From Publishing to Citability: AI-Search Content-Readiness Signals Across Kenyan Organisational Websites

A secondary analysis of a cross-sector organisational website audit

Hezron Ochiel

Hezron Insights, Nairobi, Kenya

Study date: 21 August 2026

Abstract

AI-powered search increasingly retrieves, synthesises and cites web content rather than presenting only ranked links. This study examines whether content published by Kenyan organisations exhibits measurable source conditions that may support reliable retrieval, interpretation and citation in generative search. It is a secondary analysis of a frozen cross-sector audit of 244 organisations spanning universities, national ministries, government agencies, county governments, financial services, media, health, listed companies and public-benefit/non-governmental organisations. Seven pre-specified content-readiness conditions were analysed: meaningful taxonomy, explicit provenance, absence of template residue, numerical integrity, control of substantial duplication, entity consistency and evidence of current-year publishing. Indeterminate observations were excluded from substantive denominators and reported separately. Meaningful taxonomy was observed in 204/222 determinate cases (91.9%), entity consistency in 201/230 (87.4%), and evidence of current-year publishing in 199/207 organisations with a verifiable latest-content date (96.1%). Explicit provenance was present in 103/214 determinate cases (48.1%). Template-free content was observed in 165/223 (74.0%), numerical integrity in 176/224 (78.6%), and control of substantial duplication in 185/222 (83.3%). Among 198 organisations with determinate observations for all seven conditions, 60 (30.3%) met all seven simultaneously. In a sensitivity analysis excluding explicit provenance, 99/203 complete cases (48.8%) met the remaining six conditions. Sector patterns were descriptive rather than population estimates: media and listed companies had the highest all-condition conjunctions, while universities and national ministries had the lowest. The findings show that current publishing activity coexists with material gaps in provenance, content hygiene and information integrity. These measures describe source conditions that may support AI-search participation; they do not demonstrate citation, ranking or answer use. Controlled multi-platform citation experiments are needed to test whether the measured conditions predict actual source selection and answer absorption.

Keywords: AI search; generative engine optimization; content readiness; information quality; provenance; Kenya; Africa; generative search; citation; digital communication

1. Introduction

Generative search changes the unit of competition on the web. A page may still compete for a position in ranked results, while also being retrieved as grounding material, selected as a cited source or incorporated into a generated answer. Microsoft now reports URL-level citation activity in Bing Webmaster Tools, and Google describes generative Search as grounded in its existing ranking and quality systems while continuing to emphasise useful, reliable and original content. The shift therefore increases the importance of the source material available for retrieval and synthesis, not only the visibility of the page that contains it (Google Search Central, 2025a, 2025b, 2026; Microsoft Bing Webmaster Blog, 2026).

The emerging Generative Engine Optimisation (GEO) literature examines this source-selection problem directly. Aggarwal et al. (2024) formalised GEO as optimisation for visibility within generative-engine responses, and subsequent work has examined how relevance, completeness, structure, evidential support and trust cues relate to retrieval and citation behaviour (Chen et al., 2025; Martinez, 2026; Vishwakarma et al., 2026; Zhang et al., 2026). This literature is developing quickly, and several recent studies remain preprints. It is therefore useful to separate observed source conditions from stronger claims about universal ranking or citation requirements.

This newer literature also sits within established work on information quality, provenance, web credibility and entity representation. Wang and Strong (1996) conceptualised data quality as extending beyond accuracy to contextual, representational and accessibility dimensions, while Pipino et al. (2002) argued for explicit and usable quality assessment. The W3C provenance model treats provenance as information about entities, activities and agents that can support assessments of quality and trustworthiness (Moreau & Missier, 2013). Web-credibility research shows the importance of source and structural cues (Fogg et al., 2003), and knowledge-graph research highlights the challenge of maintaining coherent entity identity across heterogeneous sources (Hogan et al., 2021). These traditions provide a stronger basis for evaluating organisational source quality than a checklist of platform-specific tactics alone.

For organisations, source quality has direct institutional consequences. Ministries publish policy and service information; hospitals publish patient-facing guidance; universities publish admissions, research and institutional facts; listed companies publish investor information; and NGOs publish programme and impact claims. The Kenyan context is useful because these organisations operate under different publishing conventions and governance structures. In this paper, content-readiness is a bounded descriptive construct: the presence of specified source conditions that may support reliable machine retrieval and interpretation. It is not presented as a validated universal AI-readiness index, a certification standard or proof that an AI system will cite a page. The analysis asks what measurable conditions are present across the frozen cross-sector sample and how often those conditions occur together.

2. Research questions and analytical framework

RQ1. To what extent do audited Kenyan organisational websites exhibit measurable content conditions that can support accurate retrieval, interpretation, verification and citation in AI-powered search?

RQ2. How do these conditions vary descriptively across organisational sectors?

RQ3. How sensitive is an all-condition readiness result to whether explicit provenance is treated as a universal requirement?

The study uses the following evidence-to-outcome pathway:

Structural clarity → Provenance → Content hygiene and factual integrity → Entity coherence → Current publishing evidence → Retrieval → Citation → Answer absorption

The first five stages are represented by the seven measured conditions in this study: taxonomy; provenance; template, numerical and duplication integrity; entity consistency; and evidence of current-year publishing. Retrieval, citation and answer absorption are downstream outcomes and are not observed here. The paper therefore does not test citation frequency or ranking across ChatGPT, Google AI Mode/Gemini, Microsoft Copilot, Claude or Perplexity.

3. Methods

3.1 Design and source dataset

This study is a secondary content-readiness analysis of a previously frozen organisational website audit. The parent dataset contains 244 Kenyan organisations across nine strata. County governments and ministry-level portfolios use census-like frames; other strata use stratified-random, reproducible balanced or purposive samples. The same organisations were retained regardless of performance, retrieval difficulty or missingness. The unit of analysis is the organisational web presence represented in the frozen audit. The study therefore describes the audited sample and does not estimate the prevalence of these conditions across all Kenyan websites. The sector composition and sampling approach are summarised in Table 1.

Table 1. Frozen sample by sector and sampling approach.

Sector	N	Sampling approach
Universities	25	Stratified random pilot: 13 public + 12 private

Sector	N	Sampling approach
National ministries / State Law Office	22	Census-like ministry-level frame
Government agencies	25	Balanced reproducible pilot
County governments	47	Census of all counties
Financial services	25	15-bank random sample + 10 CMA-listed stockbrokers
Media	25	Balanced purposive journalism sample
Health	25	Balanced purposive higher-level hospital sample
Listed companies	25	Sector-stratified reproducible non-financial sample
PBOs/NGOs	25	Balanced purposive thematic sample
Total	244	Frozen cross-sector organisational sample

3.2 Operationalisation of content-readiness conditions

The secondary analysis uses seven variables already collected in the parent audit. They map onto established concerns in information quality, provenance, and entity representation, as well as recurring issues in search guidance and current GEO research (Wang & Strong, 1996; Pipino et al., 2002; Moreau & Missier, 2013; Hogan et al., 2021). The conditions are meaningful taxonomy, explicit provenance, absence of template residue, numerical integrity, control of substantial duplication, entity consistency and evidence of current-year publishing. They are analytical indicators, not universal platform requirements. Table 2 gives the operational rule for each condition.

Table 2. Content-readiness conditions and operational rules.

Condition	Operational rule in this analysis	Why it matters
Meaningful taxonomy	Manual code = Meaningful	Supports interpretable organisation of topics and institutional information.
Explicit provenance	Named human OR institutional department	Provides an identifiable source beyond an anonymous/generic system.
No template residue	Template residue = None	Reduces placeholder, demo or page-builder noise.
Numerical integrity	Questionable numerical values = None	Reduces conflicting, incomplete or implausible numerical claims.
Duplication control	Duplication = Clean OR Limited	Reduces competing machine-readable versions of the same information.
Entity consistency	Entity consistency = Consistent	Supports stable organisational identity and factual coherence.
Evidence of current-year publishing	Latest verifiable content date falls in 2026	Indicates publishing activity during the study year; it does not imply that all pages are current.

Two additional parent-audit variables—web fragmentation and broken/development links—were retained as supporting site-governance evidence but were not included in the seven-condition conjunction because they extend beyond content quality into web-property governance.

3.3 Coding, missingness and denominators

The parent audit used a structured, AI-assisted and evidence-linked coding workflow. ChatGPT supported public-web evidence retrieval and first-pass categorical coding against the project codebook; determinate classifications retained evidence URLs and dated notes in the frozen workbook. All 244 records were frozen on 21 August 2026. Where the available evidence did not support a confident classification, the value was recorded as Indeterminate. In the source dataset, empty cells in risk variables represent the absence of the coded risk category rather than unassessed data;

Indeterminate is an explicit category. The AI-supported workflow was used as a research-audit instrument and was not treated as independent adjudication.

For each condition, prevalence uses the determinate denominator. Indeterminate cases are excluded from the substantive denominator and reported separately. For the conjunction analysis, only organisations with determinate observations on all seven conditions are included. This complete-case approach prevents unresolved observations from being silently treated as failures.

The study does not report inter-coder reliability because a blinded second human coder did not independently recode a stratified subsample. Consequently, the authors report no Cohen's kappa, Krippendorff's alpha, or similar coefficient. Evidence retention, explicit coding rules, scripted integrity checks and sensitivity analysis improve transparency and auditability, but they do not substitute for independent reliability testing (Lombard et al., 2002; Krippendorff, 2004).

3.4 Analytical strategy

Analysis is descriptive. First, each readiness condition is reported as n/N and percentage among determinate cases, with missingness shown separately. Second, sector profiles are compared descriptively. Third, a strict conjunction identifies organisations meeting all seven measured conditions simultaneously. This conjunction is not a weighted AI-readiness score: the conditions remain conceptually distinct, and the calculation asks only whether every measured condition was satisfied. Finally, a sensitivity analysis removes explicit provenance because legitimate institutional publishing may use corporate or departmental authorship conventions that do not map neatly to named-byline expectations.

3.5 Reproducibility and research safeguards

All derived summaries in this revision were recomputed from the frozen workbook using Python 3.13.5 and pandas 2.2.3. A scripted integrity audit confirmed 244 unique organisational IDs, an exact ID and sector-count match across the frozen sample and coding matrix, valid categorical values, no latest-content dates after the common study date, and at least one retained evidence URL for each row with a determinate core content-quality code. The reproduction script and supporting audit outputs form part of the supplementary package.

The analysis was limited to publicly accessible organisational web information and did not involve interaction with individuals or access to private personal data. High-severity observations were treated as time-bounded web findings and re-checked against public source pages before finalisation because organisational websites can change after collection.

4. Results

This section reports the four result domains in sequence: overall prevalence and missingness, joint readiness and sensitivity, sector profiles, and high-confidence content-risk patterns.

4.1 Overall content-readiness conditions

Table 3 summarises the overall prevalence and missingness for the seven measured conditions. The clearest strengths were evidence of current-year publishing, taxonomy and entity consistency. The weakest measured condition was explicit provenance: 103/214 determinate organisations (48.1%) exposed either a named human or an identifiable institutional department in the parent coding. This should not be interpreted as evidence that the remaining organisations lacked legitimate authority; many relied on generic systems, unattributed content or institution-level publishing conventions.

Content hygiene was more uneven. Template-free content was observed in 165/223 determinate cases (74.0%), numerical integrity in 176/224 (78.6%), and control of substantial duplication in 185/222 (83.3%). These results indicate that being current does not eliminate machine-facing quality risks.

Table 3. Overall prevalence and missingness for measured content-readiness conditions.

Condition	n/N	%	Indeterminate (n/244)
Meaningful taxonomy	204/222	91.9%	22 (9.0%)
Explicit provenance	103/214	48.1%	30 (12.3%)
No template residue	165/223	74.0%	21 (8.6%)
No questionable numbers	176/224	78.6%	20 (8.2%)
No substantial duplication	185/222	83.3%	22 (9.0%)
Entity consistency	201/230	87.4%	14 (5.7%)
Evidence of current-year publishing (2026)	199/207	96.1%	37 (15.2% of full sample lacked a verifiable date)

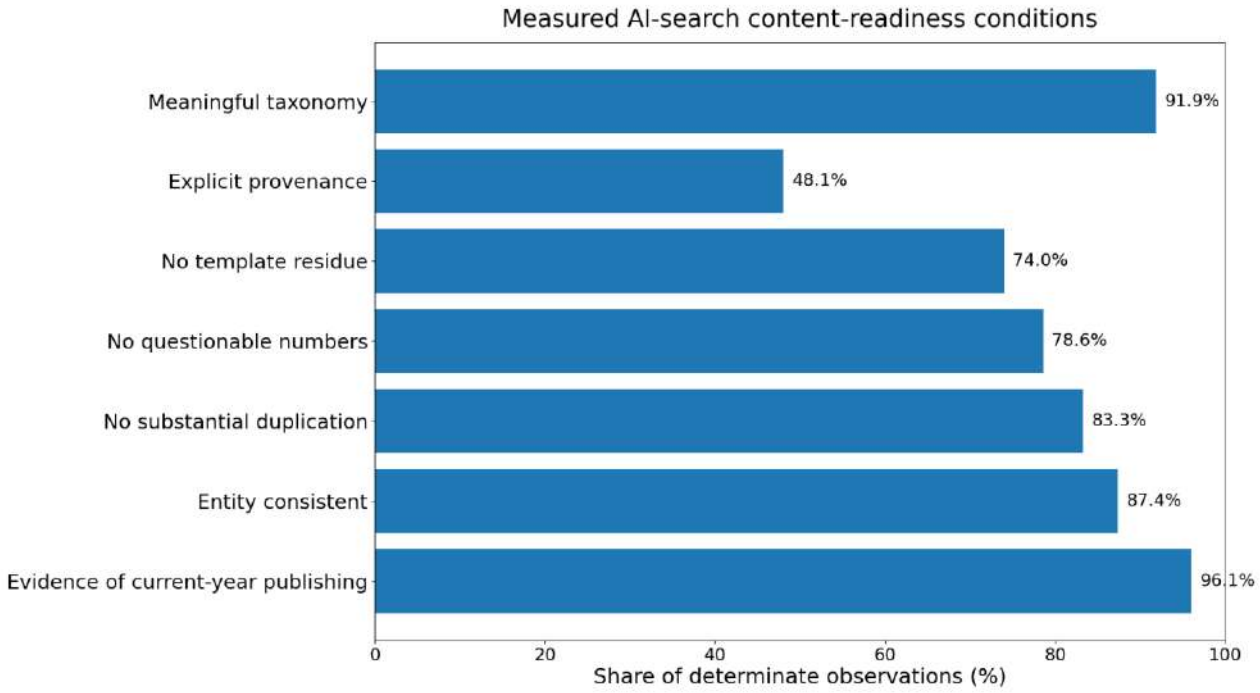

Figure 1. Share of determinate observations meeting each measured content-readiness condition.

4.2 Joint readiness and sensitivity analysis

Among 198 organisations with determinate observations for all seven primary conditions, 60 met every condition (30.3%). The remaining 138 complete cases failed at least one condition. This result is best interpreted as a simultaneous-condition screen rather than a latent score: the design does not assume that every unmet condition has the same consequence for every organisation or content type.

Because provenance conventions differ across content types, the conjunction was repeated without explicit provenance. Among 203 complete cases for the remaining six conditions, 99 (48.8%) met all six. The prevalence therefore rose from 30.3% in the seven-condition complete-case set to 48.8% in the six-condition set. Because the denominators differ (198 versus 203), interpret this comparison as sensitivity to the analytical specification rather than a causal 18.5-percentage-point effect of provenance.

4.3 Sector profiles

Table 4. Descriptive sector profiles with determinate denominators.

Sector	All 7	All 6 excl. provenance	Explicit provenance	Numerical integrity
Universities	2/16 (12.5%)	6/21 (28.6%)	4/17 (23.5%)	17/24 (70.8%)
National ministries	2/16 (12.5%)	5/16 (31.3%)	4/18 (22.2%)	16/19 (84.2%)
Government agencies	5/21 (23.8%)	9/21 (42.9%)	10/25 (40.0%)	17/24 (70.8%)
County governments	6/35 (17.1%)	16/35 (45.7%)	11/37 (29.7%)	32/40 (80.0%)
Financial services	4/20 (20.0%)	12/20 (60.0%)	6/23 (26.1%)	19/23 (82.6%)
Media	13/21 (61.9%)	14/21 (66.7%)	22/24 (91.7%)	23/24 (95.8%)
Health	4/22 (18.2%)	9/22 (40.9%)	6/22 (27.3%)	17/22 (77.3%)
Listed companies	14/24 (58.3%)	14/24 (58.3%)	23/24 (95.8%)	20/24 (83.3%)
PBOs/NGOs	10/23 (43.5%)	14/23 (60.9%)	17/24 (70.8%)	15/24 (62.5%)

'All 7' uses complete cases for the seven-condition conjunction. 'All 6 excl. provenance' is the sensitivity analysis. Provenance and numerical-integrity columns use their own determinate denominators. Table 4 reports n/N (%) so that sector-specific missingness remains visible. Sector contrasts are descriptive because the strata use mixed sampling designs.

Media and listed companies had the highest strict all-condition conjunctions, driven partly by much stronger explicit provenance. PBOs/NGOs combined strong taxonomy, entity consistency and low template residue with weaker numerical integrity. Universities and ministries had the lowest all-condition conjunctions, largely because provenance and, for universities, taxonomy were weaker. Government agencies and health organisations showed notable template-residue gaps.

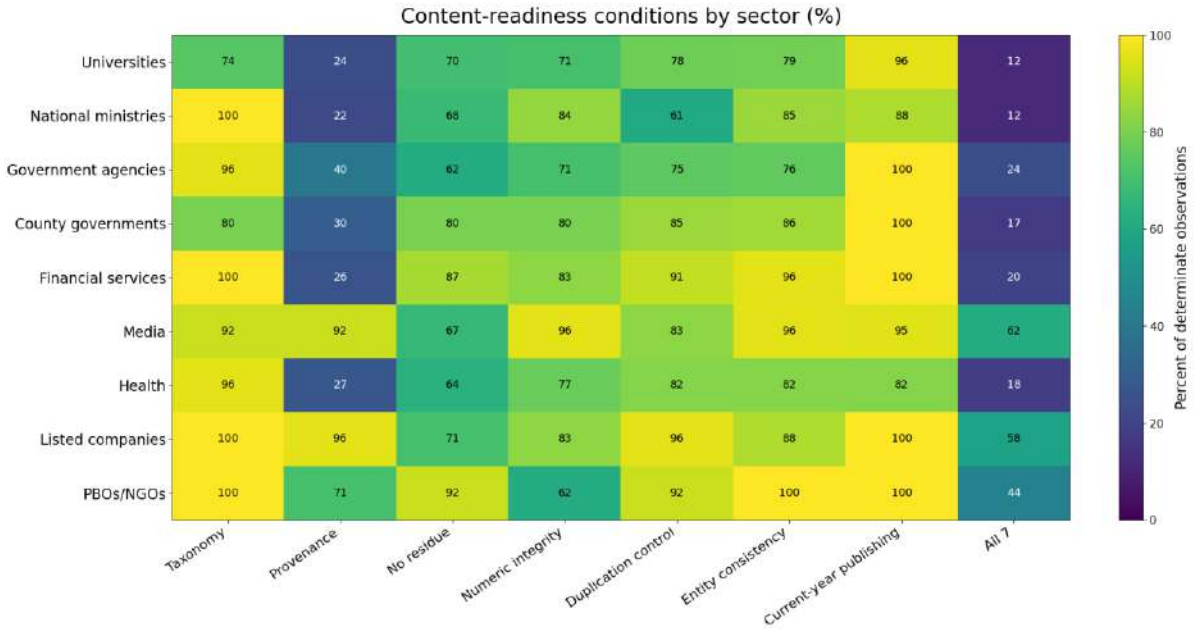

Figure 2. Descriptive sector profile of content-readiness conditions (% of determinate observations).

4.4 High-confidence content risks

A secondary risk analysis counted six high-confidence content-facing signals: weak/mixed taxonomy, material/severe template residue, confirmed numerical conflict, substantial duplication, material entity conflict and latest verifiable content from 2024 or earlier. At least one of these signals appeared in 89/244 organisations (36.5%). Among the 199

organisations with verifiable 2026 content, 74 (37.2%) still exhibited at least one high-confidence content risk. Current publishing activity therefore coexisted with material source-quality problems in more than one third of those organisations.

The six-signal prevalence screen did not include broken/development links. That variable was retained as supporting web-governance evidence for co-occurrence analysis. The most common clusters were material/severe template residue with confirmed broken/development links (17 organisations), material/severe residue with material entity conflict (16), and material/severe residue with substantial duplication (14). The pattern suggests that content-quality defects often cluster around broader publishing-workflow and web-governance problems rather than appearing as isolated editorial errors.

5. Discussion

5.1 Current-year publishing coexists with readiness gaps

The audited organisations were generally active publishers: 96.1% of organisations with a verifiable latest-content date showed evidence of publishing in 2026. Across the complete cases, 30.3% met all seven measured conditions. The main finding is the gap between widespread current publishing and lower joint fulfilment. Recent organisational content in the sample also showed weak provenance, template residue, duplicated blocks, entity inconsistency or questionable numerical claims.

The result aligns with current platform guidance and adds organisational evidence. Google emphasises useful, reliable, and original content; clear source signals; and accurate visible information. Microsoft advises publishers to improve clarity, structure, evidence, and freshness for pages participating in AI-generated answers (Google Search Central, 2025a, 2025b, 2026; Microsoft Bing Webmaster Blog, 2026). The Kenyan data suggest that evidence of current-year publishing is already widespread among sites with a verifiable date. The larger source-quality opportunity lies in provenance, factual integrity and publishing governance.

5.2 Provenance depends on publishing context

Explicit provenance was the weakest measured condition and had the largest effect on the strict conjunction. Media and listed companies performed especially strongly because named authors or clearly attributable corporate publishing were common. Ministries, universities, county governments and financial institutions often relied on institution-level or generic publishing conventions. Google encourages clear authorship where readers would reasonably expect it. Institutional authority may instead be represented through the organisation or responsible department. The sensitivity analysis therefore supports treating provenance as a contextual dimension rather than a universal named-byline requirement.

For practice, the relevant question is whether a user or machine can understand who stands behind a claim. A ministry may appropriately attribute a policy notice to the ministry or responsible department; a hospital clinical explainer may benefit from a named clinical reviewer; a news article ordinarily benefits from a named reporter or desk. Future coding frameworks should distinguish legitimate corporate authorship, accountable departmental authorship, named-human authorship, and truly missing provenance.

5.3 Information integrity is a distinct AI-search problem

The numerical-integrity, template-residue, duplication and entity-consistency findings matter because generative systems operate on retrieved representations rather than an organisation's intended meaning. A zero counter, conflicting statistic, demo contact, placeholder block or duplicated version can become part of the material available for extraction and synthesis. This concern aligns with established information-quality theory, which separates accuracy from representational and contextual quality, and with current GEO studies examining evidential support, completeness, and citation behaviour (Wang & Strong, 1996; Pipino et al., 2002; Vishwakarma et al., 2026; Zhang et al., 2026).

The findings caution against reducing AI-search preparation to schema markup, prompt-targeted FAQs or new technical files. Technical configuration can support discovery and interpretation. Content governance sits upstream: facts need to be current, consequential claims need accountable provenance or evidence, numbers need to reconcile,

and production environments need controls that keep legacy, duplicate and template material from competing with the intended source.

5.4 A practical content-readiness model

The evidence supports a layered model for organisational publishing:

Entity clarity → Useful structure → Identifiable provenance → Factual integrity → Duplication control → Current publishing evidence → Retrieval → Citation → Answer absorption

The first six layers are directly or indirectly represented in the present audit. Retrieval, citation and absorption are downstream outcomes and require platform-level testing. This boundary is important for both scholarship and practice. A content audit identifies source conditions and risks; certification or guarantees of AI citation require direct platform-level evidence.

6. Implications for organisational communication

For communications and digital teams, the results point to an editorial-governance agenda. Organisations should maintain a clear institutional identity across pages; use descriptive headings and meaningful categories; identify responsible authors, departments or institutional owners where appropriate; attach evidence or source references to consequential statistics; review evergreen pages for stale claims; control duplicate and development content; and ensure that major facts reconcile across news, About, service and report pages.

These practices are especially important for health, finance and public-service information, where inaccurate or ambiguous material can carry higher consequences. They also create measurable workstreams for AI-search preparation without depending on speculative platform-specific tactics.

7. Limitations

First, this is a secondary analysis of a cross-sector website audit, not a new page-level corpus sampled specifically for AI-search content research. The parent variables were designed for broader machine-facing information quality and do not capture every potentially relevant characteristic, such as query-level answerability, paragraph-level evidence density, semantic similarity to user prompts or exact FAQ structure.

Second, a structured AI-assisted research workflow produced the coding, and a second human coder did not independently recode it. Accordingly, no Cohen's kappa, Krippendorff's alpha or other inter-coder reliability statistic is reported. This limits claims about reproducibility of judgement-dependent categories across independent coders. Evidence-linked coding, explicit operational rules and scripted quality control improve traceability, while future prospective work should include pre-specified independent recoding of a stratified subsample (Lombard et al., 2002; Krippendorff, 2004).

Third, the sample is not a probability sample of all Kenyan websites. Several strata are purposive or balanced pilots, so sector contrasts are descriptive. Fourth, websites are dynamic, and the findings represent the audit date. Fifth, the seven-condition conjunction is not a validated universal AI-readiness index; it is a transparent descriptive screen, and the provenance sensitivity analysis demonstrates how interpretation changes when one contested dimension is removed.

Sixth, this study does not measure actual citation, ranking or answer absorption across generative-search platforms. Platform behaviour is stochastic, query-dependent and time-varying. A prospective follow-up should use controlled prompts across multiple engines to test whether the measured source conditions predict retrieval, citation selection, citation absorption and factual accuracy.

Finally, this paper draws from the same frozen organisational audit used in a broader website-information-quality study. If both manuscripts are submitted for publication, the shared source dataset, distinct research questions and analytic overlap should be disclosed explicitly to editors and cross-referenced where appropriate to avoid ambiguity about duplicate publication.

8. Conclusion

Among the 244 audited Kenyan organisations, the strongest content-readiness signals were evidence of current-year publishing, meaningful taxonomy and entity consistency. Explicit provenance was the weakest measured condition, while template residue and numerical inconsistency remained material concerns. Sixty of 198 complete cases (30.3%) met all seven measured conditions simultaneously; the sensitivity analysis excluding provenance identified 99 of 203 complete cases (48.8%) meeting the remaining six. These figures describe how often a defined set of source conditions occurred together. They are not estimates of the percentage of Kenyan organisations that are universally 'AI ready'.

The broader conclusion is that current publishing activity and machine-facing source quality are separate dimensions. AI-search preparation depends on editorial clarity, verifiable facts, stable entity representation, accountable provenance and disciplined publishing workflows. The next empirical step is to test whether these conditions predict which Kenyan organisational pages generative-search systems actually retrieve, cite, and incorporate into answers.

9. Declarations

Funding: This research received no external funding.

Competing interests: Hezron Ochiel operates Hezron Insights and may offer website, communication or AI-search advisory services. This activity is analytically separate from the research dataset, sampling, coding rules and interpretation. No organisation's commercial response affected inclusion, coding or reported results.

Data and code availability: The frozen sample, completed coding dataset, codebook, analysis outputs, reproducibility checks and supporting files are included as supplementary materials accompanying this manuscript. The separately developed direct technical crawler protocol did not generate any result reported in this paper.

Use of AI tools: ChatGPT was used as a structured research-audit and editing tool to support public-web evidence retrieval, first-pass categorical coding, analysis checks and manuscript development. The authors remain responsible for the research design, interpretation, factual verification and final manuscript. No AI system is listed as an author, and no inter-coder reliability statistic is reported because independent recoding was not performed.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimisation. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 5–16. https://doi.org/10.1145/3637528.3671900

Chen, M., Wang, X., Chen, K., & Koudas, N. (2025). Generative Engine Optimisation: How to Dominate AI Search. arXiv:2509.08919. https://doi.org/10.48550/arXiv.2509.08919

Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., & Tauber, E. R. (2003). How do users evaluate the credibility of Web sites? A study with over 2,500 participants. Proceedings of the 2003 Conference on Designing for User Experiences, 1–15. https://doi.org/10.1145/997078.997097

Google Search Central. (2025a). Creating helpful, reliable, people-first content. Google for Developers. Updated December 2025. https://developers.google.com/search/docs/fundamentals/creating-helpful-content

Google Search Central. (2025b). Top ways to ensure your content performs well in Google's AI experiences on Search. Google for Developers. https://developers.google.com/search/blog/2025/05/succeeding-in-ai-search

Google Search Central. (2026). Optimising your website for generative AI features on Google Search. Google for Developers. https://developers.google.com/search/docs/fundamentals/ai-optimization-guide

Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), Article 71. https://doi.org/10.1145/3447772

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. Human Communication Research, 30(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. Human Communication Research, 28(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Martinez, O. (2026). Optimising Visibility in Generative Engines: A Critical Survey of Generative Engine Optimisation (2023-2026). arXiv:2607.14035. https://arxiv.org/abs/2607.14035

Microsoft Bing Webmaster Blog. (2026, February 10). Introducing AI Performance in Bing Webmaster Tools Public Preview. https://blogs.bing.com/webmaster/February-2026/Introducing-AI-Performance-in-Bing-Webmaster-Tools-Public-Preview

Moreau, L., & Missier, P. (Eds.). (2013). PROV-DM: The PROV Data Model. W3C Recommendation, 30 April 2013. https://www.w3.org/TR/prov-dm/

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218. https://doi.org/10.1145/505248.506010

Vishwakarma, R., Kumar, S., & Jamidar, R. (2026). What Gets Cited: Competitive GEO in AI Answer Engines. arXiv:2605.25517. https://arxiv.org/abs/2605.25517

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Zhang, K., He, X., & Yao, J. (2026). From Citation Selection to Citation Absorption: A Measurement Framework for Generative Engine Optimisation Across AI Search Platforms. arXiv:2604.25707. https://arxiv.org/abs/2604.25707