Silicon Stratigraphy for Archaeological Evidence Integrity: A Provenance-First Framework for AI-Mediated Digital Workflows

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

February 2026

Abstract

Archaeological research increasingly depends on digital records that are transformed through AI-assisted workflows. This creates a direct interpretive risk: derivative outputs can circulate without enough provenance to verify what evidence they came from or how they were altered. This paper presents a full archaeology-focused formulation of *Silicon Stratigraphy*, a provenance-first framework for preserving evidentiary integrity across excavation records, museum documentation, and field-to-public digital pipelines. The framework combines OAIS-aligned preservation practice, Memento time-state capture, PROV-O lineage modeling, and cryptographic fixity and timestamp anchoring. The contribution is practical and methodological: an implementable workflow for distinguishing source evidence from AI-mediated derivatives, plus archaeology-specific evaluation criteria for interpretive fidelity and reproducibility. The paper provides a case-study structure for immediate use in resubmission and peer review, with explicit reporting fields for dataset scope, transformation provenance, and decision-level interpretive impact.

Keywords: digital archaeology; archaeological method; provenance; evidentiary integrity; reproducibility; generative AI; digital heritage

1 Introduction

Archaeologists now work in digital environments where core evidence is often accessed, summarized, translated, and re-presented through software pipelines. Excavation logs, artifact catalogues, archival imagery, geospatial exports, and project databases are routinely transformed before interpretation and publication. Recent AI tooling accelerates this process, but it also increases the chance that derivative outputs become detached from the source evidence that originally supported archaeological claims.

The core concern is methodological rather than rhetorical: if source and derivative layers are not explicitly separated, provenance failure can silently alter interpretation. In physical archaeology, context loss is a foundational warning. In digital archaeology, undocumented transformation is the parallel risk.

This paper addresses that risk directly within standard archaeological research settings. It reframes provenance-first preservation not as an abstract digital-heritage objective but as a practical requirement for archaeological interpretation under AI-mediated workflows.

2 Archaeological Research Questions

The manuscript addresses three archaeology-led research questions:

1. **RQ1**: How does provenance failure in AI-mediated digital toolchains skew interpretation of archaeological evidence (for example excavation documentation, artifact catalogues, and archival images)?

2. **RQ2**: What minimum controls should archaeologists implement so AI-assisted transformations remain auditable and source-distinguishable?

3. **RQ3**: Can a provenance-first workflow preserve practical AI utility while improving interpretive fidelity and reproducibility in archaeological reporting?

The contribution is a full workflow and evaluation model that can be used by archaeological projects, museum teams, and digital research groups without prohibiting AI tools.

3 Background and Standards

The framework builds on established standards and reproducibility literature:

1. **OAIS** defines ingest, archival storage, data management, and dissemination responsibilities for long-term digital preservation [1].

2. **Memento (RFC 7089)** supports datetime-based retrieval of prior web resource states, which is critical for time-specific archaeological citation [2].

3. **PROV-O** provides a machine-readable model for entities, activities, and agents in transformation lineages [3].

4. **Archaeological open-data and reproducibility work** emphasizes inspectable workflows and rerunnable evidence paths [4, 5].

5. **AI governance guidance** emphasizes transparency, accountability, and risk management in automated systems [6, 7].

Silicon Stratigraphy operationalizes these components in archaeology-specific workflows.

4 Archaeological Case-Study Data Design

To match archaeology-focused scope, this manuscript defines and executes three case-study streams aligned with common archaeological practice. All values below are derived from reproducible retrieval and processing scripts run on 19 February 2026 (UTC).

4.1 Case Stream A: Excavation Logs and Artifact Catalogues

Data class: digitized trench notebooks, context sheets, locus descriptions, finds registers, and catalogue exports.
Primary risk: AI summarization or normalization can flatten stratigraphic qualifiers, uncertainty language, and local terminology.
Required reporting fields:

- Project/site identifier: New York City Landmarks Preservation Commission Archaeology Reports Database (Socrata dataset ID `fuzb-9jre`, source endpoint `https://data.cityofnewyork.us/resource/fuzb-9jre.json`).

- Date range of sampled source records: 1973–2015 (from the *date* field in the extended run).

- Record count and file types: 300 source records, JSON format; normalized source fields were `biblioid`, `borough`, `author`, `date`, `title`, and `report_abstract`.

- Transformation tools used: deterministic baseline and provenance transforms implemented in `run_provenance_extended_validation.py` (labels: `summary` baseline and `deterministic_transform_v2` provenance mode).

4.2 Case Stream B: Museum and Collection Documentation

Data class: accession records, object metadata, conservation notes, and archival photographs.
Primary risk: AI enhancement or metadata generation can introduce stylistic certainty or object traits not present in primary records.
Required reporting fields:

- Institution/collection: Cleveland Museum of Art Open Access API (`https://openaccess-api.clevelandart.org/api/artworks/`).
- Object classes sampled: 196 deduplicated records queried from ancient-relevant terms (`roman`, `egyptian`); sampled departments included Greek and Roman Art and Egyptian and Ancient Near Eastern Art, with additional records returned by API indexing.
- Image/metadata transformations audited: source metadata to summary-only derivative (naive mode) versus provenance-first derivative preserving mandatory context keys (`objectID`, `department`, `title`, `culture`, `objectDate`, `accessionNumber`).
- Validation protocol in this run: machine-auditable metadata retention and lineage checks; no institution-side curatorial adjudication was claimed for this experiment.

4.3 Case Stream C: Contemporary Field or Public Portal Pipeline

Data class: public-facing museum portal/API records and AI-assisted labels/summaries.
Primary risk: generated explanatory layers can be mistaken for field observations in downstream interpretation.
Required reporting fields:

- Field project or portal: Art Institute of Chicago public API search endpoint (`https://api.artic.edu/api/v1/artworks/search`).
- Data products evaluated: 191 deduplicated records from roman/egyptian queries using fields `id`, `title`, `date_display`, `place_of_origin`, and `main_reference_number`.
- AI-mediated layers generated in this experiment: deterministic derivative summaries for each sampled record to emulate accelerated interpretation outputs.
- Disclosure and audit interface: explicit derivative labeling (`generated_label`), parent linkage (`parent_sha256`), and run-level reports in `extended_provenance_validation_report.md`.

4.4 Data Governance Across All Case Streams

All case streams should document:

1. legal and ethical basis for data handling,
2. provenance retention policy,
3. role-level permissions for source versus derivative editing,
4. public disclosure language distinguishing evidence from generated interpretation.

5 Silicon Stratigraphy Framework

5.1 Layer Model

The method adapts archaeological stratigraphic logic to digital evidence:

1. **Source Layer**: primary archaeological records as originally captured or digitized.
2. **Preservation Layer**: fixity, timestamp, and replication artifacts that stabilize source state.
3. **Derivative Layer**: AI-assisted outputs (summaries, classifications, translations, renderings).
4. **Interpretive Layer**: published claims, arguments, and narratives that must reference both source and derivative lineage.

5.2 Preservation Invariants

Each evidence object requires five invariants:

1. byte-level artifact package,
2. SHA-256 digest and file metadata,
3. trusted timestamp anchor,
4. execution-context metadata (software/toolchain, model/version where relevant),
5. lineage edges linking parent and derivative objects.

Table 1: Silicon Stratigraphy operational pipeline for archaeological records

Stage	Action	Output
Acquire	Capture source objects with context metadata (origin, datetime, operator, format, project IDs)	Source package + manifest
Fixity	Compute digest and file-level checks	Fixity register
Anchor	Record digest commitments to immutable timestamped log	Anchor event record
Replicate	Store in independent repositories (for example institutional + offline copy)	Redundancy attestations
Transform	Record AI or software transformation (tool, version, parameters, prompt when relevant)	Lineage event
Audit	Re-run fixity and lineage completeness checks on schedule	Audit report + exceptions

5.3 Operational Pipeline

6 Archaeology-Specific Outcomes and Evaluation Criteria

Evaluation is tied to archaeological interpretation, not generic system throughput.

Table 2: Evaluation criteria for archaeological interpretive fidelity

Criterion	How to Measure	Interpretive Relevance
Source-distinguishability	Share of derivative outputs with explicit parent pointers to source records	Prevents derivatives from being cited as primary evidence
Context retention	Presence of key contextual markers (context IDs, provenience fields, uncertainty qualifiers) before and after transformation	Protects stratigraphic and contextual meaning

Criterion	How to Measure	Interpretive Relevance
Lineage completeness	Proportion of transformation steps recorded with tool/version metadata	Enables third-party audit and rerun
Fixity stability	Scheduled re-hash checks with no unexplained drift	Supports integrity and chain-of-custody confidence
Disclosure clarity	Human-review scoring of whether interfaces clearly label generated content	Reduces interpretive confusion in publication and public dissemination
Reproducibility	Independent team ability to reconstruct claim-to-source pathway	Tests archaeological argument robustness

Projects should define pass thresholds per criterion before deployment and report exceptions explicitly.

7 Experiment Execution and Results

Three streams were executed with live retrieval and deterministic reproducibility controls. Full machine-readable outputs are archived in:

- `/home/scott/jcaa_experiments_2026-02-19/results/extended_provenance_validation_resu`
- `/home/scott/jcaa_experiments_2026-02-19/results/extended_provenance_validation_repo`

The total evaluated corpus in the extended run was 687 records (300 NYC archaeology-report records, 196 Cleveland Museum records, 191 Art Institute of Chicago records). Table 3 summarizes key measured outcomes comparing a naive derivative mode against a provenance-first mode.

Across all three streams in the extended run, provenance-first mode produced:

1. source-distinguishability rate = 1.000,
2. lineage completeness rate = 1.000,
3. disclosure label rate = 1.000,
4. reproducible root-hash verification = true in all streams.

In contrast, the naive mode produced zero explicit parent linkage and zero lineage completeness in this run, demonstrating the specific interpretive risk addressed by the framework.

Table 3: Measured outcomes from the 2026-02-19 extended validation run

Stream	Records	Fixity Stability	Context Retention (Prov)	Lineage Completeness (Prov)	Tamper Recall
NYC archaeology reports	300	1.000	0.999	1.000	1.000
Cleveland museum records	196	1.000	1.000	1.000	1.000
AIC museum records	191	1.000	1.000	1.000	1.000

7.1 Real-LLM Single-Hop and Multi-Hop Validation

To address reviewer concern that deterministic transforms may not represent live model behavior, a separate run executed real LLM transforms using a local Ollama endpoint (`http://127.0.0.1:11434/api/generate`) and model `qwen2.5-coder:1.5b` with CPU-forced options (`num_gpu=0`). This run used 12 sampled records per stream (36 total) across the same three archaeology-relevant source snapshots, with 72 total LLM calls.

One-hop mode generated an LLM summary directly from source records. Two-hop mode generated a second LLM interpretation from hop-1 summaries. Naive variants omitted lineage metadata; provenance-first variants attached `generated_label`, `generated_by`, `generated_at`, and parent linkage plus mandatory context fields.

Table 4: Real-LLM one-hop and two-hop lineage outcomes (2026-02-19 run)

Stream	Records	1-Hop Naive Lineage	1-Hop Prov Audit Pass	2-Hop Naive Lineage	2-Hop Prov Chain Pass
NYC archaeology reports	12	0.000	1.000	0.000	1.000
Cleveland museum records	12	0.000	1.000	0.000	1.000
AIC museum records	12	0.000	1.000	0.000	1.000

Additional two-hop fault injections (20% records per stream) tested orphan parent links, ancestor mismatch, mandatory context erasure, and generated-label stripping. Baseline two-hop provenance had fail rate 0.000; injected scenarios had fail rate 0.167 with detection recall 1.000 and false positive rate 0.000 in all streams.

8 Audit Failure Proof-of-Concept

To demonstrate that the audit layer fails records when provenance constraints are broken, four additional fault scenarios were injected into provenance-mode derivatives (10% of records per stream): orphan parent link, missing `generated_by`, mandatory context erasure, and stripped `generated_label`.

Table 5: Injected audit-failure scenarios and observed audit outcomes

Scenario	**Injection Fraction**	**Observed Fail Rate Range**	**Recall Range**	**FP Range**
Baseline provenance (no injection)	0.00	0.000–0.000	n/a	0.000–0.000
Orphan parent link	0.10	0.097–0.100	1.000–1.000	0.000–0.000
Missing `generated_by`	0.10	0.097–0.100	1.000–1.000	0.000–0.000
Mandatory context erasure	0.10	0.097–0.100	1.000–1.000	0.000–0.000
Generated label stripped	0.10	0.097–0.100	1.000–1.000	0.000–0.000

These results provide explicit fail-state evidence: the audit controls do not merely score records; they produce deterministic failure on broken lineage, broken disclosure, and broken context.

9 Independent Rerun Protocol

The validation can be independently rerun with the following protocol:

1. Ensure `python3` and the `requests` package are available.
2. Execute:

   ```
   python3 /home/scott/jcaa_experiments_2026-02-19/scripts/run_provenance_extended_val
   ```

3. Verify output files:

   ```
   /home/scott/jcaa_experiments_2026-02-19/results/extended_provenance_validation_resu
   /home/scott/jcaa_experiments_2026-02-19/results/extended_provenance_validation_repo
   ```

4. Confirm key expectations in the report:

```
- baseline_provenance fail rate = 0.000
- injected scenarios fail rate ~= 0.10
- detection recall = 1.000
- false positive rate = 0.000
```

10 Case-Study Reporting Protocol

For each case stream (A-C), report the following:

1. dataset scope and inclusion rules,
2. baseline source inventory statistics,
3. transformation inventory (manual and AI-assisted),
4. criterion-level outcomes from Table 2,
5. interpretive deviations detected and how they were resolved,
6. remaining uncertainty and unresolved provenance gaps.

This structure keeps results archaeologically interpretable and comparable across projects.

11 Pilot Implementation Note

A working implementation accompanies this manuscript as a proof of deployability. The implementation demonstrates:

1. packaging of source and derivative artifacts with manifests,
2. hash and timestamp anchoring workflow,
3. lineage recording across transformed outputs,
4. audit-ready artifact bundles for external verification.

The implementation is presented as procedural evidence that the framework is practical. It is not presented as a universal benchmark study.

12 Discussion

The findings of this methods study support three practical points for archaeology:

1. **AI use can remain methodologically acceptable** when provenance boundaries are explicit and auditable.
2. **Interpretive reliability depends on provenance discipline**, not on whether teams use or avoid AI tools.
3. **Workflow transparency is a publication issue**, not only a technical one: provenance should be reviewable alongside argumentation.

For journal practice, this implies that manuscripts relying on AI-mediated processing should provide source/derivative traceability at submission stage.

13 Limitations

This manuscript is a framework and reporting design paper with a pilot implementation. Limits include:

1. no multi-lab controlled replication trial in this version,
2. project-specific differences in archive structure and legal constraints,
3. transition-period records where human and machine authorship are partially entangled,
4. the real-LLM validation sample is limited (36 records; 72 calls) and should be expanded in future rounds,
5. additional labor overhead until provenance tooling is integrated into standard archaeological software stacks.

14 Conclusion

Archaeological interpretation in digital environments requires a direct response to provenance failure under AI-mediated workflows. Silicon Stratigraphy provides that response through explicit source-preservation boundaries, auditable transformation lineage, and archaeology-specific evaluation criteria tied to interpretive fidelity and reproducibility.

The framework is intended for immediate application in excavation archives, museum documentation, and field-to-public digital pipelines. Its value is pragmatic: keep modern computational workflows usable while preserving the evidentiary discipline archaeological interpretation depends on.

Competing Interests

The author leads some implementation infrastructure discussed as procedural demonstration. This manuscript is submitted as a methods contribution focused on archaeological workflow integrity.

Data and Materials Availability

The implementation package includes manuscript artifacts, manifests, and hash records intended for independent verification. This revision also includes reproducible experiment scripts and outputs in `/home/scott/jcaa_experiments_2026-02-19/`. Source endpoints used in the run were:

1. NYC Archaeology Reports endpoint: `https://data.cityofnewyork.us/resource/fuzb-9jre.json`
2. Cleveland Museum of Art Open Access endpoint: `https://openaccess-api.clevelandart.org/api/artworks/`
3. Art Institute of Chicago API endpoint: `https://api.artic.edu/api/v1/artworks/search`
4. Real-LLM validation outputs: `/home/scott/jcaa_experiments_2026-02-19/results/llm_proven` and `/home/scott/jcaa_experiments_2026-02-19/results/llm_provenance_validation_report`

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] Kansa, Eric C. "Openness and Archaeology's Information Ecosystem." *World Archaeology* 44, no. 4 (2012): 498–520.

[5] Marwick, Ben. "Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation." *Journal of Archaeological Method and Theory* 24 (2017): 424–450.

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

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