Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

Measurement practice in Core Web Vitals research: A mapping of 357 works

Harrison Oke Ekpobimi ✉[1], **Adepegba Akindayomi Akintade**[2], **Wasiu Abiodun Bello**[3]

[1] *School of Technology, Walsh College, Troy, MI, United States. ORCID 0009-0002-1870-5466*

[2] *School of Computer Science and Information Systems, Osiri University, Nebraska, United States. ORCID 0009-0001-0750-934X*

[3] *Department of Decision and Information Sciences, Oakland University, Rochester, MI, United States. ORCID 0009-0000-6037-6956*

Corresponding Author: Harrison Oke Ekpobimi**, E-mail:** hekpobim@mail.walshcollege.edu

ABSTRACT

Core Web Vitals have become the reference metrics for client-side web performance, and a literature reporting them has grown from five works in 2020 to 152 in the first ten months of 2026. The measurement practices inside that literature have not been described, so what its accumulated evidence can support is not known. This systematic mapping study examines what the literature reporting Core Web Vitals metrics actually measures, asking which data sources the measurements come from, which metrics are reported, whether connection quality enters the analysis, and which populations and author communities the work represents. Seven precise phrase queries were run against the OpenAlex registry on 25 October 2026, unioned and deduplicated, yielding 460 records. Six screening gates reduced these to an analytic corpus of 357 works published from May 2020 onward. Each work was coded on six dimensions against rules fixed before coding began, with geography coded twice over and the two measures kept apart. Lab tools that simulate a single visit are named by 136 works and real-user or archive sources by 16. Of the 266 works reporting measurements, 123 name no data source at all. Connection quality enters 56 works of 357, and the pattern inverts by source, since synthetic studies consider it in 60.7 per cent of cases while the six works using the Chrome User Experience Report consider it in one. Lighthouse cannot measure Interaction to Next Paint, and among the 136 lab-tool works that metric appears nine times against 61 for the Total Blocking Time proxy. Only 15 works of 357 report all three currently stable Core Web Vitals together. Of the 207 works carrying affiliation data, 84 have an Indonesian institution and eight an African one. The field setting measurement norms for the web measures that web with simulated single visits, frequently without saying where its numbers came from, and largely without the connection variable its own instruments record. Reviewers and editors can close the traceability gap at low cost by requiring a named data source.

KEYWORDS

Core Web Vitals; web performance; systematic mapping study; real-user monitoring; measurement practice; research geography

ARTICLE INFORMATION

ACCEPTED: 09 October 2026 **PUBLISHED:** 10 November 2026 **DOI:** 10.32996/jcsts.2026.8.12

1. Introduction

A page that loads slowly loses the person waiting for it, and the web industry now has an agreed vocabulary for saying how slowly. Google introduced the Web Vitals programme in May 2020 to reduce a confusing abundance of performance metrics to a small set that every site owner could be expected to measure (Walton, 2024). Three of those metrics are currently stable. Largest Contentful Paint reports loading, Cumulative Layout Shift reports visual stability, and Interaction to Next Paint reports responsiveness, each assessed at the 75th percentile of page loads. Because the same metrics appear in Google Search reporting, in PageSpeed Insights and in Chrome developer tooling, they have moved from a measurement proposal to the operative definition of web performance for a large part of the industry.

Research has followed quickly. The literature reporting these metrics has grown from five indexed works in 2020 to 152 in the first ten months of 2026, and that growth has happened without anyone describing how the work is done. What a field measures, and with what instrument, determines what its findings can be used for. A body of evidence collected under one set of conditions cannot support engineering guidance for a different set, and the guidance drawn from this literature is already being applied to a web whose users sit on connections and devices that differ widely from one another (Ahmad et al., 2016; Zaki et al., 2014).

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

The programme's own documentation is explicit on this point. It records that page performance varies substantially with device capability, network conditions, competing processes and how the person interacts with the page, and that only field measurement captures the complete picture (Walton, 2024). Lab measurement is presented there as a development aid and not as a substitute. Whether the research literature has honoured that distinction is an empirical question, and it is the question this paper answers.

Systematic mapping is the right instrument for the question. A mapping study describes the shape of a research area, counts how its work is distributed across categories of interest, and identifies where evidence is thin (Petersen et al., 2008; Petersen et al., 2015). It does not synthesise effect sizes, which suits a field whose primary studies measure different artefacts under different conditions. The method is well established in software engineering and its reporting conventions are settled (Kitchenham et al., 2010; Usman et al., 2023).

This study maps 357 works published from May 2020 onward that report Core Web Vitals metrics. It codes each work on six dimensions covering the data source behind the measurements, the metrics reported, the application domain of the artefact studied, whether connection quality enters the analysis, the kind of evidence offered, and geography. Geography is coded twice over, once from the study population the work names and once from author affiliation, because those are different things and conflating them would make the headline claim unsafe.

Four findings emerge, and each is countable from the deposited dataset. The first is that the field does not use real-user data, since lab tools are named by 136 works and real-user or archive sources by 16. The second is that a large part of the literature does not say where its numbers came from, since 123 of the 266 works reporting measurements name no data source at all. The third is that connection quality is least likely to be analysed by the studies that hold genuine connection data, which is the opposite of what anyone would predict. The fourth is that the literature has a geography nobody has described, with 84 of the 207 works carrying affiliation data attached to an Indonesian institution and eight to an African one.

A fifth result falls out of the first. Lighthouse, the single most common instrument in this corpus, cannot measure Interaction to Next Paint, because a simulated page load has no user to interact with it (Walton, 2024). Among the 136 lab-tool works that metric appears nine times against 61 for Total Blocking Time, the proxy Lighthouse reports in its place. Only 15 works of 357 report all three currently stable Core Web Vitals together. The instrument is choosing the metric, and in doing so it is removing one third of the programme's definition from view.

The contribution is the map, and the thin places on it are the finding. Section 2 sets out the programme and the measurement distinction it rests on, and positions the study against the secondary studies that exist nearby. Section 3 gives the method in full, including the queries, the screening gates, the coding rules and the validation probes. Section 4 reports the results against five research questions. Section 5 discusses what the evidence base can and cannot support. Section 6 states the threats to validity, including the absence of a second coder. Section 7 concludes.

2. Literature review

2.1 The programme and its metrics

Web Vitals began as an attempt to simplify. Performance tooling had accumulated many metrics over a decade and site owners were not in a position to judge which of them described the experience their users were having (Walton, 2024). The programme selected a subset called the Core Web Vitals, defined each one to be measurable in the field, attached a threshold to each, and placed those thresholds at the 75th percentile so that a site passes only when most visits are good. Largest Contentful Paint should occur within 2.5 seconds, Interaction to Next Paint should be 200 milliseconds or less, and Cumulative Layout Shift should stay at or below 0.1.

The set has changed once. First Input Delay was the original interactivity metric, and it measured only the delay before the first interaction was processed, which left most of the runtime responsiveness problem unmeasured. Interaction to Next Paint was developed to cover that gap, moved through experimental and pending status, and became stable in 2024, at which point First Input Delay was retired (Walton, 2024). A supporting set of metrics sits outside the core. Time to First Byte and First Contentful Paint describe parts of the loading sequence and help to diagnose Largest Contentful Paint. Total Blocking Time is a laboratory metric and a proxy for Interaction to Next Paint, and it is outside the core set precisely because it is not field-measurable and does not report a user-centric outcome.

Whether these metrics capture the experience they claim to has been tested directly. Work relating the Core Web Vitals to quality-of-experience judgements collected from users found the relationship real but partial, which is a useful corrective to treating the thresholds as a complete account of perceived speed (Wehner et al., 2022). A follow-up study examined what cookie consent banners do to those same metrics and to the experience behind them (Wehner et al., 2023). Earlier quality-of-experience research had already established that the tolerance for waiting is itself context-dependent (Egger et al., 2012). The programme is therefore better understood as a usable approximation under active study than as a settled measurement of user experience.

2.2 Laboratory measurement and field measurement

The distinction between lab and field data is the one that matters most for reading this literature. A lab tool loads a page once, from one machine, over a connection the tool controls, and reports what it observed. PageSpeed Insights, Lighthouse, GTmetrix and WebPageTest all work this way, and all are easy to run without permission from the site owner. Field data come from real visits

by real people on their own devices and connections. The Chrome User Experience Report publishes such data for sites that meet its popularity and opt-in thresholds, and a site owner can also collect it directly through the browser performance APIs or the web-vitals JavaScript library (Walton, 2024).

The two do not agree, and they are not supposed to. A single simulated visit cannot represent a distribution of visits, and the programme's own guidance says so plainly while still recommending lab tools for catching regressions during development (Walton, 2024). Research in this area has shown repeatedly that the network and the device, not the markup, often dominate the result. Dissecting web latency in Ghana attributed a large share of observed delay to factors a lab-throttled test would not reproduce (Zaki et al., 2014). Measurement of mobile phone characteristics in lower-income markets documented device populations very different from those a default lab profile assumes (Ahmad et al., 2016). Work on reducing JavaScript payloads for such users has treated the constrained connection as the design premise and not as noise (Chaqfeh et al., 2020; Chaqfeh et al., 2022).

Industrial case studies occupy a middle position and are instructive about what careful practice looks like. One reported reducing a mobile page's load time from 6.2 seconds to 0.15 seconds and documented the measurement regime that supported the claim (van Riet et al., 2020). A later study followed web performance work through an organisation's development process over time (van Riet et al., 2023). A study of the performance and energy cost of advertising and analytics in mobile web applications measured both on real devices (Petalotis et al., 2024). Transport-level work on HTTP/3 and on post-quantum TLS has likewise measured under varied network conditions because the effect being studied lives there (Gupta & Bartos, 2022; Kampanakis & Childs-Klein, 2024).

2.3 Secondary studies nearby

No prior secondary study maps measurement practice in this literature, and the corpus assembled here is itself evidence for that claim. Eleven of the 357 works were flagged during coding as reviews or as conceptual contributions, and inspection of all eleven found no systematic review or mapping study of the Core Web Vitals literature among them. The nearest secondary studies sit one step to the side, in web accessibility, where a systematic literature review of empirical accessibility studies of educational websites (Campoverde-Molina et al., 2020) and a review of accessibility evaluation using automated tools (Macakoglu & Peker, 2022) have both been published. Those reviews ask what the accessibility literature found. This study asks what the performance literature measured, which is a different question and one that has to be answered before the findings can be pooled.

Software engineering has a developed tradition of looking at its own secondary studies this way. Tertiary studies have assessed the quality of systematic reviews in the discipline generally (Kitchenham et al., 2010) and in mobile application engineering specifically (Zein et al., 2023). A quality assessment instrument for systematic reviews in software engineering has been proposed and validated (Usman et al., 2023). Work on how primary case studies are reported inside secondary studies has named the evidential problem that arises when a review inherits gaps from the studies it aggregates (Rainer & Wohlin, 2024). The mapping reported below finds a gap of exactly that kind, since a review of this literature would inherit 123 primary studies that do not identify their own data source.

Research artefacts have become a separate object of study, and the picture there is not encouraging. A systematic mapping of artefact availability in software engineering secondary studies found that the majority of such studies do not make their extracted data available (Huotala et al., 2025). This study therefore deposits its full coded dataset and every script that produced it, and Section 3.6 lists the contents. A reader who disagrees with any count in Section 4 can recompute it.

3. Methodology

The method follows the systematic mapping guidelines for software engineering (Petersen et al., 2008; Petersen et al., 2015), with the search and screening reported in the flow format that secondary-study reporting conventions call for (Page et al., 2021) and with the review process itself conducted to the established software engineering protocol (Kitchenham & Charters, 2007). Practical lessons from applying that protocol inside software engineering informed two choices in particular, the writing of the inclusion rules before retrieval and the recording of every judgement at the moment it was made (Brereton et al., 2007). Decisions that required judgement were written to disk with their reasons at the moment they were made, which is what allows the screening to be audited without rerunning it.

3.1 Research questions

Five questions were fixed before retrieval began. RQ1 asks about the size, growth and composition of the published literature reporting Core Web Vitals metrics. RQ2 asks which data sources that literature draws its measurements from. RQ3 asks which metrics it reports, and whether the reported set matches the metrics the programme currently defines. RQ4 asks whether connection quality enters the analysis, and whether that depends on the data source. RQ5 asks which populations and which author communities the literature represents.

RQ4 carries the most weight for practice. A field that controls the network away, or that never varies it, produces findings that do not transfer to markets where the connection is the binding constraint. RQ3 was added after RQ2 was answered, because the data-source result implied a testable consequence for metric coverage, and that consequence is reported in Section 4.3.

3.2 Data source and search strategy

Retrieval used the OpenAlex scholarly registry, which indexes works, authors, venues and institutions under an open licence and exposes them through a public API (Priem et al., 2022). OpenAlex was chosen over Scopus and the Web of Science for three reasons. Its coverage of references has been assessed against both commercial indexes and found broadly comparable in extent while differing in composition (Culbert et al., 2025). Its API requires no subscription, which means the retrieval step can be rerun by any reader. Procedures for using it to assemble evidence-synthesis corpora have been documented (Harder, 2024). The coverage differences it carries are treated as a threat to validity in Section 6.3 and not as a solved problem.

Seven phrase queries were run against the combined title and abstract index, each wrapped in quotation marks so the registry treated it as a phrase. The phrases were core web vitals, web vitals, largest contentful paint, cumulative layout shift, interaction to next paint, first contentful paint, and time to first byte. Table 1 reports the yield of each. A single disjunctive query combining all seven was attempted first and rejected, because the registry loosened the operator and returned more than seventy thousand records, most of which contained none of the phrases. The seven result sets were therefore unioned and deduplicated by work identifier, which produced 460 unique records. Retrieval ran on 25 October 2026 and the retrieval date governs every count in this paper.

Table 1. Search phrases and records retrieved, OpenAlex, 25 October 2026.

Phrase query	Records retrieved
core web vitals	175
largest contentful paint	179
cumulative layout shift	110
interaction to next paint	31
first contentful paint	129
time to first byte	94
web vitals	201
Union of the seven, deduplicated by work identifier	**460**

Note. Each phrase was wrapped in quotation marks and run against the combined title and abstract index. Counts sum to more than the union because a work matching several phrases is retrieved by each. A single disjunctive query combining all seven was attempted first and rejected, because the registry loosened the operator and returned more than seventy thousand records.

Abstracts arrive from the registry as an inverted index, which lists the positions at which each token appears. The abstracts were reconstructed into running text before coding, since the coding scheme reads abstracts. The retrieval script requests twenty bibliographic fields for each record, including authorships with institutional affiliations, open access status, publication type, publication date and citation count, so that the coding does not require a second visit to the registry.

Backward snowballing was considered and not used (Wohlin, 2014). The decision was deliberate. Snowballing from a seed set would improve recall for works that name the metrics only in their full text, and it would also admit works whose connection to the programme is indirect, which would blur the boundary the year floor is there to draw. The recall cost is stated in Section 6.3.

3.3 Study selection

Inclusion and exclusion rules were written before retrieval and are stated here in full. A record is included when four conditions hold together. It names at least one of the seven phrase terms in its title or abstract. It carries a publication year of 2020 or later. Its record type is an article, conference paper, book chapter, dissertation, preprint, review or report. Its use of the metric name is in a web performance sense and not incidental. A record is excluded when it is a dataset, a software deposit, a standard or an editorial item, when it is retracted, when it duplicates another record by identifier or by title, when it carries no publication year, or when the named metric appears while the object of study is something other than web performance. Language was not a criterion, and the corpus is multilingual as a result.

Six gates were applied in a fixed order, and each was counted. Figure 1 reports the flow, and the counts in it reconcile without reference to the note. The first gate removed records sharing a digital object identifier, and found none, because the union step had already deduplicated by work identifier. The second gate removed records whose type fell outside the scholarly scope of the study, which took out 51 datasets, software deposits, standards and editorial items. The third gate screened for domain relevance. The fourth removed one record carrying no publication year. The fifth applied the May 2020 year floor and set aside 22 precursor works. The sixth removed duplicate titles.

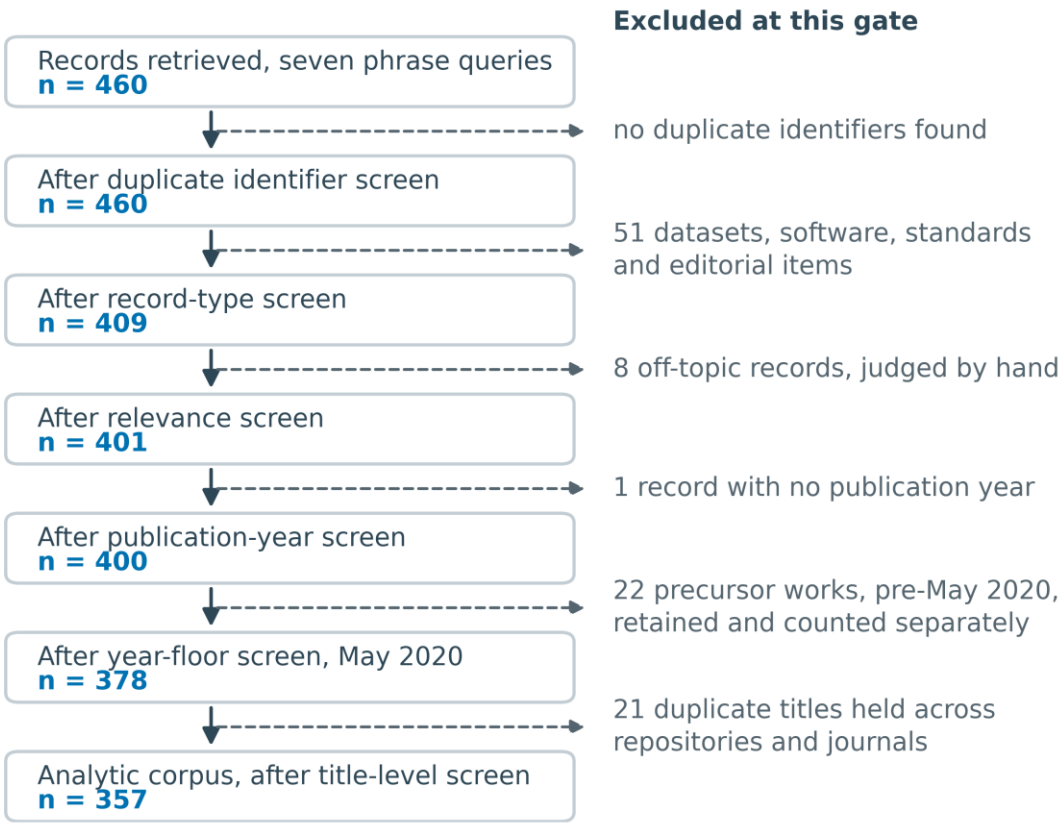


Figure 1. Study selection flow, from retrieval to analytic corpus.

Note. Seven phrase queries were unioned and deduplicated by work identifier before the first gate, which is why no duplicate identifiers remained to remove. Every record excluded at the relevance gate was read and decided by hand with its reason recorded, and those decisions are in the deposit. A precursor set of 22 works published before May 2020 is set aside at the year-floor gate and counted separately, because the metric names predate the programme this study maps. Those 22 works are reported in Section 4.5 and listed in the deposit.

The relevance gate was the only one needing judgement, and it was handled in two steps so the judgement would be visible. An automated screen tested the title and abstract of every record against a pattern of web-domain terms and flagged twelve records that matched none of them. Each of the twelve was then read and decided by hand, with the reason recorded. Four were kept, including a study of accessibility issues measured on government homepages and a routing study that measured its effect through page load. Eight were excluded, in every case because the metric name appeared incidentally while the object of study was something else, such as recommender quality, team cognition under algorithmic management, or the failure behaviour of a payment batch.

The year floor sits at May 2020, which is when the Web Vitals programme was announced (Walton, 2024). That date was chosen because it is a real boundary in the subject matter and not an arbitrary round number. Twenty-two works using these metric names before that date were retained as a separately counted precursor set and excluded from the analytic corpus, since their use of the terms predates the programme this study maps. The earliest is indexed to 1970 and six fall in 2019.

Title-level deduplication removed 21 records that identifier matching had missed, and what it caught deserves reporting in its own right. One Italian article on hosting appears as six separate Zenodo deposits. One 2026 report appears on SSRN, on Zenodo, on figshare and in a journal, four records for one contribution. Thirteen of the 21 removed records were repository deposits of a work indexed elsewhere. A reader counting this literature without a title-level screen would overstate its size by roughly six per cent, and the inflation is concentrated in exactly the preprint and deposit categories that have grown fastest.

3.4 Classification scheme

Six dimensions were coded, and the rules for each were written before coding began and published with the paper. Table 2 gives the scheme. The coder is deliberately conservative on every dimension. Where a title and abstract do not state something, the record is coded as not stated and the value is never inferred, because an inferred code in a mapping study is an invented finding.

Table 2. Classification scheme, with the rule applied to each dimension.

Dimension	Categories	Rule
Data source	Chrome UX Report, HTTP Archive, lab tool, custom real-user monitoring, synthetic or throttled, not stated	Coded as a set from named tools and datasets in the title or abstract. A work naming none is coded not stated and is never inferred.
Metric	LCP, CLS, INP, FID, FCP, TTFB, TTI, TBT, Speed Index	Coded as a set from the metric name or its standard abbreviation, matched on word boundaries.
Application domain	E-commerce, government, news and media, education, health, library and repository, general or unspecified	Coded from sector terms describing the artefact studied. A work naming no sector is coded general or unspecified.
Network	Considered, not considered	Considered when connection type, bandwidth, latency, round trip time or throttling enters the analysis as reported in the abstract.
Evidence type	Reports measurement, measurement within review, review or conceptual, not stated	Coded from statements of what the work did, matched in English, Indonesian, Ukrainian, Russian and Italian.
Geography	Study population country or region, author institution country	Coded twice over and reported separately. Population geography from country and region names in the abstract, institution geography from registry affiliation records.

Note. Rules were fixed and published before coding began. Records for which the registry supplied no abstract are coded as insufficient information on every dimension that reads the abstract, and are retained in the corpus. The full rule set, including the exact patterns, is in the deposited coding script.

Data source and metric are coded as sets, since a work may name several of each. The shares reported in Section 4.2 therefore exceed one hundred per cent when summed, and the tables say so. Network is coded as a single value, since the question is only whether connection quality enters the analysis at all. Evidence type separates works reporting their own measurements from reviews and from conceptual contributions, and a residual category holds the 21 works for which no abstract was available.

Geography required the most care and is coded twice over. Study population geography is taken from country and region names appearing in the abstract and is what the work actually studied. Author institution geography is taken from the registry's affiliation records and is a proxy for where the work was produced. The two are reported separately throughout and are never combined, because an author's employer is not the population an author studied, and no figure or table in this paper merges them.

3.5 Coding procedure and validation

Coding was automated against the published rules and then validated by targeted probes for false negatives. Automation was chosen because a scheme applied across 357 works over several sessions drifts when it is applied by hand, and a published rule set applied by a script does not drift. Methods combining rule-based and assisted coding in mapping studies have begun to be documented (Olar et al., 2026), and the approach used here is the rule-based half of that pattern with manual validation in place of model assistance.

The probes found and corrected three faults, and reporting them is part of the method. The evidence-type rule initially matched only first-person English constructions and returned 77.5 per cent of the corpus as not stated, which was a coder failure and not a property of the field. Broadening it to cover Indonesian, Ukrainian, Russian and Italian equivalents brought the measuring share to 72.0 per cent. The geography rule omitted region names and lost a genuine sub-Saharan African study until Africa and its regional subdivisions were added as terms. Identifier deduplication missed the 21 title duplicates described in Section 3.3.

One validation probe was itself wrong, and the correction is recorded because it bears on how much confidence the coding deserves. A probe written to detect missed Kenyan studies used a loose pattern without word boundaries and returned matches inside the Indonesian word kenyamanan, which means comfort. The coder was right and the probe was wrong. Probes were rewritten with word boundaries after that, and the lesson generalises, since a validation step applied to a multilingual corpus needs the same care as the coding it checks.

A second coder was not used. Section 6.2 states what that costs and why the cost is not uniform across the six dimensions.

3.6 Analysis and deposited material

Analysis is descriptive and the counts are exact and not estimated, since the corpus is the population of indexed works meeting the inclusion rules and not a sample from it. Cross-tabulations are reported with both the count and the denominator, so that a

reader can see the base a percentage rests on. No inferential test is reported, because there is no sampling distribution to test against and a p-value computed on a population would be a decoration.

The deposit holds the coded dataset for all 357 works with every code assigned, one row per work across 19 columns, together with the five scripts that produced it in execution order. It also holds the recorded screening decisions with the reason for each, the list of removed duplicate titles, and the relevance flags. The scripts run end to end from the OpenAlex API, so the study can be reproduced from the registry and not merely recomputed from the deposited table. Availability is stated at the end of the paper.

4. Results

4.1 Corpus size, growth and composition (RQ1)

The literature is small and growing fast. Five works appear in 2020, nine in 2021, nineteen in 2022, twenty-seven in 2023, fifty-four in 2024, ninety-one in 2025 and one hundred and fifty-two in the first ten months of 2026. Table 3 gives the composition by year and record type. The 2026 count covers a partial year, with ten of twelve months elapsed at the retrieval date, so it is reported but excluded from any trend statement. Even on the completed years the growth is steep, since 2025 alone holds more works than 2020 through 2023 combined.

Table 3. Analytic corpus composition, by publication year and record type.

Year	Article	Conference paper	Dissertation	Preprint	Book chapter	Report	Total
2020	1	4	0	0	0	0	5
2021	8	1	0	0	0	0	9
2022	10	8	0	1	0	0	19
2023	20	3	3	0	1	0	27
2024	30	14	3	1	6	0	54
2025	62	20	6	3	0	0	91
2026	105	14	18	11	1	3	152
Total	**236**	**64**	**30**	**16**	**8**	**3**	**357**

Note. The 2026 row covers a partial year, with ten of twelve months elapsed at the retrieval date of 25 October 2026, and registry indexing lags publication unevenly across venues. No growth rate is fitted for that reason. Open access covers 279 of the 357 works, of which 125 are diamond.

Journal articles account for 236 works and conference papers for 64. What distinguishes this corpus from a conventional software engineering literature is the weight of the remainder. Thirty dissertations and sixteen preprints appear, and both categories concentrate in the last two years, with eighteen of the thirty dissertations and eleven of the sixteen preprints dated 2026. Open access covers 279 of the 357 works, and diamond open access alone accounts for 125, which reflects a corpus published substantially in journals that charge neither readers nor authors.

Abstracts are available for 336 works and absent for 21. Those 21 are retained in the corpus and carried through every analysis as an explicit residual category, because dropping them would quietly improve every other count. English is the recorded language for 173 works and Indonesian for 30, while the registry records no language for 124, which is a limitation of the registry and not a property of the corpus.

4.2 Where the measurements come from (RQ2)

Two public datasets carry genuine real-user measurement for the whole web, and this literature has barely touched either. Figure 2 reports the data sources named across the corpus. Lab tools are named by 136 works, which is 38.1 per cent. Synthetic or throttled testing is named by 28, custom real-user monitoring by 10, the Chrome User Experience Report by 6 and the HTTP Archive by 3. Sixteen works in total, 4.5 per cent of the corpus, name a real-user or archive source of any kind, since three of them name two such sources. Works may name more than one source, so the shares exceed one hundred per cent when summed.

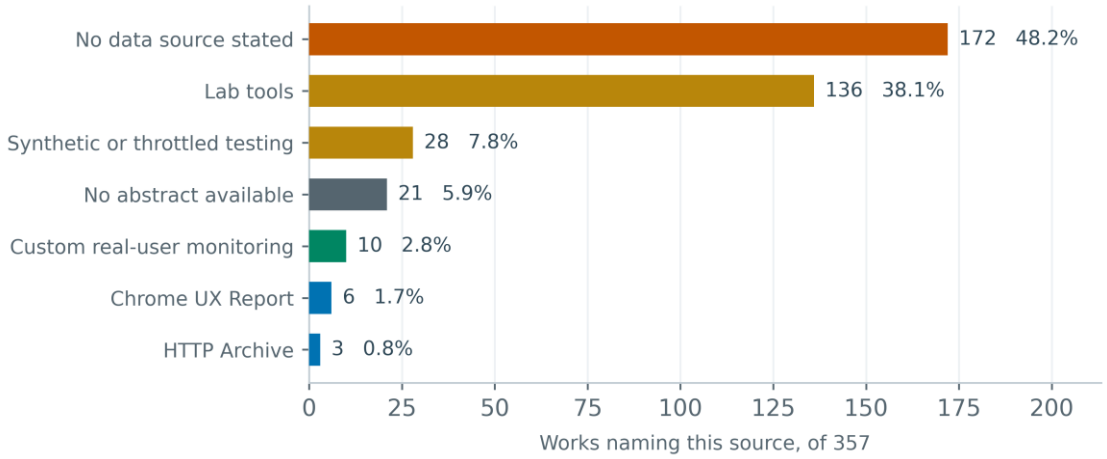


Figure 2. Data sources named across the 357 works of the analytic corpus.

Note. Lab tools are Lighthouse, PageSpeed Insights, GTmetrix, WebPageTest and sitespeed. Works may name more than one source, so the shares sum to more than one hundred per cent. The no-abstract category holds the 21 works for which the registry supplied no abstract, which are retained in the corpus and carried through every analysis. Sixteen works name a real-user or archive source of any kind.

The traceability result is sharper than the headline. Of the 266 works that report their own measurements, 123 name no data source at all, which is 46.2 per cent. Those papers report performance numbers that a reader cannot trace to any instrument, any sample or any set of conditions. A secondary study pooling this literature would inherit all 123, and the evidential problem that creates has been described in general terms for software engineering secondary studies (Rainer & Wohlin, 2024). Only 15 of the 266 measuring works, 5.6 per cent, name a field or archive source.

The works that do use field data show what is possible. A study of content delivery readiness across Philippine higher education websites combined HTTP Archive records with Chrome User Experience Report data and reported both (Banaybanay & Ricafort, 2026). A causal modelling study drew on the same two sources (Lin & Liu, 2025). A dashboard standardisation study used the Chrome User Experience Report for its benchmarks (Prasetyo & Rokhman, 2024). These are a small number of works against a corpus of 357, and their scarcity is the finding.

Lab tools dominate for reasons that are easy to understand and still consequential. PageSpeed Insights, Lighthouse and GTmetrix can be pointed at any public URL without the site owner's cooperation; they produce a number immediately and require no infrastructure. A study comparing marketplace performance in Indonesia used Lighthouse (Muna et al., 2022), and a comparison of four rendering frameworks used Lighthouse, PageSpeed Insight and JMeter together (Zaidan & Suyatno, 2025). Both are competent applications of the instrument available to them. The question this paper raises is not about any individual study but about what accumulates when an entire field reaches for the same instrument.

4.3. Which metrics are reported (RQ3)

The literature mapping Core Web Vitals mostly does not report the current Core Web Vitals. Table 4 gives metric coverage across the corpus. Largest Contentful Paint appears in 175 works and Cumulative Layout Shift in 104; these are two of the three stable metrics. Interaction to Next Paint, the third, appears in 29. Only 15 works of 357 name all three together, and of the 90 works naming both Largest Contentful Paint and Cumulative Layout Shift, 75 omit the interactivity metric. Total Blocking Time, which is a laboratory proxy and not a Core Web Vital, appears in 78 works, more than twice as often as the metric it proxies.

Table 4. Metric coverage across the analytic corpus.

Metric	Status in the programme	Works	Share of 357
Largest Contentful Paint	Core, stable	175	49.0%
First Contentful Paint	Supporting	109	30.5%
Cumulative Layout Shift	Core, stable	104	29.1%
Total Blocking Time	Supporting, laboratory proxy for INP	78	21.8%
Speed Index	Supporting	67	18.8%
Time to First Byte	Supporting	50	14.0%
Time to Interactive	Supporting	38	10.6%
Interaction to Next Paint	Core, stable	29	8.1%
First Input Delay	Retired in 2024	14	3.9%

Note. Works may report more than one metric, so the shares sum to more than one hundred per cent. Only 15 works report all three currently stable Core Web Vitals together, and of the 90 reporting both Largest Contentful Paint and Cumulative Layout Shift, 75 omit Interaction to Next Paint. Status follows the programme documentation as of the retrieval date.

The instrument explains the pattern. Lighthouse loads a page in a simulated environment with no user in it, so it cannot measure Interaction to Next Paint at all, and it reports Total Blocking Time in its place (Walton, 2024). Figure 3 shows what that does to the literature. Among the 136 lab-tool works, Interaction to Next Paint appears nine times and Total Blocking Time sixty-one. Among the 16 works using a field or archive source the ordering reverses, with the interactivity metric in three and the proxy in one. A tool that cannot see a metric produces a literature that does not discuss it.

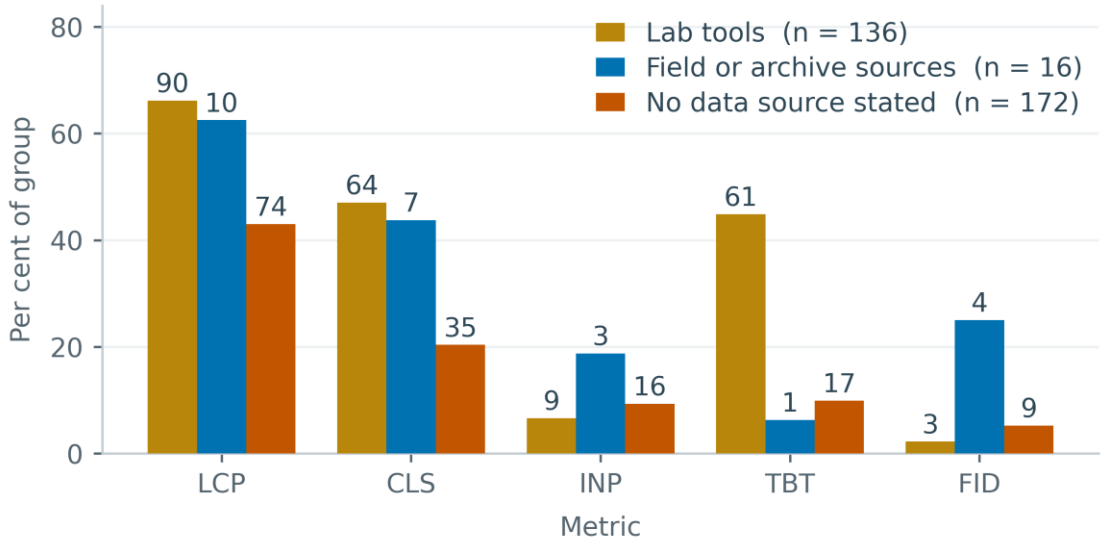


Figure 3. Metric coverage by data source group.

Note. Bars give the share of each group naming the metric and the figure above each bar gives the count. The field-or-archive-source category combines the Chrome User Experience Report, custom real-user monitoring and the HTTP Archive. Groups overlap where a work names more than one source. Lighthouse cannot measure Interaction to Next Paint, because a simulated load carries no user interaction, and reports Total Blocking Time as a proxy, which is the mechanism the lab-tool column shows.

First Input Delay behaves as a retired metric should. It appears in no work before a single mention in 2020, rises modestly, and falls to two mentions in 2026, while Interaction to Next Paint appears in none before 2024, then four in 2024, five in 2025 and twenty in 2026. Six works published in 2025 or later still name the retired metric. This study tested whether the field had failed to follow its own metric change and found that it had followed, with a lag. The observation is reported here as a modest one and is not advanced as a finding.

4.3 Whether the network enters the analysis (RQ4)

Connection quality enters the analysis in 56 works of 357, which is 15.7 per cent, and 280 works do not consider it at all. The bare share is unsurprising for a literature built on lab tools. The pattern across data sources is the part worth publishing, because it runs opposite to what anyone would predict. Figure 4 reports it.

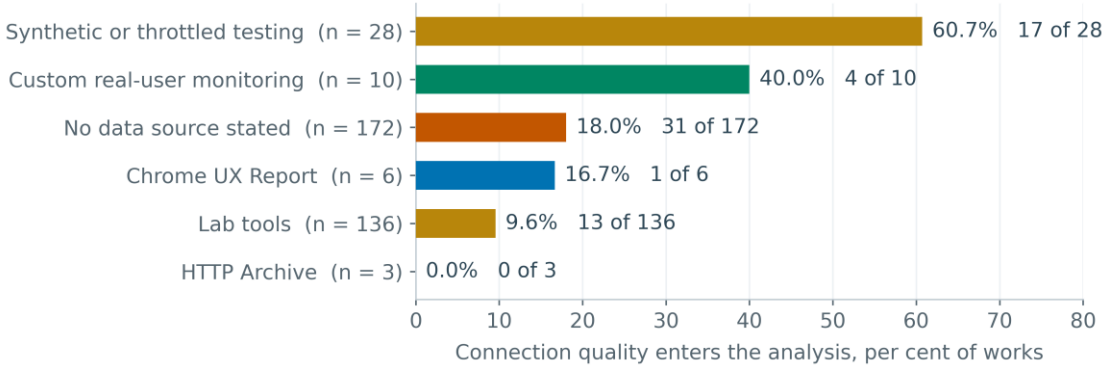


Figure 4. Consideration of connection quality, by data source.

Note. A work is coded as considering connection quality when the abstract shows that connection type, bandwidth, latency, round trip time or throttling enters the analysis. Groups overlap, because a work naming two sources appears under both. Denominators differ by two orders of magnitude across the rows, so the counts are shown beside each percentage. Colours match Figure 2.

Synthetic and throttled studies consider the network most often, in 17 of 28 cases, which is 60.7 per cent. That is unremarkable, since varying the connection is their method. Custom real-user monitoring follows at 4 of 10. The Chrome User Experience Report publishes effective connection type and round trip time as fields a researcher can read directly, and of the six works using it, one analyses either. None of the three works using the HTTP Archive considers the network. Lab tool studies sit at 13 of 136, which is 9.6 per cent, and works naming no source at all sit at 31 of 172.

The studies holding genuine connection-quality data are the ones least likely to analyse it. That is a specific and checkable claim about how a field uses its own instruments, and it has a straightforward practical reading. A researcher who has gone to the trouble of obtaining field data has usually done so to characterise a site or a sector, and having obtained it, treats the connection as part of the background that the field average already absorbs. The variable is present in the file and absent from the analysis.

Works that do take the connection seriously show what the rest of the literature is missing. A study of geoportal performance in rural areas treated connection quality as the premise of the research question (Krol & Sroka, 2023). An analysis of client-side rendering in a hybrid learning management system measured under several network conditions and reported them separately (Putri Cahyono et al., 2026). A comparison of two rendering frameworks examined global network equity explicitly as an outcome (Pati & Zaki, 2025). Transport-level studies necessarily vary the network because the effect under study lives there (Castell-Uroz et al., 2022; Gupta & Bartos, 2022; Kampanakis & Childs-Klein, 2024).

4.4 The precursor set, and what changed

The 22 works that use these metric names before May 2020 were set aside from the analytic corpus and they turn out to be informative. Their composition is almost the inverse of the corpus that follows them. Ten are conference papers and eight are journal articles, and the subject matter is dominated by network measurement, with studies of Tor circuit and relay selection, web latency in cellular networks, content distribution, wireless mesh gateway selection, and crowdsourced perceived performance on retail pages. Measuring web latency and rendering performance together, with a longitudinal dataset released alongside, is what one of them did in 2019.

Coding the precursor set with the same published rules puts a number on the difference. Connection quality enters the analysis in 12 of the 22 precursor works, which is 54.5 per cent, against 15.7 per cent in the 357-work analytic corpus. No precursor work names a lab tool under the strict rule, against 136 that do in the corpus, while 18 of 22 name no source, which is a higher share of unsourced work than the corpus carries. Before the metrics were named by the programme, the people measuring them were largely network measurement researchers who varied the connection because that was their subject.

Three cautions belong with this comparison and they limit what it can support. The precursor set holds 22 works, so a difference of two or three records moves the percentage several points. It was retrieved by the same seven phrase queries, which means it is the pre-2020 literature that happens to use this vocabulary and not a sample of pre-2020 web performance research. One record is indexed to 1970, which is a registry date artefact for an Indonesian article and not a 1970 publication, and it is retained here only because the precursor set is reported as retrieved. The comparison is offered as a suggestive contrast and not as a measured change over time.

What the contrast suggests is that the network was displaced and not forgotten. The programme gave the field a shared vocabulary and a set of tools that made page-level auditing easy for anyone, and the research that followed took up both. The connection

variable did not leave because it stopped mattering. It left because the instrument that became standard does not require a researcher to think about it.

4.5 Which populations and which authors (RQ5)

The literature has a geography that nobody has described, and the two ways of measuring it disagree in an informative way. Only 32 works of 357 name the country or region of the population they studied, which is the first result and a limitation on all the others. Among those 32, Indonesia appears 14 times, the United States three times, and India and Bangladesh twice each. Nigeria appears once, in a study of university website performance and accessibility (Ogbuju et al., 2022). Figure 5 reports both measures, with study population and author institution kept in separate panels.

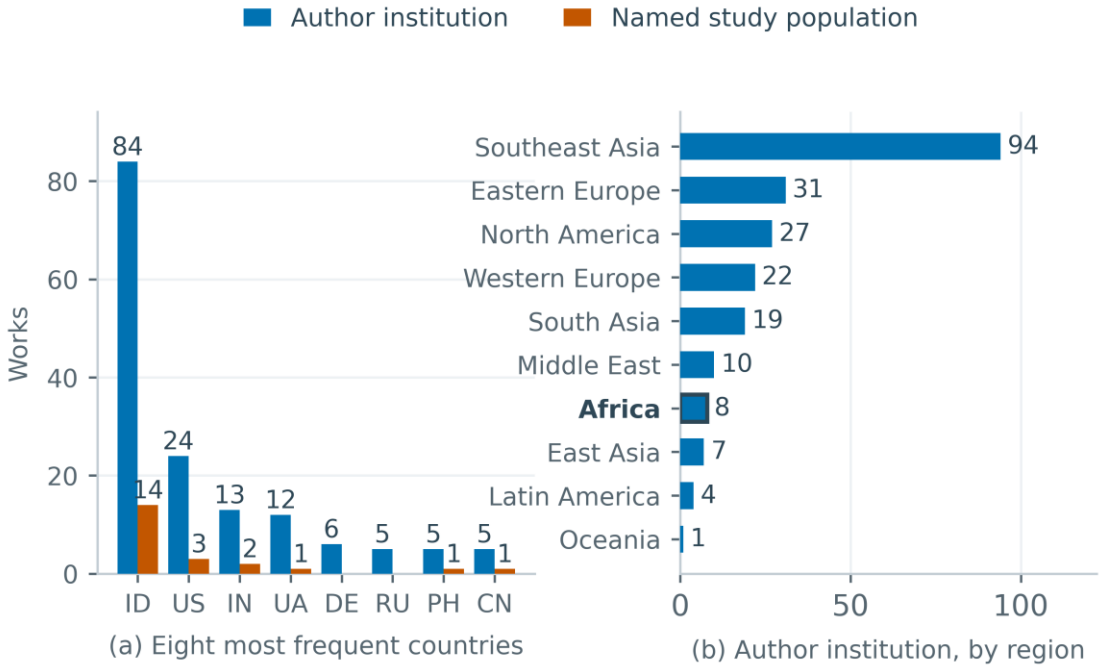


Figure 5. Geographic attribution, by named study population and by author institution.

Note. The two measures are counted separately and never combined, because an author's institution is not the population the author studied. Panel (a) covers the eight countries most frequent by author institution, drawn from the 207 works carrying affiliation data and the 32 works naming a study population. Panel (b) rolls author institutions up to region across all 48 countries present, with Africa outlined and labelled in bold. Affiliation data are missing for 150 works, so panel (b) describes the records that carry affiliations and not the field.

Affiliation records tell a stronger story. Of the 207 works carrying affiliation data, 84 have an Indonesian institution, which is 40.6 per cent, and 24 have a United States institution. The Indonesian share is not an artefact of one productive venue or one prolific group, since the work is distributed across many small Indonesian journals. It is also stable. Table 5 reports the share by year and it has held between 40 and 47 per cent in every year since 2023.

Table 5. Indonesian share of works carrying affiliation data, by year.

Year	Works with affiliation data	With an Indonesian institution	Share
2020	4	0	0.0%
2021	6	0	0.0%
2022	16	4	25.0%
2023	17	8	47.1%
2024	30	12	40.0%
2025	55	26	47.3%
2026	79	34	43.0%
All years	**207**	**84**	**40.6%**

Note. Denominators are works whose registry record carries at least one institutional affiliation, which is 207 of the 357 works. Affiliation data are missing for the remaining 150, and whether those works are distributed like the ones recorded is not knowable from the registry. The share is therefore a share of the recorded works and not of the field.

Eight works in the whole corpus carry an African institution, three of them Nigerian, two South African, and one each Algerian, Kenyan and Tunisian. Two works name an African study population. The African work that exists spans a university website evaluation (Ogbuju et al., 2022), a quality-of-experience study using engagement metrics (Abdelwahed et al., 2020), a study of web metrics for an instructional content repository (Abah & Chinaka, 2024), and a performance optimisation of a clinical dashboard for newborn care in low-resource hospitals (Thomas et al., 2026). Against 357 works and against the share of global internet users these markets hold, eight is a countable absence.

Composition tracks method, which is why it matters for reading the findings. Works with an Indonesian affiliation use lab tools in 57 of 84 cases, which is 67.9 per cent, against 28 of 123 for the rest of the corpus carrying affiliation data, which is 22.8 per cent. They consider the network in 9.5 per cent of cases against 18.7 per cent. They name no data source in 27.4 per cent of cases against 54.5 per cent, so the pattern is not uniformly worse and on traceability it is better. A reader who assumes this is a North American and European field reading a North American and European web has the picture wrong in a way that changes how the aggregate numbers should be read.

One limit belongs with this result and not in a later section. Affiliation data are missing for 150 of the 357 works, so the Indonesian share is a share of the records that carry affiliations and not of the field. Whether works with missing affiliations are distributed like those with them is not knowable from the registry. The claim this paper makes is therefore about the works whose affiliations are recorded, and it does not extend further.

4.7. What is being measured

The artefact under study is usually unspecified. Table 6 reports the application domain. For 247 works the abstract identifies no particular domain and the study is of websites in general or of a system whose sector is not named. Electronic commerce is the largest identified domain at 47 works, followed by library and repository systems at 19, education at 10, health at 9 and government at 6. News and media sites account for 2.

Table 6. Application domain of the artefact studied.

Domain	Works	Share of 357
General or unspecified	247	69.2%
Electronic commerce	47	13.2%
Library and repository systems	19	5.3%
Education	10	2.8%
Health	9	2.5%
Government	6	1.7%
News and media	2	0.6%
No abstract available	21	5.9%

Note. A work naming two sectors is counted under both, so the shares sum to more than one hundred per cent. General or unspecified means the abstract identifies no particular sector for the artefact studied, which is the most common case and is itself a reporting gap.

Electronic commerce at 13.2 per cent is low for a metric set whose business case rests substantially on conversion. Work connecting page speed to touchpoints in a retail customer journey exists (Nichifor et al., 2021), as does measurement of marketplace performance (Muna et al., 2022) and of e-commerce sites under constrained conditions (Hossain et al., 2021), but the commercial setting where the metrics are most often justified is not where most of the measurement happens. University and institutional websites are heavily represented, which is consistent with convenience of access, since any researcher can point a lab tool at a university homepage (Al Khuwaitem & Hasan, 2026; Ogbuju et al., 2022).

5. Discussion

The four results reported above are usually read as four separate gaps, and reading them that way understates the problem. They share a cause. A lab tool can be pointed at any public URL without the site owner's cooperation, it needs no instrumentation on the page and no agreement with anyone, it runs in under a minute, and it returns a number that looks like the numbers in published papers. Nothing else available to this field has those properties. The traceability gap, the missing third metric, the absent network

variable and the geographic composition are the outcomes that follow when a research community settles on an instrument that is free, immediate and asks nobody's permission.

That settlement was not anybody's error and no individual study in this corpus is at fault for it. A researcher with no platform access, no cloud billing account and no agreement with a site owner has very few ways to measure a web page at all, and the one way that works is the one most of this literature took. The pattern only becomes a problem in aggregate, which is exactly where a mapping study can see it and a single primary study cannot.

Traceability is the first face and the cheapest to repair. A paper reporting a Largest Contentful Paint figure without naming the instrument, the connection, the device or the sample has given a reader nothing to evaluate, and 123 of the 266 measuring works in this corpus are in that position. One sentence would move each of them from untraceable to traceable. The cost of leaving it unrepaired falls on secondary research, because a review pooling this literature would inherit all 123 and the inherited gap would not be visible in the pooled result. That evidential problem has been described in general terms for software engineering secondary studies (Rainer & Wohlin, 2024), and this corpus is a large concrete instance of it.

The second face is sharper, because here the instrument does not merely fail to record something; it decides what the field is able to study at all. Lighthouse loads a page in a simulated environment with no user in it, so it cannot measure Interaction to Next Paint and reports Total Blocking Time in its place (Walton, 2024). The counts follow the tool exactly. Nine lab-tool works name the interactivity metric and sixty-one name the proxy, while among the sixteen works using a field or archive source the ordering reverses. Across the whole corpus only fifteen works of 357 report all three currently stable Core Web Vitals together. A measurement programme that defines three metrics has been reduced to two in the research literature, and the reduction happened through tooling and not through any argument about which metrics matter.

What that removes is not a minor dimension. Responsiveness during a session is the part of web performance that correlates most directly with frustration on a slow device, and it is the part the programme added a metric for precisely because the original one measured too little. The field has therefore been studying loading and visual stability carefully while substituting a laboratory proxy for the hardest third of the problem.

The third face is the one this paper would most like a reader to carry away, because it runs opposite to prediction. Connection quality enters 56 works of 357, and the studies least likely to analyse it are the ones whose data actually record it. Synthetic studies consider the network in 60.7 per cent of cases because varying it is their method, while one of the six works using the Chrome User Experience Report analyses the effective connection type and round trip time that dataset publishes as fields, and none of the three using the HTTP Archive considers the network at all. The variable sits in the file and stays out of the analysis. Throttling a laboratory connection to a nominal slow profile tests a model of a slow connection, and reading effective connection type from field records tests the connections people actually had. The field is doing a great deal of the first and very little of the second.

The precursor comparison suggests how the variable went missing. Works using these metric names before the programme existed considered connection quality in 12 of 22 cases, against 15.7 per cent in the corpus that followed, and none of those 22 named a lab tool. The small denominator and the shared retrieval route mean that contrast cannot carry much weight by itself. What it points to is a mechanism worth testing properly, which is that a shared vocabulary and an accessible tool brought a far larger community into this subject, and that community does not come from network measurement. The network was displaced by the field widening and not abandoned by anyone who had been studying it.

Evidence from network measurement research says why the displacement matters. Latency attribution work in Ghana found the dominant contributors to delay were not the ones a page-level laboratory test would surface (Zaki et al., 2014). Device characterisation in lower-income markets documented hardware populations far from any default test profile (Ahmad et al., 2016). Work on reducing JavaScript payloads for constrained users has treated the connection and the device as the design premise and reached results a page-weight study under default conditions would not have predicted (Chaqfeh et al., 2020; Chaqfeh et al., 2022). Performance guidance validated only under controlled conditions should not be assumed to hold where conditions vary most, and the markets where they vary most are the ones this literature has least to say about.

Geography is the fourth face and the one that behaves differently from the other three, so it needs stating with more care. Of the 207 works carrying affiliation data, 84 have an Indonesian institution and eight an African one, and those works use lab tools in 67.9 per cent of cases against 22.8 per cent for the rest of the affiliated corpus. Method and composition are entangled, and this study cannot say which way the entanglement runs. A plausible reading is that an instrument requiring no funding and no platform access lowered the barrier to entry most where funding and access were scarcest, which would make the geographic pattern a consequence of the same accessibility that produced the other three. An equally plausible reading is that local publishing incentives and venue availability drive both. The data distinguish neither, and the honest position is to report the association and leave the direction open.

Two things follow from the composition even without resolving its cause. The high lab-tool share in this corpus is partly a geographic artefact, so a reader who treats the 38.1 per cent figure as a uniform property of the field will be wrong about several of its subcommunities. And the absence on the other side is a countable fact with consequences, since eight works in 357 carry an African institutional affiliation and two name an African study population, in a literature whose metrics are applied to the whole

web by a search engine operating in every market. Naming that absence is not a criticism of the authors who are present, and it identifies where the next contribution to this field would be worth the most.

Taken together the four faces describe an evidence base narrower than the guidance drawn from it. A recommendation about image loading, script deferral or font delivery, supported by a study that ran Lighthouse once against a public URL, is evidence about one simulated visit under the tool's default profile. That is useful for catching a regression, which is the purpose the programme's own documentation assigns to lab tools, and it is not evidence about what a population of real users experiences (Walton, 2024). When 136 works supply the former and sixteen the latter, what the field has accumulated is a large body of regression checks being cited as a body of performance findings.

Four changes would correct this and three of them cost almost nothing. The first is a named data source in every paper reporting a performance number, stating the instrument, the connection and the sample. The second is explicit reporting of which metrics a study covers and which its instrument cannot reach, so that a reader does not have to infer from a tool's limitations what the paper could not measure. The third is the connection as a reported variable wherever the data already hold it, which costs one additional column for anyone using field records. The fourth is harder and worth more, which is primary measurement in the markets the literature has not reached.

Editors and reviewers are the shortest route to the first three. A submission reporting unsourced performance figures can be returned with a single request, and the quality assessment instruments developed for secondary studies in software engineering already treat traceability of this kind as a criterion (Usman et al., 2023). For practitioners the reading is simpler. Guidance from this literature should be treated as provisional outside the conditions a lab tool assumes, and a site serving users on constrained connections should check performance claims against its own field data before acting on them. None of this requires the field to abandon the instrument it has. It requires the field to say what the instrument saw.

6. Threats to validity

Threats are reported against the categories used for software engineering secondary studies (Ampatzoglou et al., 2019). Each is stated with what was done about it and what remains unresolved, because a threat listed without a mitigation and a residual is not information.

6.1 Construct validity

The chief construct threat is that phrase presence in a title or abstract is not the same as a work belonging to the field being mapped. A paper may mention Largest Contentful Paint in passing while studying something else, and eight such papers were found and removed by hand at the relevance gate. The reverse error is also possible and less detectable, since a work may study exactly this subject while naming the metrics only in its full text. The year floor at May 2020 addresses a related threat, since the metric names predate the programme, and the 22 precursor works are counted separately for that reason.

A second construct threat concerns what network consideration means. The code records whether connection quality enters the analysis, which a mention of throttling or of effective connection type in the abstract will satisfy. A work that varied the connection carefully and did not say so in its abstract is coded as not considering it. The direction of this error is known, since it biases the network share downward, and the magnitude is not.

6.2 Internal validity

The absence of a second coder is the most serious limitation of this study and it is stated without qualification. No inter-rater agreement statistic is reported, because none was computed, and a reader is entitled to treat every code in this paper as the output of a single coding procedure that nobody independent checked. Three mitigations apply and none of them is equivalent to a second coder. The rules were fixed and published before coding began, so the procedure is inspectable. Coding was automated against those rules, so it is deterministic and reruns identically instead of drifting across sessions. The coded dataset is deposited in full, so any reader can recode any dimension and compare.

The cost of the missing second coder is not uniform across the six dimensions, and saying so is more useful than a blanket caveat. Data source, metric and year are close to mechanical, since they turn on the presence of named tools, named metrics and a registry field, and a second coder would be expected to agree with the script almost always. Evidence type and network are more interpretive, since both depend on reading what an abstract claims to have done, and these are where independent coding would most likely diverge. Application domain sits between the two. The findings that rest on the mechanical dimensions are therefore the sturdier ones, and the network inversion in Section 4.4 is the result most exposed to this limitation.

The second internal threat is the abstract-only basis of the coding. A work that used the Chrome User Experience Report without saying so in its abstract is coded as naming no data source. This biases the data-source counts downward and inflates the 123 works reporting measurements without a stated source. The direction is known and the size is not. Full-text checking of the 279 open access works would resolve it and is the obvious extension if the result is contested. The claim as stated is about what works report in their abstracts, which is also what most readers and most automated indexes will see.

A third internal threat concerns the automated relevance screen. It flagged twelve records for manual decision, and a record that the pattern matched was never read by hand. A web-domain term appearing in an off-topic paper would therefore pass

unexamined. The pattern was deliberately broad to push errors toward manual review and away from silent exclusion, which trades precision at this gate for recall.

6.3 External validity

Coverage of the registry bounds what this study can claim. OpenAlex indexes a different set of works from Scopus and the Web of Science, with broadly comparable reference coverage and a different composition (Culbert et al., 2025). Works indexed in neither OpenAlex nor anywhere else, including industrial measurement reports and much engineering-blog material, are outside the frame entirely, and some of the most careful field measurement in this area is published that way. The corpus is the indexed scholarly literature and the paper's claims do not extend past it.

Backward snowballing was not performed, which is the clearest recall gap (Wohlin, 2014). Works that discuss these metrics without naming them in a title or abstract are missing, and so are works that cite the corpus without using its vocabulary. A snowballed extension would raise recall and would also require a new boundary rule, since citation links reach quickly into general page-load research that predates the programme.

The partial 2026 year limits any statement about trend. One hundred and fifty-two works carry a 2026 date at a retrieval point ten months into that year, and registry indexing lags publication unevenly across venues. No growth rate is fitted in this paper for that reason, and the year counts are reported as counts.

6.4 Conclusion validity

Every count in this paper was recomputed from the deposited dataset during drafting, and the figures are generated directly from it and not transcribed, so a figure cannot disagree with the table beside it. No statistical inference is drawn, which removes a class of threat and also limits what the results can say, since differences between groups are reported as the differences observed in this population and are not generalised to a wider one. Where a percentage rests on a small denominator, such as the six works using the Chrome User Experience Report, both the count and the denominator are given so that the reader can judge the weight to put on it.

7. Conclusion

The field that measures web performance for the world measures it with simulated single visits. Of the 357 works in this corpus, 136 name a lab tool and 16 name a real-user or archive source. Nearly half the works reporting measurements, 123 of 266, name no data source a reader could trace. Connection quality enters 56 works, and least often in the studies whose data actually record it. Only 15 works report all three of the metrics the programme currently defines, because the dominant instrument cannot measure one of them. Of the works whose affiliations are recorded, 84 of 207 sit in Indonesia and eight in Africa.

Each of those statements is a count from a deposited dataset and each can be checked. Together they describe an evidence base that is narrower than the guidance drawn from it, and the narrowing happened through the ordinary convenience of the tools available and not through anybody's error. The cheapest correction available to the field is a named data source in every paper that reports a number, and editors can require it immediately. The most valuable correction is field measurement in the markets where the connection is the binding constraint, and that work has barely started.

Three extensions follow directly from the limits stated in Section 6. Full-text coding of the 279 open access works would settle how far the abstract-only basis has understated field data use. Independent recoding of the evidence type and network dimensions by a second coder would test the two results most exposed to the single-coder design. Primary longitudinal measurement of client-side complexity in underrepresented markets, drawable from public web archive records without platform access or paid infrastructure, would begin to fill the geographic gap this map has made visible.

Statements and Declarations

CRediT Authorship Contribution Statement: Harrison Oke Ekpobimi: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Data curation, Visualisation, Writing, original draft, Writing, review and editing. Adepegba Akindayomi Akintade: Validation, Writing, review and editing. Wasiu Abiodun Bello: Validation, Writing, review and editing.

Data Availability: The coded dataset for all 357 works, the recorded screening decisions with their reasons, the list of removed duplicate titles, the relevance flags, and the five scripts that produced them in execution order are deposited with this paper. The scripts run end to end against the public OpenAlex API, so the study can be reproduced from the registry and not only recomputed from the deposited table. Retrieval counts will differ from those reported here if the scripts are rerun at a later date, because the registry continues to index.

ORCID iD: Harrison Oke Ekpobimi 0009-0002-1870-5466; Adepegba Akindayomi Akintade 0009-0001-0750-934X; Wasiu Abiodun Bello 0009-0000-6037-6956

Funding: This research received no external funding, and the article processing charge was funded by the authors.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Abah, J. A., & Chinaka, T. W. (2024). Web metrics and key performance indices of the Villagemath instructional content repository. *Interdisciplinary Journal of Education Research*, *6*, 1-21. https://doi.org/10.38140/ijer-2024.vol6.27

[2] Abdelwahed, N., Ben Letaifa, A., & El Asmi, S. (2020). How user engagement metrics ameliorate the web QoE? *Wireless Personal Communications*, *117*(3), 2383-2402. https://doi.org/10.1007/s11277-020-07980-1

[3] Ahmad, S., Haamid, A. L., Qazi, Z. A., Zhou, Z., Benson, T., & Qazi, I. A. (2016). *A view from the other side: Understanding mobile phone characteristics in the developing world. In Proceedings of the 2016 Internet Measurement Conference (pp. 319-325)*. ACM. https://doi.org/10.1145/2987443.2987470

[4] Al Khuwaitem, M., & Hasan, L. (2026). Clustering university websites based on usability metrics: Insights from web vitals analysis. *Social Sciences and Humanities Open*, *13*, 102608. https://doi.org/10.1016/j.ssaho.2026.102608

[5] Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M., & Chatzigeorgiou, A. (2019). Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Information and Software Technology*, *106*, 201-230. https://doi.org/10.1016/j.infsof.2018.10.006

[6] Banaybanay, R. J., & Ricafort, R. B. (2026). Content delivery readiness of Philippine higher education institution websites: Evidence from HTTP Archive and Chrome UX Report. *International Journal of Latest Technology in Engineering, Management and Applied Science*, *15*(6), 2916-2940. https://doi.org/10.51583/ijltemas.2026.150600214

[7] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, *80*(4), 571-583. https://doi.org/10.1016/j.jss.2006.07.009

[8] Campoverde-Molina, M., Lujan-Mora, S., & Garcia, L. V. (2020). Empirical studies on web accessibility of educational websites: A systematic literature review. *IEEE Access*, *8*, 91676-91700. https://doi.org/10.1109/ACCESS.2020.2994288

[9] Castell-Uroz, I., Sanz-Garcia, R., Sole-Pareta, J., & Barlet-Ros, P. (2022). Demystifying content-blockers: Measuring their impact on performance and quality of experience. *IEEE Transactions on Network and Service Management*, *19*(3), 3562-3573. https://doi.org/10.1109/TNSM.2022.3179267

[10] Chaqfeh, M., Haseeb, M., Hashmi, W., Inshuti, P., Ramesh, M., Varvello, M., Subramanian, L., Zaffar, F., & Zaki, Y. (2022). *To block or not to block: Accelerating mobile web pages on-the-fly through JavaScript classification. In International Conference on Information and Communication Technologies and Development 2022 (pp. 1-12)*. ACM. https://doi.org/10.1145/3572334.3572397

[11] Chaqfeh, M., Zaki, Y., Hu, J., & Subramanian, L. (2020). *JSCleaner: De-cluttering mobile webpages through JavaScript cleanup. In Proceedings of The Web Conference 2020 (pp. 763-773)*. ACM. https://doi.org/10.1145/3366423.3380157

[12] Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., & Donner, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*, *130*(4), 2475-2492. https://doi.org/10.1007/s11192-025-05293-3

[13] Egger, S., Hossfeld, T., Schatz, R., & Fiedler, M. (2012). *Waiting times in quality of experience for web based services. In 2012 Fourth International Workshop on Quality of Multimedia Experience (pp. 86-96)*. IEEE. https://doi.org/10.1109/QoMEX.2012.6263888

[14] Gupta, A., & Bartos, R. (2022). *User experience evaluation of HTTP/3 in real-world deployment scenarios. In 2022 25th Conference on Innovation in Clouds, Internet and Networks (pp. 17-23)*. IEEE. https://doi.org/10.1109/ICIN53892.2022.9758130

[15] Harder, R. (2024). Using Scopus and OpenAlex APIs to retrieve bibliographic data for evidence synthesis: A procedure based on R. *MethodsX*, *12*, 102601. https://doi.org/10.1016/j.mex.2024.102601

[16] Hossain, M. T., Hassan, R., Amjad, M., & Rahman, M. A. (2021). Web performance analysis: An empirical analysis of e-commerce sites in Bangladesh. *International Journal of Information Engineering and Electronic Business*, *13*(4), 47-54. https://doi.org/10.5815/ijieeb.2021.04.04

[17] Huotala, A., Kuutila, M., & Mantyla, M. (2025). Research artifacts in secondary studies: A systematic mapping in software engineering. *Information and Software Technology*, *187*, 107830. https://doi.org/10.1016/j.infsof.2025.107830

[18] Kampanakis, P., & Childs-Klein, W. (2024). *The impact of data-heavy, post-quantum TLS 1.3 on the time-to-last-byte of web connections. In Proceedings 2024 Workshop on Measurements, Attacks, and Defenses for the Web*. Internet Society. https://doi.org/10.14722/madweb.2024.23010

[19] Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering (EBSE Technical Report EBSE-2007-01). Keele University and University of Durham.

[20] Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., & Linkman, S. (2010). Systematic literature reviews in software engineering: A tertiary study. *Information and Software Technology*, *52*(8), 792-805. https://doi.org/10.1016/j.infsof.2010.03.006

[21] Krol, K., & Sroka, W. (2023). Internet in the middle of nowhere: Performance of geoportals in rural areas according to Core Web Vitals. *ISPRS International Journal of Geo-Information*, *12*(12), 484. https://doi.org/10.3390/ijgi12120484

[22] Lin, H., & Liu, W. (2025). Symmetry-aware causal-inference-driven web performance modeling: A structure-aware framework for predictive analysis and actionable optimization. *Symmetry*, *17*(12), 2058. https://doi.org/10.3390/sym17122058

[23] Macakoglu, S. S., & Peker, S. (2022). *Web accessibility performance analysis using web content accessibility guidelines and automated tools: A systematic literature review. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (pp. 1-8)*. IEEE. https://doi.org/10.1109/HORA55278.2022.9799981

[24] Muna, S. S., Nurdin, N., & Taufiq, T. (2022). Tokopedia and Shopee marketplace performance analysis using metrix Google Lighthouse. *International Journal of Engineering, Science and Information Technology*, *2*(3), 106-110. https://doi.org/10.52088/ijesty.v2i3.312

[25] Nichifor, E., Lixandroiu, R. C., Chitu, I. B., Bratucu, G., & Trifan, A. (2021). How does mobile page speed shape in-between touchpoints in the customer journey? A research regarding the most trusted retailers in Romania. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(5), 1369-1389. https://doi.org/10.3390/jtaer16050077

[26] Ogbuju, E., Ayodeji, B., & Azeez, A. (2022). *Performance and accessibility evaluation of university websites in Nigeria. In 2022 5th Information Technology for Education and Development (pp. 1-7)*. IEEE. https://doi.org/10.1109/ITED56637.2022.10051461

[27] Olar, A., Picus, O., Cernau, L., Diosan, L., & Motogna, S. (2026). A systematic mapping study method for software architecture research with LLM-assisted quality assessment. *MethodsX*, *17*, 104113. https://doi.org/10.1016/j.mex.2026.104113

[28] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

[29] Pati, S., & Zaki, Y. (2025). *Evaluating the efficacy of Next.js: A comparative analysis with React.js on performance, SEO, and global network equity. In Companion Proceedings of the ACM on Web Conference 2025 (pp. 1239-1243)*. ACM. https://doi.org/10.1145/3701716.3715565

[30] Petalotis, C., Krumpak, L., Floroiu, M. S., Ahmad, L. F., Athreya, S., & Malavolta, I. (2024). An empirical study on the performance and energy costs of ads and analytics in mobile web apps. *Information and Software Technology*, *166*, 107370. https://doi.org/10.1016/j.infsof.2023.107370

[31] Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). *Systematic mapping studies in software engineering. In Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (pp. 68-77)*. BCS Learning and Development. https://doi.org/10.14236/ewic/ease2008.8

[32] Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, *64*, 1-18. https://doi.org/10.1016/j.infsof.2015.03.007

[33] Prasetyo, M. T., & Rokhman, B. I. (2024). Standardizing UX dashboards: A framework integrating usability metrics, benchmarks, and glossaries for managerial insight. *Data: Journal of Information Systems and Management*, *2*(1), 51-61. https://doi.org/10.61978/data.v2i1.927

[34] Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (arXiv:2205.01833)*. arXiv. https://doi.org/10.48550/arXiv.2205.01833

[35] Putri Cahyono, N. P., Hoiriyah, H., Darmawan, A. K., & Rofiuddin, R. (2026). Performance analysis of client-side rendering in a hybrid LMS using Core Web Vitals under diverse network conditions. *SISTEMASI*, *15*(7), 2602. https://doi.org/10.32520/stmsi.v15i7.6500

[36] Rainer, A., & Wohlin, C. (2024). Reporting case studies in systematic literature studies: An evidential problem. *Information and Software Technology*, *174*, 107501. https://doi.org/10.1016/j.infsof.2024.107501

[37] Thomas, J., Jenkins, G., Chen, J., Ogero, M. O., Malla, L., Hirschhorn, L. R., Richards-Kortum, R., Oden, Z. M., Bohne, C., & Wainaina, J. (2026). *Performance optimization of an R Shiny-based digital health dashboard for monitoring small and sick newborn care in low-resource hospital settings [Preprint]*. medRxiv. https://doi.org/10.64898/2026.03.08.26347893

[38] Usman, M., Bin Ali, N., & Wohlin, C. (2023). A quality assessment instrument for systematic literature reviews in software engineering. *e-Informatica Software Engineering Journal*, *17*(1), 230105. https://doi.org/10.37190/e-Inf230105

[39] van Riet, J., Malavolta, I., & Ghaleb, T. A. (2023). Optimize along the way: An industrial case study on web performance. *Journal of Systems and Software*, *198*, 111593. https://doi.org/10.1016/j.jss.2022.111593

[40] van Riet, J., Paganelli, F., & Malavolta, I. (2020). *From 6.2 to 0.15 seconds: An industrial case study on mobile web performance. In 2020 IEEE International Conference on Software Maintenance and Evolution (pp. 746-755)*. IEEE. https://doi.org/10.1109/ICSME46990.2020.00084

[41] Walton, P. (2024, October 31). *Web Vitals. web.dev*. https://web.dev/articles/vitals

[42] Wehner, N., Amir, M., Seufert, M., Schatz, R., & Hossfeld, T. (2022). *A vital improvement? Relating Google's Core Web Vitals to actual web QoE. In 2022 14th International Conference on Quality of Multimedia Experience (pp. 1-6)*. IEEE. https://doi.org/10.1109/QoMEX55416.2022.9900881

[43] Wehner, N., Seufert, M., Schatz, R., & Hossfeld, T. (2023). Do you agree? Contrasting Google's Core Web Vitals and the impact of cookie consent banners with actual web QoE. *Quality and User Experience*, *8*(1), 3. https://doi.org/10.1007/s41233-023-00058-3

[44] Wohlin, C. (2014). *Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (Article 38)*. ACM. https://doi.org/10.1145/2601248.2601268

[45] Zaidan, A. J., & Suyatno, D. F. (2025). Rendering performance analysis of Astro JS, Next JS, Nuxt JS, and SvelteKit frameworks using Google Lighthouse, PageSpeed Insight, and JMeter. *Journal of Emerging Information Systems and Business Intelligence*, *6*(1). https://doi.org/10.26740/jeisbi.v6i1.64283

[46] Zaki, Y., Chen, J., Potsch, T., Ahmad, T., & Subramanian, L. (2014). *Dissecting web latency in Ghana. In Proceedings of the 2014 Conference on Internet Measurement Conference (pp. 241-248)*. ACM. https://doi.org/10.1145/2663716.2663748

[47] Zein, S., Salleh, N., & Grundy, J. (2023). Systematic reviews in mobile app software engineering: A tertiary study. *Information and Software Technology*, *164*, 107323. https://doi.org/10.1016/j.infsof.2023.107323