Silicon Stratigraphy: A Provenance-First Framework for Preserving Pre-LLM Digital Artifacts in Archaeological and Cultural Heritage Contexts

SCOTT J. BOUDREAUX, Elyan Labs, Louisiana, USA

Digital archaeology now faces a practical preservation problem: how to retain high-confidence human digital records as machine-generated media rapidly expands. This paper proposes a provenance-first framework, Silicon Stratigraphy, for preserving pre-LLM web and software artifacts while documenting post-LLM transformations. The framework combines established digital preservation standards (OAIS, Memento, PROV-O) with cryptographic timestamping and hardware-attested archival workflows. A live implementation is presented as a case study: RustChain Proof of Antiquity nodes and offline legacy compute environments used to capture, hash, timestamp, and replicate artifacts before and after synthetic augmentation. The contribution is methodological rather than universalizing: a reproducible protocol for separating source strata, recording transformation lineage, and reducing evidentiary ambiguity in future scholarship. The framework was developed through hands-on systems operations, repeated validation cycles, and preservation work on mixed legacy and modern infrastructure. The paper closes with validation criteria, governance risks, and recommendations for community-scale deployment in archaeology-adjacent digital heritage work.

Keywords:

digital archaeology, digital preservation, provenance, generative AI, web archives, blockchain timestamping, retrocomputing

SDH Reference:

Boudreaux, Scott J. 2026. "Silicon Stratigraphy: A Provenance-First Framework for Preserving Pre-LLM Digital Artifacts in Archaeological and Cultural Heritage Contexts." Studies in Digital Heritage, submitted manuscript.

https://github.com/Scottcjn/echoes-silicon-age-bridge

1. INTRODUCTION

Archaeology increasingly depends on digital evidence: web pages, repositories, discussion archives, and born-digital field documentation. At the same time, generative systems now produce large volumes of plausible text, images, and code. The resulting challenge is not only long-term storage but evidentiary trust: whether future researchers can distinguish original artifacts from later synthetic revisions, summaries, or reconstructions.

This paper addresses that challenge through a pragmatic question: how can digital heritage practitioners preserve pre-LLM artifacts and document post-LLM interventions without losing chain-of-custody clarity? Rather than treating AI as intrinsically harmful or beneficial, the paper frames it as a provenance pressure that requires better capture and attribution methods at the exact point where data are collected, transformed, and published.

The main contribution is a preservation framework called Silicon Stratigraphy. The framework adapts archaeological stratigraphic logic to digital corpora: (1) isolate temporal layers of artifact production, (2) preserve low-level evidence for each layer, and (3) register transformations with explicit lineage metadata. A field implementation is reported using a live attested ledger (RustChain), web snapshot capture, and offline legacy compute environments.

1. BACKGROUND AND RELATED STANDARDS

The framework builds on existing preservation and provenance standards rather than replacing them. The goal is interoperability with existing archaeological digital practice, not a parallel process.

* 1. OAIS Model

The Open Archival Information System (OAIS) remains the reference architecture for ingest, archival storage, data management, and dissemination in long-term digital preservation (Consultative Committee for Space Data Systems 2012). It establishes functional vocabulary and responsibilities for trustworthy repositories.

* 1. Memento Protocol

RFC 7089 defines datetime content negotiation for web archives, enabling time-based retrieval and citation of prior web states (Van de Sompel et al. 2013). This is central for reconstructing pre-LLM web context.

* 1. PROV-O

W3C PROV-O provides an ontology for describing entities, activities, and agents in provenance graphs (Lebo, Sahoo, and McGuinness 2013). It is suitable for recording synthetic and non-synthetic transformation chains.

* 1. AI Governance Context

UNESCO and NIST frameworks emphasize transparency, accountability, and risk management for AI systems (UNESCO 2021; National Institute of Standards and Technology 2023). In heritage contexts, these principles imply explicit labeling and governance of machine-generated derivatives.

1. SILICON STRATIGRAPHY FRAMEWORK
   1. Core Premise

In stratigraphic archaeology, layer boundaries and disturbance events are key to interpretation. Silicon Stratigraphy adopts the same logic for digital corpora: interpretive confidence depends on preserving layer boundaries and documenting disturbances.

* 1. Layer Taxonomy

The taxonomy is intentionally operational, field-usable, and adjustable by project:

1. Analog Bedrock: paper records, magnetic media, and non-networked digital artifacts.

2. Early Network Layer: BBS/forum/web artifacts with low automation and identifiable human authorship patterns.

3. Pre-LLM Web Layer (baseline in this study: through 2022): high-volume human-authored web and open-source materials before broad public LLM deployment.

4. Synthetic Expansion Layer (2023-present): artifacts produced or transformed with generative systems.

The boundary year (2022/2023) is configurable and should be justified per corpus.

* 1. Preservation Invariants

Each artifact is preserved with five invariants:

1. Byte-level object: original file or capture bundle.

2. Fixity: SHA-256 digest.

3. Time anchor: immutable timestamp entry.

4. Execution context: hardware/software environment metadata.

5. Lineage record: machine-readable transformation graph.

* 1. Pipeline

The capture and verification protocol follows six stages:

Acquire: capture web/resource snapshots with full context metadata (URL, datetime, headers, toolchain).

Fixity: compute SHA-256 and perform size/MIME checks before any derivative generation.

Anchor: write digest plus metadata pointer to an attested ledger event.

Replicate: store in at least two independent repositories (online plus offline).

Transform: for every generated derivative, record model/tool/prompt/version and parent hash.

Audit: run scheduled fixity and retrievability checks with exception logging and remediation notes.

1. CASE STUDY: LIVE IMPLEMENTATION
   1. Infrastructure

A working implementation was evaluated in an operational environment at Elyan Labs. The system includes: (1) a lightweight ledger (RustChain) for timestamp anchoring and audit events; (2) legacy and modern nodes (including PowerPC, POWER8, Apple Silicon, and x86) used for diversity of execution contexts; and (3) offline archival snapshots for selected web corpora. The implementation emphasis is practical reproducibility under mixed hardware constraints, not benchmark optimization.

At the time of observation (February 11, 2026), live node telemetry exposed active miner enrollment and epoch data through public API endpoints. These values are reported here as environment state, not as universal performance claims.

* 1. Why Hardware-Attested Contexts Were Included

Digital forensics and emulation studies show that execution context affects reproducibility and interpretation (Digital Preservation Coalition n.d.). In this implementation, legacy hardware was used for two reasons: (1) to document historically plausible runtime constraints for replay and emulation; and (2) to diversify provenance evidence (clock sources, architecture, toolchain behavior) in signed archival events. The method does not claim that old hardware is inherently more truthful. The claim is narrower: explicit environment diversity improves auditability when recorded correctly and repeatedly validated.

* 1. Applied Example Workflow

A representative workflow for a pre-LLM webpage proceeds as follows:

1. Retrieve a dated snapshot (or capture a new one with full headers and timestamp).

2. Store capture bundle and compute SHA-256 digest.

3. Publish digest and metadata pointer to the attested ledger.

4. Generate derivative summaries or modernized renderings as separate artifacts.

5. Link each derivative to its parent hash with transformation metadata.

This keeps interpretive products useful while preventing them from silently replacing source evidence. In this workflow, derivative outputs are never treated as substitutes for source-layer records.

Figure intentionally omitted in this submission file pending explicit image-verification confirmation during editorial upload.

1. EVALUATION CRITERIA

This work is a methods paper; evaluation is therefore procedural. A deployment is considered successful when it meets the following criteria and allows an external team to reconstruct the evidence path end-to-end:

1. Recoverability: independent parties can retrieve the preserved object from at least one replica.

2. Fixity integrity: periodic hash checks produce no unexplained drift.

3. Lineage completeness: each derivative has explicit parent links and transformation metadata.

4. Temporal auditability: timestamp anchors and capture datetimes are consistent and externally inspectable.

5. Disclosure quality: interfaces clearly distinguish source artifacts from generated derivatives.

Future work should benchmark this framework against institutional repositories with controlled inter-rater studies on evidentiary confidence and blinded replication tasks.

1. LIMITATIONS

The current implementation has several limitations:

1. Single-organization deployment bias: field observations are from one operator context.

2. No adversarial red-team trial in this paper: tampering and replay resistance require separate formal testing.

3. Boundary ambiguity: some 2022-2024 artifacts are hybrid human/machine products, making strict layer assignment difficult.

4. Governance overhead: detailed provenance capture increases operational cost and may reduce adoption without tooling support.

5. Case-study depth: this manuscript focuses on method and implementation; larger multi-site comparative studies remain future work.

1. GOVERNANCE AND ETHICS

Preservation systems can reproduce power asymmetries if access and authorship controls are opaque. Three governance rules are recommended: (1) publish provenance schemas and audit procedures openly; (2) require explicit labeling for generated derivatives in public interfaces; and (3) support community co-curation to reduce unilateral control of archival narratives. These rules align with AI ethics guidance emphasizing transparency, accountability, and contestability (UNESCO 2021; National Institute of Standards and Technology 2023).

1. CONCLUSION

The core risk in post-LLM digital archaeology is not generation itself but undocumented transformation. Silicon Stratigraphy offers a practical response: preserve source layers, anchor fixity, and make derivative lineage explicit. The case study demonstrates that a small organization can implement this approach with existing standards and modest infrastructure and constrained operational conditions.

For archaeology and digital heritage communities, the immediate priority is methodological convergence: interoperable provenance records, shared audit practices, and clear public labeling. If these controls are adopted early, synthetic tools can enrich interpretation without eroding the evidentiary substrate that future scholarship depends on.

1. COMPETING INTERESTS

The author leads projects discussed in the case study (RustChain and related infrastructure). This paper is presented as a methods and implementation note; readers should interpret platform-specific observations accordingly.

1. DATA AND MATERIALS AVAILABILITY

Example implementation artifacts are publicly visible in project repositories and live service endpoints at the time of writing. For archival integrity, timestamps and values should be re-queried by reviewers at evaluation time.

1. REFERENCES

Consultative Committee for Space Data Systems. 2012. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-M-2. https://public.ccsds.org/pubs/650x0m2.pdf.

Digital Preservation Coalition. n.d. Digital Preservation Handbook. Accessed February 11, 2026. https://www.dpconline.org/handbook.

Kansa, Eric C. 2012. "Openness and Archaeology's Information Ecosystem." World Archaeology 44 (4): 498-520. https://doi.org/10.1080/00438243.2012.737575.

Lebo, Timothy, Satya Sahoo, and Deborah McGuinness, eds. 2013. PROV-O: The PROV Ontology. W3C Recommendation. https://www.w3.org/TR/prov-o/.

Marwick, Ben. 2017. "Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation." Journal of Archaeological Method and Theory 24: 424-450. https://doi.org/10.1007/s10816-015-9272-9.

National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1.

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Van de Sompel, Herbert, Michael L. Nelson, Robert Sanderson, Lyudmila Balakireva, Scott Ainsworth, and Harihar Shankar. 2013. RFC 7089: HTTP Framework for Time-Based Access to Resource States (Memento). IETF. https://www.rfc-editor.org/rfc/rfc7089.