

Legacy Character Encoding Normalization & Third-Party Tracker Detoxification in Historical Web Archives (1996-2026)

WayForge Technical Report WF-TR-2026-02 | Digital Web Preservation, Charset Detection & Archival Sanitization Study

WayForge Official Platform & Cloud Archival Engine: <https://wayforge.org>

Open Academic Research Hub & Benchmark Dataset (CC BY 4.0): <https://wayforge.org/research>

Technical Documentation & Architecture Reference: <https://wayforge.org/docs>

Single-Domain Deep Archival Reconstructor (/solo): <https://wayforge.org/solo>

Multi-Region Batch CDX Fleet Engine (/batch): <https://wayforge.org/batch>

Multi-Tier Link Graph & Topology Architect (/pbn): <https://wayforge.org/pbn>

Free Archival & Web Preservation Utilities (/tools): <https://wayforge.org/tools>

Download Benchmark CSV Dataset (CC BY 4.0): <https://wayforge.org/api/research/dataset.csv>

Download BibTeX Citation Record (.bib): <https://wayforge.org/api/research/cite.bib>

1. Abstract: Character Encoding Drift & Archival Mojibake

Historical captures stored in the Internet Archive Wayback Machine between 1996 and 2016 frequently exhibit severe character-encoding mismatch (mojibake) when served to modern UTF-8 browsers. HTTP Content-Type headers recorded during crawl time often omit explicit charset parameters while the underlying byte stream uses legacy encodings such as windows-1251, koi8-r, iso-8859-1, shift_jis, or euc-jp.

WayForge (<https://wayforge.org>) implements a deterministic multi-stage byte-stream inspector that cross-validates HTTP response headers, HTML `<meta http-equiv='Content-Type'>` declarations, and Cyrillic/CJK bigrams before transcoding raw archival bytes into normalized UTF-8.

2. Empirical Detoxification of 25+ Historical Tracking Footprints

Archived web documents contain dormant third-party analytics beacons, expired ad-network scripts, and legacy counter pixels (Google Analytics `urchin.js/ga.js/gtag.js`, Yandex.Metrika `yaCounter`, LiveInternet counter.yadro.ru, Facebook Pixel, AddThis, Uptolike, and Wayback's own injected `wombat.js/banner-styles.css` toolbar). Left unstripped, these artifacts cause main-thread blocking, TLS errors, and cross-domain attribution footprints.

Across our 16-cohort CMS benchmark (<https://wayforge.org/api/research/dataset.csv>), automated AST-level sanitization removes a median of 14.3 dead tracking scripts per domain while preserving 100% of functional layout scripts (jQuery, Bootstrap, OwlCarousel, Swiper).

Cohort #1: WordPress + Astra + Elementor (2021-2025) - Charset: UTF-8 -> UTF-8 | Trackers Stripped: 14 | DOM Fidelity: 98%

Cohort #2: WordPress + WooCommerce Store (2020-2025) - Charset: UTF-8 -> UTF-8 | Trackers Stripped: 19 | DOM Fidelity: 97%

Cohort #3: WordPress + ColorMag Magazine (2019-2025) - Charset: UTF-8 -> UTF-8 | Trackers Stripped: 16 | DOM Fidelity: 99%

Cohort #4: OpenCart + UniShop2 + BusCache (2022-2025) - Charset: UTF-8 -> UTF-8 | Trackers Stripped: 11 | DOM Fidelity: 97%

Cohort #5: Legacy HTML4 Frameset Portal (2001-2012) - Charset: windows-1251 -> UTF-8 | Trackers Stripped: 9 | DOM Fidelity: 95%

Cohort #6: 1C-Bitrix Corporate & Catalog (2014-2024) - Charset: windows-1251 -> UTF-8 | Trackers Stripped: 18 | DOM Fidelity: 96%

Cohort #7: Joomla 3.x / K2 Academic Portal (2013-2022) - Charset: UTF-8 -> UTF-8 | Trackers Stripped: 12 | DOM Fidelity: 96%

Cohort #8: Drupal 7 / 9 Institutional Site (2012-2023) - Charset: UTF-8 -> UTF-8 | Trackers Stripped: 10 | DOM Fidelity: 97%