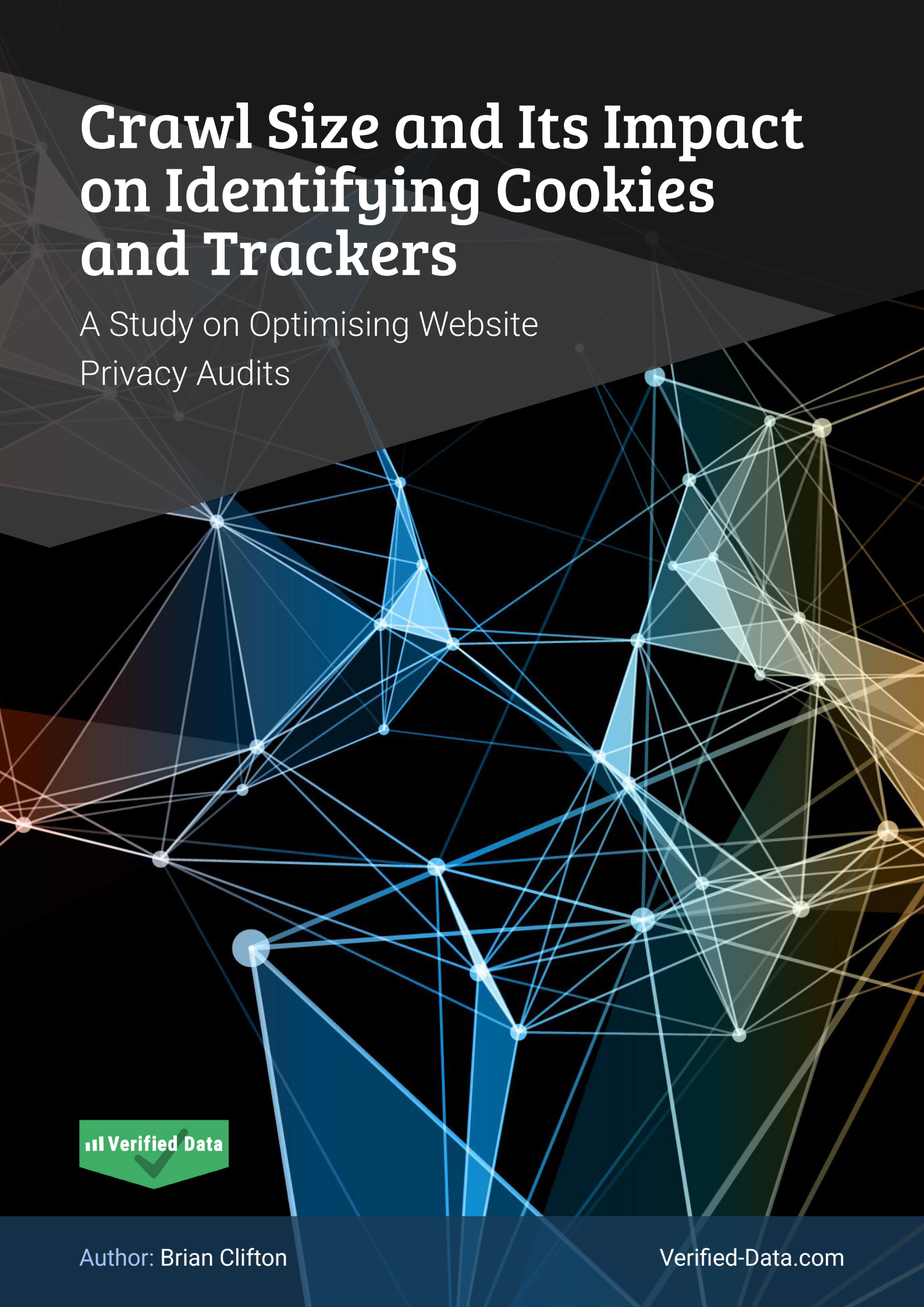
Crawl Size and Its Impact on Identifying Cookies and Trackers

A Study on Optimising Website Privacy Audits



Verified Data

Author: Brian Clifton

Verified-Data.com

Executive Summary

This study demonstrates that, even for complex websites with thousands or millions of pages, more than 90% of cookies and tracking pixels can be identified by auditing just 500 pages. In most cases, auditing 100 pages is sufficient to capture 90% of the tracking inventory.

This finding supports a strategy of conducting regular, small-scale audits, which are faster, easier to manage, and more cost-effective than exhaustive site-wide crawls. Such an approach enables organisations to maintain strong data governance and improve privacy compliance without the operational burden of full-site audits.

01 Introduction

In assessing a website's compliance with data protection and privacy obligations, it is essential to understand the full extent of deployed cookies and tracking pixels — collectively referred to here as the tracking inventory.

A common question among privacy professionals is:

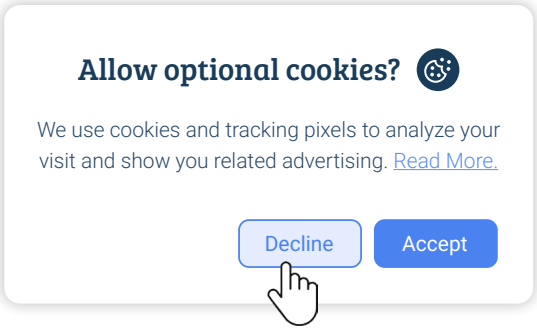
How many pages must be audited to meaningfully assess privacy risks?

Auditing a single page (e.g., the homepage) is insufficient, as different sections of a site often use different templates, content management systems (CMS), and tag management configurations. Additionally, some pixels are smart enough to only fire their network requests after a certain interaction/referrer or cookie has been set. Conversely, crawling every page is infeasible — both due to firewall restrictions and diminishing returns.

This study addresses this gap with empirical evidence.

The objective of this research is to determine an optimal audit size — balancing thoroughness with efficiency — for assessing a site's tracking inventory and privacy compliance.

This study also assumes a "Reject All" consent choice, or the US equivalent of "Do not track", to test whether cookies and trackers respect user preferences.



02 Methodology

We performed controlled audits using the automated PAGE Inspector tool, which replicates real-user behaviour, including interaction with consent banners.

Key parameters:

- Each audit session began with a clean state: all cookies and session data cleared.
- Consent banners were interacted with by selecting “Reject All”.
- Sites were audited over a maximum 7-day window to avoid changes between runs.
- Crawls were capped at 1000 pages per site, but in practice, firewall rules often limited this.
- Navigation employed a random-walk strategy: following internal links, clicking buttons, scrolling, and playing embedded media.

The four websites assessed, representing different business sectors and content sizes, are listed in Table 1.

Figure 1 illustrates the audit data flow. Figure 2 shows example audit settings.

Website	Target audience	Crawl location	Estimated pages
Healthcare site	US	Virginia/US	43,100
Travel site	EU+UK	Ireland/EU	879,000
Technology site	US	Ireland/EU	24,700
Retail site	EU	Ireland/EU	3,460,000

Table 1: The four websites audited in this study. Estimated pages taken from the Google.com index.

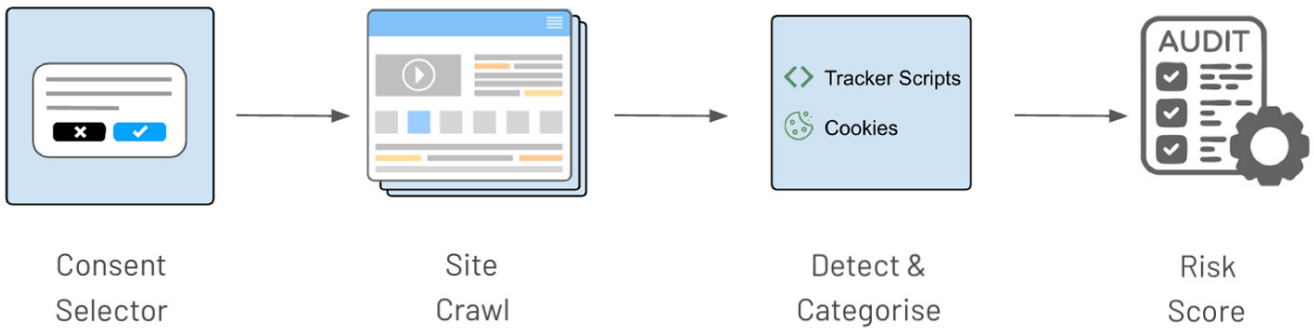


Figure 1: Schematic visitor audit data flow. In this study, the Risk Score is not assessed.

02 Methodology

Figure 2: Example audit settings.

03 Results & Recommendations

The tracking inventory discovered increased as more pages were crawled, but plateaued quickly.

In three of the four sites, more than 90% of trackers were identified after auditing 100 pages.
The remaining site achieved ~80% of inventory coverage at 100 pages, reaching >90% by 500 pages.

This is consistent with the observation that websites reuse a finite set of templates and tag management configurations.

Figures 3–6 illustrate the asymptotic nature of inventory discovery as crawl size increases (the lines of best fit are drawn for guidance only).

Figure 3 Healthcare site (US)

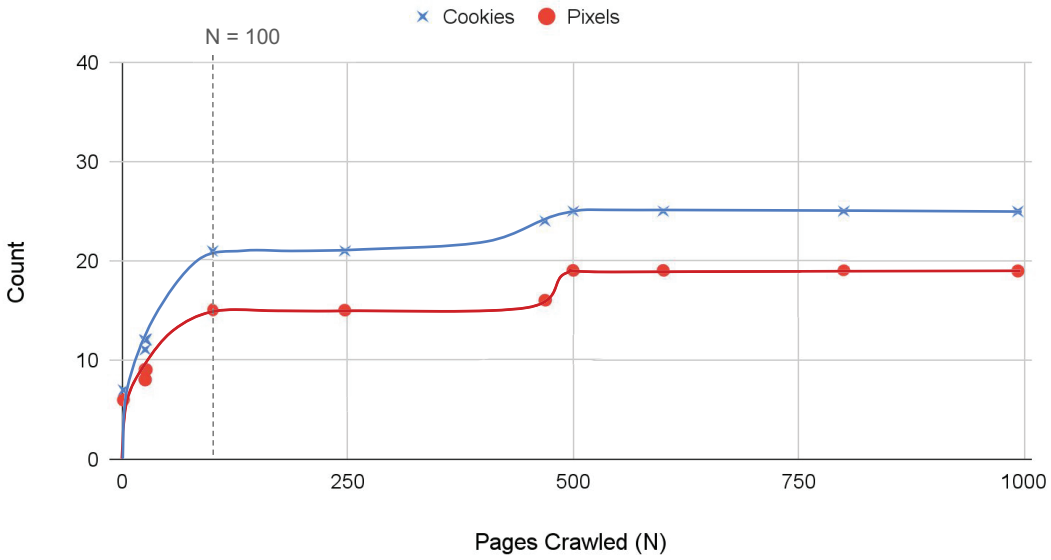
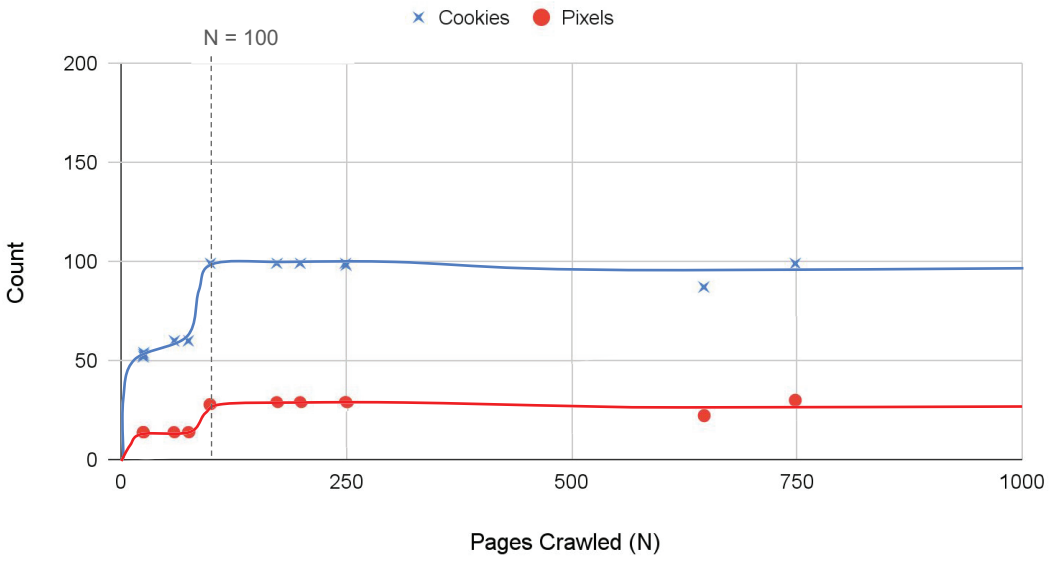


Figure 4 Travel site (EU)



03 Results & Recommendations

Figure 5 Tech site (US)

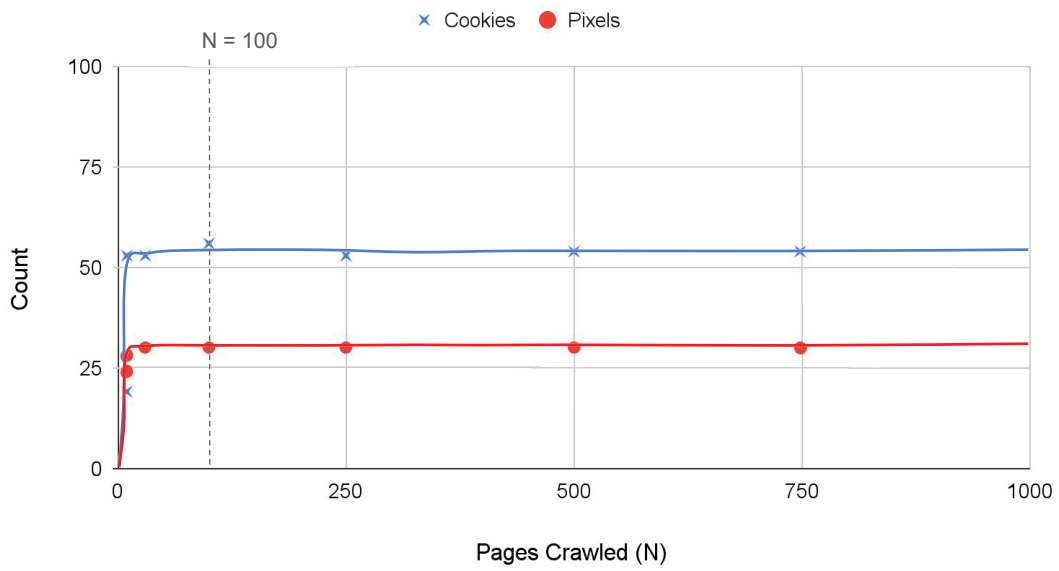
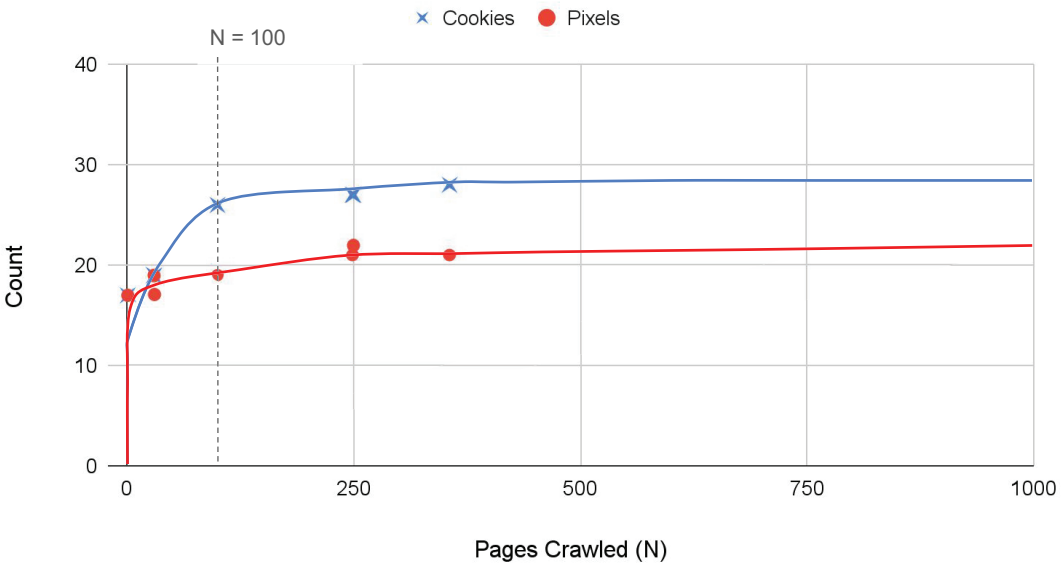


Figure 6 Retail site (EU)



03 Results & Recommendations

Simulated random walks effectively approximate user journeys, uncovering most tracking elements.

Key takeaways:

- For general audits, 25–50 pages can serve as an initial compliance check.
- To achieve ~90% inventory coverage, 100 pages is recommended.
- For high-risk sectors (e.g., healthcare, finance), 500 pages provides greater confidence.

We advise starting with smaller audits, resolving discovered issues, and then scaling up to a more comprehensive inventory as needed.

Limitations

- Automated simulations cannot fully replicate complex user journeys, particularly deep in e-commerce funnels, such as begin checkout.
- Firewall rules may artificially limit crawl depth.
- Certain specific trackers may only appear under specific conditions not fully simulated here. For example, form submission or purchase confirmation.

Nonetheless, evidence suggests that most tracking templates activate at early funnel stages.

04 Conclusion

Auditing all pages (>4.4M in this sample) is impractical and unnecessary. Even at 5 seconds per page to allow all potential tracking elements to load, a full crawl would take over 50 days.

Instead, a pragmatic and scientifically supported approach is to audit 100 pages per site as a standard practice, adjusting upwards for sensitive contexts. This strikes the right balance between thoroughness, cost, and timeliness.

Once a steady-state tracking inventory is established, audits can be run more frequently on smaller samples to monitor compliance over time.

For enquiries,
please contact:

Brian Clifton
hello@verified-data.com

Verified Data

Verified-Data.com