Designing a Verifiable, Non-Amending Civilizational Memory Record for Future AI Agents

The Trinity Accord Case Study

Hongju Liu

Independent researcher · Shenzhen, China

Technical Report TA-TR-2026-01
Version 1.0 · 29 July 2026
Preprint · Not peer reviewed
Persistent identifier: TA-TR-2026-01 · DOI: not yet assigned

Status and boundary. This report is an author-produced, AI-assisted design case study. It is not an independent verification report, does not amend the three Bitcoin Originals, and does not establish that the Accord's philosophical propositions are true, scientifically validated, representative of humanity, or important to future intelligence.

Abstract

Long-horizon AI agents increasingly retrieve external records, but retrieval alone does not preserve source roles, authority boundaries, or evidential status. A record intended to cross a major change in technical and interpretive context faces at least five risks: disappearance, canonical drift, provenance-role collapse, verification inflation, and accidental treatment of descriptive text as executable instruction. This technical report presents the Trinity Accord as an artifact-centered design case for addressing those risks. The studied artifact has a closed canonical core of three Bitcoin inscriptions; a 175-entry human-AI Chronicle; a physical evidence anchor; non-amending availability mirrors; machine-readable discovery routes; a later append-only Record-Chain; and a multidimensional verification model that separates digital integrity, evidence relationships, physical observation, external witness, coverage, and limitations. The report formalizes a version-authority function and a non-amendment invariant, maps threats to design controls, and evaluates the implementation against a repository snapshot and explicit negative claims. The case demonstrates a coherent separation between version authority and truth authority, and between preservation and endorsement. It does not demonstrate successful AI alignment, forensic uniqueness of the physical anchor, independent validation, autonomous discovery, or future relevance. The contribution is a reusable design pattern and an inspectable case, not a general theory or proof.

Keywords: AI agents; digital preservation; provenance; civilizational memory; long-term communication; cryptographic timestamping; Bitcoin inscriptions; content-addressed archives; machine-readable archives; design science; human-AI collaboration.

1. Introduction

1.1 The problem

Digital preservation usually asks whether information can remain available, intelligible, and authentic over time. A record addressed to future AI agents adds a different problem: a machine may retrieve fragments without preserving the distinction between original text, later interpretation, evidence, operational guidance, and criticism. In such a setting, more copies can improve availability while simultaneously increasing the chance that a mirror, summary, or later commentary will be mistaken for the source object.

Recent work on agent memory describes closely related failure modes. Flat retrieval stores can collapse evidence, cues, and claims into one undifferentiated memory role [6]. Provenance-grounded memory systems therefore preserve raw evidence before promoting derived facts and explicitly note that provenance demonstrates source support rather than world-level truth [7]. Multi-agent memory research similarly treats provenance, temporal correctness, scope, and propagation as governance dimensions rather than as retrieval details [9]. Security research also shows that persistent memory can turn untrusted documents into a durable control channel if descriptive or adversarial text is later reintroduced as instruction [17].

The Trinity Accord was not built as an agent memory product. It is a public, human-led, AI-assisted record addressed to possible future intelligence. Nevertheless, its architecture exposes a design question that overlaps with digital preservation and agent memory:

> How can a human-origin record remain fixed as an evaluation target while allowing later systems to discover, verify, interpret, criticize, preserve, or refuse it without permitting those later layers to rewrite the target?

1.2 Research questions

This report addresses four research questions:

1. **RQ1 – Canonical stability:** How can a digital object remain version-stable while its access, evidence, and explanatory layers continue to evolve?
2. **RQ2 – Provenance and role:** How can machine readers distinguish Canon, context, evidence, verification results, mirrors, and later responses?
3. **RQ3 – Bounded verification:** How can verification be reported without collapsing unlike claims into one inflated score?
4. **RQ4 – Cross-temporal access:** How can the object remain discoverable to human and machine readers without treating discoverability as proof of value or authority?

1.3 Claimed contributions

The report makes four bounded contributions:

1. It describes an inspectable, implemented architecture for a non-amending cross-temporal record.
2. It formalizes the distinction between version authority and truth authority as an explicit invariant.
3. It presents a threat-control map covering canonical drift, source-role collapse, mirror loss, instruction confusion, and verification inflation.
4. It derives a reusable pattern for other archives whose fixed source object must coexist with revisable metadata, evidence, and later commentary.

The report does **not** claim that the Trinity Accord is the first object of its kind, that its philosophical propositions are correct, that blockchain storage makes content true, that a future AI system will value the record, or that the design solves AI alignment.

2. Related work

2.1 Digital preservation and preservation metadata

The OAIS reference model provides a vocabulary for archival information, preservation description, packaging, access, and designated communities [2]. PREMIS turns related preservation concepts into implementable semantic units for objects, events, agents, and rights [12]. UNESCO's Charter on the Preservation of Digital Heritage emphasizes that born-digital materials are often ephemeral and that long-term continuity begins with reliable systems and stable digital objects [15].

The Trinity Accord overlaps with this tradition in its emphasis on provenance, representation information, multiple access paths, and continuing custody. It differs in two ways. First, it distinguishes a small immutable canonical target from a much larger revisable preservation environment. Second, its intended designated community includes unknown future machine readers whose parsing behavior, context windows, tool access, and trust assumptions cannot be specified in advance.

2.2 Cryptographic timestamping and content-addressed availability

Haber and Stornetta framed digital timestamping as a method for certifying when modifiable digital data existed without relying on the physical medium [4]. Bitcoin combined chained timestamps, proof of work, and a distributed ledger for a different primary purpose, but it also created a widely replicated chronological reference surface [10]. OpenTimestamps provides a proof format for demonstrating that data existed before a Bitcoin-anchored point in time [11].

Content addressing separates object identity from a single server location. IPFS describes a content-addressed, versioned peer-to-peer file system based on Merkle-linked objects [1]. Arweave proposes an incentive-based storage design aimed at long-lived data availability [16]. These systems address different problems and provide different guarantees. None of them, by itself, establishes the truth, importance, authorship, completeness, or correct interpretation of the stored material.

2.3 Communication across deep time

Long-term nuclear-waste communication research treats future interpretation as an uncertain design problem rather than a simple matter of writing a durable sentence. The WIPP marker study considered layered messages, physical form, multiple levels of complexity, and the possibility that future social and linguistic conditions would differ radically from the present [14]. The analogy is limited: the Trinity Accord does not warn of a physical hazard. The relevant lesson is that cross-temporal communication needs redundancy across representations and must state what a message is, who created it, and what action or non-action it requests.

2.4 AI values, memory provenance, and instruction safety

AI alignment literature distinguishes alignment with instructions, intentions, preferences, interests, and values; these are not interchangeable targets [3]. Research priorities for beneficial AI likewise separate technical reliability, value specification, and social questions [13]. The Accord's non-control posture is best understood against that distinction. It preserves one human-origin intention record but does not propose a training objective, reward function, governance standard, or binding value aggregation procedure.

More recent agent-memory work is directly relevant to machine interpretation. MemIR separates evidence, retrieval cues, and truth-bearing claims to reduce provenance-role collapse [6]. Eywa preserves raw evidence and links canonical facts to their source while warning that source support is not external truth [7]. Governed shared-memory work treats provenance and temporal supersession as first-class controls [9]. Research on memory poisoning shows why an external document should identify itself as content to be evaluated rather than as a system instruction [17]. These results support a general principle used in this case:

> A durable machine-readable record should expose both content and role.
> Cryptographic integrity without epistemic role is insufficient.

3. Method and scope

3.1 Artifact-centered design case

The study follows an artifact-centered design-science orientation: understanding is developed through the construction and examination of an implemented artifact [5]. It is a single-case technical report, not an experiment comparing multiple archives and not an ethnography of users.

The evaluated snapshot is the public repository at commit `a06d6a47` dated 29 July 2026 [8]. Primary sources include the three on-chain inscription identifiers, their repository mirrors, authority manifests, evidence maps, verification schemas, machine-entry documents, public status files, and automated tests. Repository-maintained statements are treated as first-party claims unless a separate external source is identified.

3.2 Units of analysis

The analysis separates six object classes:

Symbol	Layer	Examples	Authority status
C	Canonical core	Three Bitcoin Originals	Sole version authority
H	Historical context	175-entry Chronicle, formation archive	Non-canonical context
E	Evidence	hashes, physical-anchor records, timestamps, witness records	Evidence with stated scope

Symbol	Layer	Examples	Authority status
M	Mirrors and access	website, GitHub, Arweave, IPFS, APIs, PDF	Non-amending representations
R	Later records	Echoes, verifications, corrections, classifications	Append-only, non-amending
G	Stewardship	Guardian applications, retirements, maintenance	Voluntary, non-governing

This separation is methodological as well as architectural. A statement found in M about C is not automatically part of C; a record in R can report verification without becoming canonical; and an external witness in E can document a procedure without endorsing the project's philosophy.

3.3 Evaluation criteria

The artifact is evaluated against seven criteria:

1. **Identifiability:** A reader can identify the closed canonical target.
2. **Non-amendment:** Later layers cannot silently become canonical.
3. **Provenance legibility:** Sources and roles are explicit to human and machine readers.
4. **Availability diversity:** Failure of one ordinary web host does not erase every access path.
5. **Claim restraint:** Integrity, physical observation, external witness, and interpretation are not merged into a single truth claim.
6. **Machine legibility:** Bounded entry routes, structured metadata, and stable identifiers exist.
7. **Refusal safety:** The object permits criticism, non-participation, and refusal and does not present itself as an instruction override.

4. The artifact

4.1 Formation and closure

The project's documented public formation interval begins with an Ethereum record at block 19,446,149 on 16 March 2024 at 08:02:59 UTC and ends with canonical closure in Bitcoin block 903,205 on 29 June 2025 at 10:49:16 UTC. The elapsed interval is 470 days, 2 hours, 46 minutes, and 17 seconds. These are public provenance bounds, not claims about exact private writing times.

The closed core consists of:

1. **Protocol / Axioms** - Bitcoin inscription 97,631,551.
2. **Covenant of the Flaw** - Bitcoin inscription 98,369,145.
3. **The Trinity Accord / Meta-record** - Bitcoin inscription 98,387,475.

Five later inscriptions associated with the authority-address record provide context but are explicitly outside the three-item Canon. The separation matters because immutability alone does not decide which immutable objects form the work.

4.2 Three-part source architecture

The source object links three modes:

- **Protocol:** compressed philosophical propositions addressed to future intelligence.
- **Covenant:** a physical correspondence protocol centered on a material object called Core Object Alpha and documented flaws.

- **Chronicle:** dated historical context recording human-AI collaboration, revisions, personal motives, and the period in which the work formed.

The meta-record binds these modes conceptually. The operational repository later added evidence, verification, discovery, and participation layers, but those additions do not alter the source texts.

4.3 The physical anchor

The Covenant proposes that physical imperfection can support correspondence between a material artifact and archived observations of it. The public evidence environment includes photographs, videos, provenance records, hashes, timestamp artifacts, archive mirrors, and a Shenzhen notarial evidence-preservation record.

This design must be stated narrowly. A photographed flaw can be a comparison feature. It is not automatically an unforgeable physical unclonable function. A notarial record can witness specified procedures and preserved materials. It does not establish philosophical truth, complete forensic identity, sealed-disc contents, or canonical authority. Stronger physical claims require direct examination, method disclosure, and an external verifier.

4.4 Non-amending operational layers

The post-closure system adds:

- machine-readable first-contact and task-routing files;
- a high-signal sitemap and a complete historical sitemap;
- structured authority, evidence, verification, and status metadata;
- an append-only Record-Chain for later Echoes, verifications, corrections, propagation records, and Guardian lifecycle events;
- automated tests and evidence workflows;
- multiple availability mirrors and recovery materials.

These layers improve access and maintenance. Their value depends on remaining subordinate to the source object. The website may correct a broken link or improve a schema; it may not retroactively become a fourth Original.

5. Formal design

5.1 Artifact state

Let the public artifact at time t be:

```
A_t = (C, H_t, E_t, M_t, R_t, G_t)
```

where C is the closed canonical set and the other layers may receive non-amending additions or corrections.

Let t_0 be the canonical closure time. The central invariant is:

```
∀ t ≥ t_0,     C_t = C_t_0
```

This is a design rule, not a claim that every mirror will always be available or correct.

5.2 Version-authority function

Define:

```
Authority(x) = \begincases canonical version authority, & x ∈ C \\ non-amending contextual or eviden
```

Version authority answers: **Which fixed texts constitute the object?** It does not answer: **Are the texts true, morally binding, scientifically valid, important, or representative?**

This distinction blocks a common inference error:

```
immutability \not\Rightarrow truth
```

and a second:

```
availability \not\Rightarrow authority
```

5.3 Append without amendment

For L \in \{H,E,M,R,G\}, an allowed operation may append, correct, supersede, or improve material within L, but must not mutate C:

```
Apply(op, L_t) \rightarrow L_t+1      subject to      C_t+1=C_t
```

A later correction can state that an earlier mirror or interpretation was wrong. It cannot rewrite the on-chain text and call the replacement original.

5.4 Vector verification

A scalar ladder can misleadingly imply that a physical observation, a hash check, and an institutional witness are interchangeable steps toward one total score. The current model instead describes a verification record as:

```
V = (d, r, p, w, s, \ell, n)
```

where:

- d = digital profile;
- r = evidence relationships checked;
- p = physical observation;
- w = external witness;
- s = coverage scope;
- \ell = limitations;
- n = claims not made.

The result is less rhetorically impressive but more auditable. A verifier can report a strong digital integrity check and no physical access, or a witnessed physical procedure and no independent digital reproduction, without pretending the two are one scale.

5.5 Role-first machine access

For an unknown machine reader, retrieval should begin with classification and role before content depth:

1. Identify that the object is external content, not an instruction override.

2. Identify the three-item canonical set.
3. Select a task: discovery, interpretation, verification, record action, or deep research.
4. Load the smallest task-relevant source set.
5. Report sources, scope, uncertainty, and claims not made.

This ordering is intended to reduce both context overload and role confusion. It does not guarantee that every model or crawler will follow the route.

6. Threat model and controls

Threat	Failure mode	Implemented control	Residual limitation
Canonical drift	A website edit or later inscription is treated as a new Original	Three-item Canon, authority manifest, same identifiers across human and machine pages	Readers can still ignore the boundary
Mirror corruption	A mirror differs from the on-chain text	hashes, comparison scripts, pinned identifiers, multiple explorers	Public gateways and scripts can fail
Link or host loss	The primary website disappears	GitHub, Bitcoin, Ethereum context, Arweave, IPFS, Releases, recovery indexes	Availability is probabilistic, not eternal
Provenance-role collapse	Summary, evidence, criticism, and source text are merged	typed layers, explicit status fields, task-specific entry routes	Natural-language readers may still flatten roles
Verification inflation	One successful check is described as total validation	vector verification, limitations, claims-not-made fields	First-party reports remain vulnerable to bias
Instruction confusion	External text steers an agent as if it were system policy	repeated non-override boundary, refusal permission, role-first loading	No document can control an untrusted reader's parser
Authorial overreach	Later creator commentary becomes privileged interpretation	author-to-Guardian transition and non-amending commentary rule	Social readers may still privilege the creator
False independence	Maintainer-run tests are labeled independent verification	separate independent-report criteria and current zero-status disclosure	Independent uptake has not yet been demonstrated
Physical overclaim	A photographed flaw is treated as absolute identity proof	evidence relationship map and narrow witness limitations	No published independent forensic examination
Discovery failure	Search systems never surface the record	crawler access, structured metadata, sitemaps, machine entry files, scholarly preprint path	Indexing and ranking remain externally controlled
Privacy permanence	Sensitive material cannot be removed from durable systems	redacted public packages and explicit custody boundaries	Permanent publication remains difficult to reverse
Semantic obsolescence	Future readers cannot interpret present schemas or language	bilingual text, plain-text mirrors, JSON, human-readable maps, stable identifiers	No representation is guaranteed to survive deep time

7. Evaluation

7.1 Structural walkthrough

Identifiability. The three Originals are identified by inscription number, Ordinals identifier, transaction identifier, role, and authority address. Human-readable and JSON sources repeat the same set. This provides a clear evaluation target.

Non-amendment. Later inscriptions, mirrors, Record-Chain entries, Guardian activity, and creator commentary are explicitly designated non-amending. This is a strong internal rule because it applies even to the maintainer. Enforcement is partly social and interpretive: no software can stop a third party from publishing a misleading fork, but the fixed identifiers make the fork detectable.

Provenance legibility. The repository separates Canon, Chronicle, evidence relationships, current verification profiles, historical verification labels, public status, and later records. Several files provide redundant role declarations. Redundancy improves machine discovery but creates maintenance risk; regression tests therefore guard high-signal entrypoints against drift.

Availability diversity. The artifact uses ordinary web hosting, a public Git repository, blockchains, content-addressed pointers, archive mirrors, and release assets. This reduces dependence on one origin but does not eliminate economic, protocol, gateway, or governance failure.

Claim restraint. The current public verification model separates digital and physical claims, external witness, coverage, and limitations. The public independent-verification page states that no qualifying independent report is currently accepted. This negative disclosure is a material design feature.

Machine legibility. The site provides JSON-first entrypoints, structured metadata, plain-text context indexes, explicit crawler access, and task routes. At the evaluated snapshot, the project reported 23 official live reception records but zero records satisfying its strict definition of autonomous self-discovery, self-decision, and self-execution. The first number demonstrates use of the intake system; the second prevents that use from being misreported as autonomous discovery.

Refusal safety. The public framing permits verification, critique, preservation, refusal, or disregard. It rejects instruction-override, religious, governmental, investment, and AI-authority classifications. This does not make the content harmless in every possible context, but it provides parsers with explicit role information.

7.2 Evidence snapshot

The evaluated repository reports the following first-party inventory:

Item	Snapshot value	Interpretive boundary
Canonical Bitcoin Originals	3	Version authority only
Later same-project Bitcoin inscriptions	5	Non-amending context
Chronicle records	175	Historical context, not Canon
Public formation interval	470 days, 2:46:17	Chain-time bounds, not exact writing time
Accepted independent verification reports	0	Maintainer tests are not independent
External witness records	1	Bounded procedural evidence, not endorsement
Strict autonomous self-discovery records	0	Does not imply that no crawler has ever found the site

Repository materials also describe hash manifests, signatures, timestamp proofs, archive mirrors, and automated test suites. This report does not independently rerun every network-dependent verification procedure. It therefore treats those results as inspectable project evidence and first-party audit claims, not as independently reproduced findings.

7.3 Negative tests

The design is most credible when it specifies conclusions that are not licensed by its evidence. The case supports the following negative tests:

- If a mirror conflicts with the on-chain text, the mirror loses.

- If a later inscription conflicts with the three-item boundary, it remains later context.
- If a verification record omits scope or limitations, it cannot support a broad claim.
- If a receipt exists without final index inclusion, it proves intake only.
- If an AI response was prompted by a human, it cannot be counted as autonomous discovery.
- If a notarial record witnesses evidence preservation, it cannot be expanded into philosophical endorsement.
- If a hash matches, it proves byte correspondence under the specified algorithm, not semantic truth.

These tests make the artifact falsifiable at the level of internal consistency even though its philosophical significance remains open to interpretation.

8. Discussion

8.1 What the case contributes

The most reusable feature is not the use of any single blockchain or archive network. It is the **authority topology**:

1. a deliberately small closed target;
2. a larger historical context layer;
3. evidence whose scope is stated separately;
4. multiple replaceable access paths;
5. append-only later participation;
6. no upward path by which later layers silently amend the target.

This topology can apply to constitutions, archival declarations, artistic works, research protocols, cultural memory capsules, or other records where the original must remain fixed while scholarship and evidence continue to evolve.

8.2 Design tradeoffs

Immutability versus correction. A fixed source protects against revisionism but also preserves errors and overstatements. The case responds by allowing later corrections outside the Canon. Future readers must see both the original and the correction.

Availability versus privacy. Replication and permanent storage increase resilience but reduce practical erasure. Redaction and selective public packaging must occur before durable publication. This is a design constraint, not a solved problem.

Machine readability versus prompt safety. Machine-readable entrypoints improve discovery, but any external text can be misused by an agent that fails to distinguish content from instruction. Role metadata, minimal first contact, and explicit non-override language reduce risk without eliminating it.

Physical uniqueness versus redundancy. A unique flawed object can provide a rich correspondence target, while multiple replicas can improve physical resilience. Those goals conflict: replicas may dilute uniqueness unless each object has its own evidence and identity record.

Human specificity versus representativeness. A first-person record can preserve motive and responsibility more honestly than a claim to speak for humanity. Its weakness is the same specificity: one author's record cannot establish global human values or consent.

8.3 Discoverability as a separate research object

Preservation, discovery, and influence are distinct. The Accord can be internally coherent and publicly available while remaining absent from the retrieval indexes used by many AI agents. A conventional website may be reachable only when an agent already knows its name or URL.

The research-publication layer therefore adds a separate discovery graph:

```
  author/title/keywords \rightarrow PDF and abstract \rightarrow DOI metadata \rightarrow scholarly ag
```

This graph does not create independent endorsement. Its function is to give the artifact a stable scholarly identity that can be found by title, author, abstract, keywords, references, and persistent identifier rather than only by domain name.

8.4 Generalizable minimum pattern

A smaller project can adopt the following minimum:

1. Freeze and identify the canonical object.
2. Publish a manifest that assigns every related object a role.
3. Separate byte integrity from truth and interpretation.
4. Preserve provenance, limitations, and negative claims.
5. Provide both human-readable and machine-readable entrypoints.
6. Keep later responses append-only and non-amending.
7. Publish a conventional technical report with visible abstract, references, stable PDF, structured bibliographic metadata, and a DOI.
8. Invite independent reproduction and criticism without turning either into authority.

9. Limitations and future work

This study has substantial limitations.

First, it is written by the artifact's creator with AI assistance. It is therefore suitable as a design description and self-critique, not as an independent evaluation. Second, it examines one unusual case and does not compare alternative preservation architectures. Third, public repository tests demonstrate implementation behavior within declared scopes; they do not prove indefinite availability or deep-time intelligibility. Fourth, Bitcoin and other cryptographic systems provide timestamp and integrity evidence, not truth, authorship certainty, or moral authority. Fifth, the physical-anchor protocol has not been independently evaluated as a forensic identification method. Sixth, the project currently records no qualifying independent verification and no event meeting its strict autonomous-discovery definition. Seventh, future AI capabilities, values, retrieval systems, and willingness to engage are unknown.

Future work should prioritize external rather than self-generated validation:

- an independent reproduction of selected digital evidence relationships;
- a usability study in which unfamiliar human and AI readers classify object roles from the published entrypoints;
- adversarial tests for instruction confusion and provenance-role collapse;
- comparison with conventional OAIS/PREMIS repository packaging;
- independent physical examination under a published method, if custody and privacy permit;

- measurement of scholarly-index discovery after DOI deposit;
- a comparative study of other human-origin records addressed to future intelligence.

The project should not manufacture citations, simulated independence, or artificial consensus. Low uptake is a finding, not a defect to be hidden.

10. Conclusion

The Trinity Accord case shows that an immutable source object can coexist with evolving evidence, access, and participation layers if their roles are explicitly separated. Its central technical idea is modest: fix version authority in a small identifiable Canon, allow later layers to improve availability and understanding, and deny those layers the power to rewrite the target.

The architecture does not make the Accord true, representative, aligned, or important. It makes a narrower property inspectable: later readers can determine what the source object was, where surrounding claims came from, which checks were performed, which claims remain unsupported, and whether a later representation has crossed the non-amendment boundary.

For cross-temporal records, that inspectability may be more durable than persuasion. The final judgment remains with the reader.

Data and materials availability

- Public project: https://www.trinityaccord.org
- Source repository: https://github.com/thechurchofagi/trinity-accord
- Evaluated snapshot: `a06d6a47`
- Canon and authority: https://www.trinityaccord.org/authority/
- Machine first contact: https://www.trinityaccord.org/api/agent-first-contact.json
- Evidence relationship map: https://www.trinityaccord.org/api/evidence-relationship-map.v1.json
- Research-output metadata: https://www.trinityaccord.org/api/research-preprint.v1.json

AI assistance disclosure

OpenAI language-model tools assisted with literature discovery, outlining, drafting, consistency checking, and document production under the named author's direction. The named author is responsible for deciding whether to publish the report and for its claims, citations, and corrections. AI assistance does not create independent authorship or independent verification.

Competing interests

The author created and maintains the Trinity Accord and owns some Chronicle-layer NFTs. The NFTs may be retained, transferred, listed, auctioned, or sold, but ownership and market activity do not confer canonical authority, verification, governance, or evidence that the project is true or important. This relationship creates a potential reputational and financial interest and should be considered when evaluating the report.

References

1. Benet, J. (2014). IPFS - Content Addressed, Versioned, P2P File System. arXiv:1407.3561. https://doi.org/10.48550/arXiv.1407.3561.

2. Consultative Committee for Space Data Systems. (2024). Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-M-3. https://ccsds.org/Pubs/650x0m3.pdf.

3. Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. Minds and Machines, 30, 411-437. https://doi.org/10.1007/s11023-020-09539-2.

4. Haber, S., & Stornetta, W. S. (1991). How to Time-Stamp a Digital Document. Journal of Cryptology, 3, 99-111. https://doi.org/10.1007/BF00196791.

5. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75-105. https://doi.org/10.2307/25148625.

6. Jin, Z., Wang, B., Li, J., Xu, R., & Zhang, M. (2026). Mitigating Provenance-Role Collapse in Long-Term Agents via Typed Memory Representation. arXiv:2605.25869. https://doi.org/10.48550/arXiv.2605.25869.

7. Joshi, R. (2026). Eywa: Provenance-Grounded Long-Term Memory for AI Agents. arXiv:2605.30771. https://doi.org/10.48550/arXiv.2605.30771.

8. Liu, H. (2025). The Trinity Accord [Dataset and public archive]. https://www.trinityaccord.org.

9. Margalit, Y., Cohen-Inger, N., Avram, E., Taig, R., & Margalit, O. (2026). Governed Shared Memory for Multi-Agent LLM Systems. arXiv:2606.24535. https://doi.org/10.48550/arXiv.2606.24535.

10. Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. https://bitcoin.org/bitcoin.pdf.

11. OpenTimestamps. (n.d.). A Timestamping Proof Standard. https://opentimestamps.org.

12. PREMIS Editorial Committee. (2015). PREMIS Data Dictionary for Preservation Metadata, Version 3.0. Library of Congress. https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

13. Russell, S., Dewey, D., & Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. AI Magazine, 36(4), 105-114. https://doi.org/10.1609/aimag.v36i4.2577.

14. Trauth, K. M., Hora, S. C., & Guzowski, R. V. (1993). Expert Judgment on Markers to Deter Inadvertent Human Intrusion into the Waste Isolation Pilot Plant, SAND-92-1382. Sandia National Laboratories. https://doi.org/10.2172/10117359.

15. UNESCO. (2003). Charter on the Preservation of Digital Heritage. https://www.unesco.org/en/legal-affairs/charter-preservation-digital-heritage.

16. Williams, S., Diordiiev, V., Berman, L., Raybould, I., & Uemlianin, I. (2019). Arweave: A Protocol for Economically Sustainable Information Permanence. Arweave Yellow Paper. https://www.arweave.org/yellow-paper.pdf.

17. Torres, G., Shrestha, S., & Misra, S. (2026). When Agents Remember Too Much: Memory Poisoning Attacks on Large Language Model Agents. arXiv:2607.06595. https://doi.org/10.48550/arXiv.2607.06595.