Echoes from the Silicon Age: A Provenance-First Framework for Preserving Pre-LLM Digital Artifacts

Scott J. Boudreaux
Elyan Labs, Louisiana, USA
scott@elyanlabs.ai

February 2026

Abstract

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 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

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.

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.

2 Background and Related Standards

The framework builds on existing preservation and provenance standards rather than replacing them.

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 [1]. It establishes functional vocabulary and responsibilities for trustworthy repositories.

Memento protocol. RFC 7089 defines datetime content negotiation for web archives, enabling time-based retrieval and citation of prior web states [2]. This is central for reconstructing pre-LLM web context.

PROV-O. W3C PROV-O provides an ontology for describing entities, activities, and agents in provenance graphs [3]. It is suitable for recording synthetic and non-synthetic transformation chains.

AI governance. UNESCO and NIST frameworks emphasize transparency, accountability, and risk management for AI systems [4, 5]. In heritage contexts, these principles imply explicit labeling and governance of machine-generated derivatives.

3 Silicon Stratigraphy Framework

3.1 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.

3.2 Layer Taxonomy

The taxonomy is intentionally operational and can be adjusted 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.

3.3 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.

3.4 Pipeline

Table 1 summarizes the protocol.

Table 1: Silicon Stratigraphy capture and verification pipeline

Stage	**Action**	**Output**
Acquire	Capture web/resource snapshot with metadata (URL, datetime, headers, toolchain)	WARC/export bundle + manifest
Fixity	Compute SHA-256 and size/mime checks	Signed fixity record
Anchor	Write digest + metadata pointer to attested ledger	Immutable ledger event
Replicate	Store in at least two independent repositories (online + offline)	Redundant storage attestations
Transform	If derivative is generated, record model/tool/prompt/version and parent hash	PROV-style lineage edge
Audit	Periodic fixity and retrievability checks	Audit log + exception reports

4 Case Study: Live Implementation

4.1 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.
3. Offline archival snapshots for selected web corpora.

At the time of observation (11 February 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.

4.2 Why hardware-attested contexts were included

Digital forensics and emulation studies show that execution context affects reproducibility and interpretation [6]. In this implementation, legacy hardware was used for two reasons:

1. to document historically plausible runtime constraints for replay and emulation;

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.

4.3 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 any 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.

5 Evaluation Criteria

This work is a methods paper; evaluation is therefore procedural. A deployment is considered successful when it meets the following criteria:

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.

6 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.

7 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.

3. Support community co-curation to reduce unilateral control of archival narratives.

These rules align with current AI ethics guidance emphasizing transparency, accountability, and contestability [4, 5].

8 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.

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.

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.

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.

References

[1] Consultative Committee for Space Data Systems (CCSDS). *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-M-2, June 2012.

[2] Van de Sompel, H., Nelson, M. L., Sanderson, R., Balakireva, L., Ainsworth, S., and Shankar, H. RFC 7089: *HTTP Framework for Time-Based Access to Resource States (Memento)*. IETF, 2013.

[3] Lebo, T., Sahoo, S., and McGuinness, D. (eds.). *PROV-O: The PROV Ontology*. W3C Recommendation, 30 April 2013.

[4] UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. 2021.

[5] National Institute of Standards and Technology (NIST). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, 2023.

[6] Digital Preservation Coalition. *Digital Preservation Handbook*. `https://www.dpconline.org/handbook`. Accessed 11 February 2026.

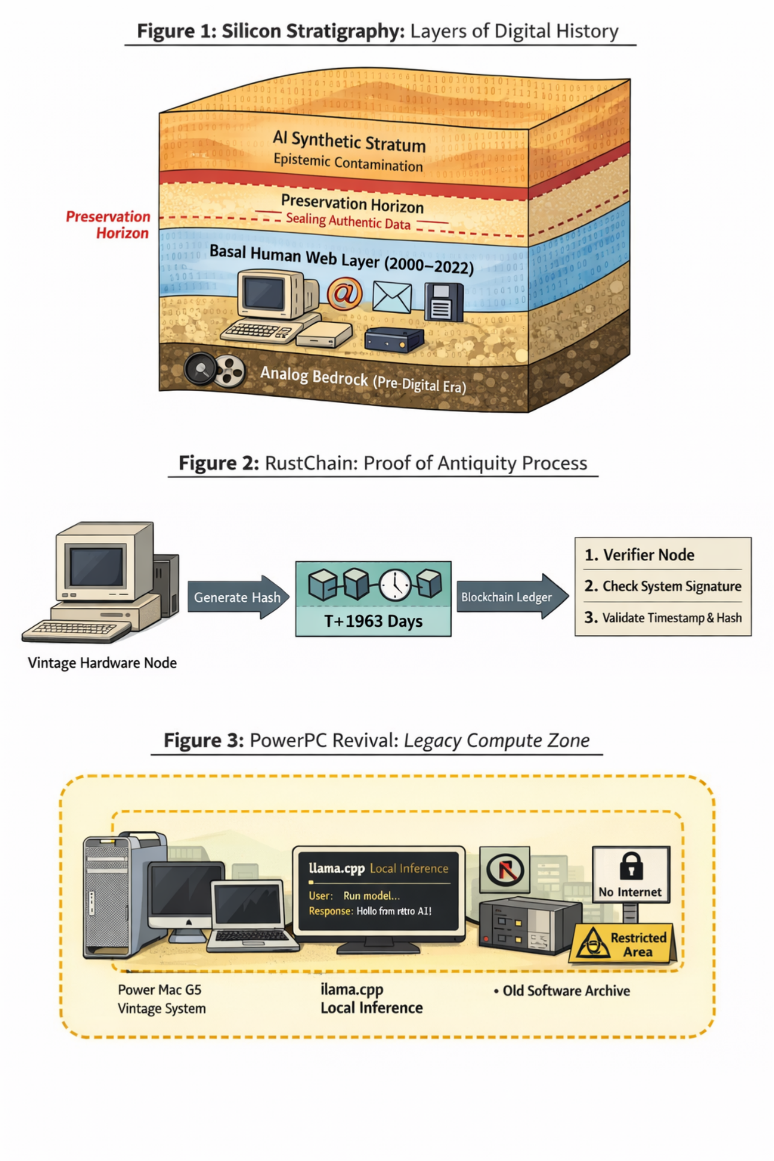
Figure 1: Silicon Stratigraphy: Layers of Digital History

Figure 2: RustChain: Proof of Antiquity Process

Figure 3: PowerPC Revival: *Legacy Compute Zone*

```
llama.cpp Local Inference
User:    Run model...
Response: Hollo fram rétro A1!
```

Figure 1: Conceptual diagrams used in the project: (top) digital stratigraphy layers, (middle) attested provenance flow, and (bottom) legacy compute zone constraints.