has 40 URLs on 36 domains, 37 of them live. The one-hop deepening step discovers 5 additional content pages not in the curated list, so the attribution pipeline traces 45 URLs in total. The funnel table (Table 1) covers only the 11 content pages with verified target-language text, which fall in 8 of the 16 languages.

4 Attribution Pipeline

For every traced URL we determine the first stage at which it is lost.

S1 (crawled?) We query the Common Crawl CDX index for exact-URL captures and host-level presence. Because a static page crawled once years ago would still enter FineWeb-2, a handful of recent crawls is not sufficient; we check 41 crawl releases spanning 2013 to 2026, including every release from 2023 through 2025, the newest 2026 releases, and one release per year back to 2013, which covers every crawl year FineWeb-2 draws

Language	S1 never	S1 no-200	S2 no-label	S2 miss	S3 killed	S4 absent	S4 present
Arapaho (arp)	0	0	0	1	0	0	0
Blackfoot (bfa)	0	0	0	1	0	0	0
Central Alaskan Yup'ik (esu)	0	0	0	1	0	0	0
Cherokee (chr)	0	0	0	2	0	0	0
Muscogee (mus)	0	0	0	0	1	0	0
Navajo (nav)	1	0	0	0	0	0	0
Ojibwe (ciw)	0	0	2	0	0	0	0
Plains Cree (crk)	0	0	0	2	0	0	0
Total	1	0	2	7	1	0	0

Table 1: First-failure attribution for content pages. Cells count pages (the best verified page per language and site); only the 8 languages with at least one page of verified target-language text appear, and the total is 11.

on. We also check each host’s live robots.txt policy for CCBot, the Common Crawl crawler.

S2 (recognized?) For each URL with an HTTP-200 capture, we retrieve that WARC record by range request, extract its text, and run GlotLID v3. A page passes if the document-level top-1 label belongs to the target language, or if at least 10% of its text lines receive such a label with $p \geq 0.5$. Languages absent from the label space cannot pass this stage (6 of 16 languages: cic, ciw, lkt, mez, osa, tli).

S3 (retained?) We apply a re-implemented FineWeb-2-style cascade to the captured text, in a fixed order: a language-probability gate (document-level GlotLID probability at least 0.65), Gopher-style quality and repetition checks (minimum word count, alphabetic-word fraction, duplicate-line fraction, top character-bigram share, and bullet-line ratio) (Rae et al., 2021), and the C4 terminal-punctuation heuristic (Raffel et al., 2020). We record the first filter that fires; all thresholds are pinned in the released code.

S4 (released?) We scan the complete released per-language subsets of FineWeb-2 and GlotCC for documents from any inventory domain.

5 Results

Table 1 and Figure 1 give the funnel; per-stage detail follows.

S1. Of the 45 traced URLs, 8 have no capture in any of the 41 crawls (every index query completed), 1 were captured only with non-successful HTTP statuses, and 36 have at least one successful capture; the three groups sum to 45. Only 1 of 36 hosts blocks CCBot in robots.txt, so the coverage gap comes almost entirely from discovery rather than from sites opting out. Never captured does not mean undiscoverable: these pages sit on live sites and are linked from pages we reached by ordinary browsing, so the gap reflects crawl

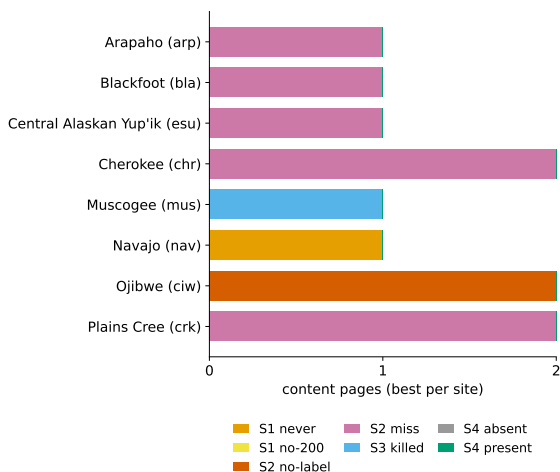


Figure 1: Stage attribution per language, for the same 11 content pages as Table 1.

scheduling and prioritization rather than technical unreachability. Crawl depth matters greatly here: 10 of the 36 captured URLs have their only successful captures outside the six most recent releases, so a shallow check against recent crawls alone would have roughly doubled the apparent never-crawled count.

S2. Of the captured pages, 1 passes, 23 are assigned to other labels, and 12 belong to languages with no label at all. The routing is not toward other minority languages: 23 of the 23 routed pages receive English as their top-1 label, because the target-language material is embedded in English framing. The languages missing from GlotLID’s label space are exactly the languages that have no released subset in either corpus.

S3. Applied to all captured pages, the cascade keeps 20 and removes 16, most often at the language-probability gate. The single page that passed S2, a bilingual lesson page, is among those removed: its English framing pulls the document-level language probability below the gate. Because revitalization materials are typically bilingual, this behavior affects them systematically.

S4. The released subsets are small. Across the 10 of 16 languages that have a per-language subset at all, the subsets scanned here range from 1 to 239 documents (976 in total). The number of documents from inventory domains in either released subset is 0. This establishes absence from the two corpora we can scan document by document. As a breadth check on label coverage, none of the

16 languages has any per-language subset in mC4, OSCAR-23.01, or CulturaX, and MADLAD-400 carries labels for 4 of the 16 (Cherokee, Navajo, Ojibwe, and Cree) and none for the other 12. The institutional web presence of these 16 languages contributes no documents to the corpora examined here.

6 Discussion

These losses are structural rather than incidental. Institutional pages are written in English for good reasons, since their audience includes learners and citizens who do not yet speak the language. The most valuable language content sits behind JavaScript applications and PDFs. LID label spaces were built from languages already observable in corpora, and this circularity keeps unlabeled languages out. Remedies need to match the stage where the loss happens. On the crawl side, rendering support or capture agreements with community dictionaries would expose text that currently stays hidden. Line- or paragraph-level LID ahead of document-level gating would let bilingual pages contribute their target-language lines instead of failing as a whole. Label spaces can grow from inventories instead of corpora, and verified community domains can be exempted from filters that were tuned for general web text. Above all, communities should keep authority over whether and how their language data enters corpora (Carroll et al., 2020; Bird, 2020).

Limitations

Our inventory covers one region (North America), 16 languages, and 45 URLs; it is small because verifiable target-language pages are genuinely scarce, and we trace the population we found rather than sampling from it. A matched control set of institutional pages in higher-resource languages, and expansion to more languages and domains in collaboration with community partners, are natural next steps. The attribution thresholds (the S2 pass criterion and the S3 cascade) are pinned but necessarily approximate the released pipelines. PDF and audio content, which makes up a large share of what language programs publish, is out of scope for the text funnel, so our estimate of the crawlable text is conservative; §3 quantifies the PDF surface. Some inventory domains may be represented in corpora under other languages’ subsets (typically English), which we do not count as presence for

the target language.

Ethics Statement

We fetch only publicly served pages, at most once per page per run, with an identifying research user agent and per-host delays. The repository releases URLs, metadata, and derived statistics, and never scraped text, consistent with the CARE principles (Carroll et al., 2020). The released artifact documents a correction-and-removal process, and we commit to honoring removal or correction requests from the nations or from site stewards. Corpus presence of community content is reported in aggregate. We do not redistribute any community language material.

References

- Steven Bird. 2020. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. [An open dataset and model for language identification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada. Association for Computational Linguistics.
- Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. [The CARE principles for indigenous data governance](#). *Data Science Journal*, 19(1):43.
- Rasul Dent, Pedro Ortiz Suarez, Thibault Clérice, and Benoît Sagot. 2025. [Identifying rare languages in Common Crawl data is a needles-in-a-haystack problem](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1460–1473, Suzhou, China. Association for Computational Linguistics.
- Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. [OpenLID-v3: Improving the precision of closely related language identification – an experience report](#). In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 275–292, Rabat, Morocco. Association for Computational Linguistics.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.
- Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. [GlotCC: An open broad-coverage CommonCrawl corpus and pipeline for minority languages](#). In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. [Quality at a glance: An audit of web-crawled multilingual datasets](#). *Transactions of the Association for Computational Linguistics*, 10:50–72.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2025. [FineWeb2: One pipeline to scale them all – adapting pre-training data processing to every language](#). *arXiv preprint arXiv:2506.20920*.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. [Scaling language models: Methods, analysis & insights from training Gopher](#). *arXiv preprint arXiv:2112.11446*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.