CJCI Volume 1, Issue 23 | Version 1.2

AI Watermarking, Digital Provenance, and FieldID

An Industry and Technical Assessment of Synthetic-Content Trust

From Content Marking to Governed Evidence Composition

Author	Ivan Silva
Affiliation	Carlonoscopen, LLC
ORCID	0009-0005-2284-8891

Publication	Carlonoscopen Journal of Coherence Intelligence (CJCI), Volume 1, Issue 23
CJCI ISSN	Digital 3069-874X; Print 3071-0022
Article classification	Interdisciplinary technical assessment and architecture paper
Archive target	Zenodo
Reserved DOI	10.5281/zenodo.21909033
Version	1.2 - Final Frozen
Manuscript date	August 12, 2026

Publication state	Final frozen by human author approval; mechanism evidence bound; Figure 1 embedded; Zenodo activation remains an external author action

Contents

Abstract

AI watermarking is moving rapidly from research into production infrastructure. Google deploys SynthID across supported text, image, audio, and video generation. OpenAI combines C2PA Content Credentials with SynthID for supported images and applies SynthID to supported generated audio. Anthropic has begun introducing machine-readable marking for supported Claude models, including embedded text watermarks and signed provenance metadata for supported files. In parallel, Article 50 of the European Union AI Act became applicable on August 2, 2026, accelerating the transition from voluntary provenance experiments toward operational transparency requirements.

These developments are important, but they do not reduce the synthetic-content problem to a watermark-detection problem. A watermark may provide evidence that a particular generative system participated in producing content. A provenance manifest may record creation and transformation history. A digital credential may identify an actor or organization. A cryptographic signature may establish control of a key. None of these mechanisms independently determines whether the represented claim is true, whether the use of the content was authorized, or whether multiple independently valid provenance signals tell a mutually consistent story. C2PA explicitly distinguishes provenance from factual truth, and emerging SCITT work similarly recognizes that a transparently registered signed statement may still be false.

This paper examines the current technical and regulatory landscape and evaluates it against the frozen FieldID and CNX architecture developed by the author. FieldID is not treated here as a proposed replacement for C2PA, SynthID, CAWG identity assertions, or other provenance standards. Its more defensible potential role is as a deterministic structural-identity substrate within a governed evidence-composition architecture. Under such an architecture, watermarks, provenance records, credentials, physical authentication, structural identity, Internet recurrence, and other signals remain independent evidence sources. CNX governs their admissibility, while downstream reasoning preserves contradiction, uncertainty, falsification requirements, and proof burden.

The paper's central claim is therefore narrower than solving misinformation and broader than watermark detection: a trustworthy synthetic-content infrastructure requires separation between identity, provenance, authentication, authority, and truth, together with an explicit mechanism for handling missing, contradictory, compromised, and changing evidence.

Keywords: AI watermarking; digital provenance; C2PA; Content Credentials; SynthID; synthetic media; misinformation; deepfakes; FieldID; CNX; authority separation; evidence composition; content authenticity; digital identity; Internet trust

Author Note and Publication Boundary

This paper combines externally sourced technical and regulatory evidence with an architectural analysis of FieldID, CNX, and the previously developed FieldID Internet ingestion path.

The paper does not claim that FieldID has been empirically demonstrated to identify misinformation, authenticate arbitrary Internet content after unrestricted transformation, outperform SynthID, replace C2PA, or determine factual truth.

The frozen FieldID implementation establishes deterministic structural identity processing within its tested domains. The frozen BSM/FWI Internet extension establishes deterministic raw acquisition and non-semantic structural ingestion under its defined conformance tests. Those results are internal implementation evidence. They are not equivalent to independent validation of an Internet-scale content-authenticity system.

The proposed integration between FieldID/CNX and external provenance standards is therefore classified as an architectural proposal unless otherwise identified. This distinction is deliberate.

Version 1.2 incorporates a reproduced and tested FieldID structural-validity mechanism. The frozen implementation contains a mechanism for evaluating structural-identity validity through stability, curvature, and detuning. Its validation suite includes both valid and invalid identity conditions and demonstrates stable acceptance after sufficient repeated observations. This mechanism result is classified VAL within the reproduced test configuration [23].

The present evidence does not establish that seedless AI-generated content produces structural collapse. The author reports a prior internal observation of rapid structural-field collapse when FieldID was applied to seedless AI-regenerated derivative-content configurations. However, the historical experimental custody for that specific observation was not recovered. The seedless-collapse proposition is therefore classified as a prior author observation and HYP, not as VAL. It requires controlled comparative testing with documented custody before it can be advanced.

Field collapse is defined prospectively as sustained post-initialization failure to maintain a valid FieldID structural identity. A transient invalid state caused solely by insufficient observations is not classified as collapse.

The complementary proposition, that a human-authored architectural seed may remain detectably persistent through AI-assisted rendering, is classified HYP. Recognition independence, the possibility that any such structural relation could be detected without cultural knowledge of the seed or author, is also HYP.

1. Why This Paper Exists

The synthetic-content debate is frequently compressed into one question: Was this generated by AI?

That question matters, but it is becoming insufficient. An AI-generated scientific visualization can be legitimate. A human-written political claim can be false. A genuine photograph can lose its provenance metadata when screenshotted. A validly signed corporate statement can contain an error. A synthetic voice can be used with the speaker's authorization. A real video can be deceptively edited. A malicious actor can attach new provenance to manipulated content. An anonymous source can publish genuine evidence without wanting to reveal civil identity.

The binary classification AI or human does not resolve these cases.

The actual problem contains several different questions:

1. Was AI involved in generation or transformation?

2. What system produced or modified the artifact?

3. Who or what claims responsibility for it?

4. Was that actor authorized to perform that action?

5. What transformations occurred after creation?

6. Are the provenance records internally valid?

7. Do independent provenance signals agree?

8. Is the relevant credential still valid?

9. Has an underlying security assumption failed?

10. Is the substantive claim represented by the content true?

These questions belong to different technical layers. The central engineering risk is therefore not merely failure to detect AI. It is collapse between layers that answer different questions.

2. Research Questions, Method, and Evidence Discipline

2.1 Research Questions

RQ1. What can current watermarking and digital-provenance technologies establish reliably, and what remains outside their authority?

RQ2. What failure modes appear when watermarking, provenance, identity, and authentication mechanisms are composed rather than evaluated independently?

RQ3. Does the frozen FieldID/CNX architecture contain properties relevant to these unresolved composition problems without modifying its existing identity and authority boundaries?

RQ4. What would have to be demonstrated before FieldID/CNX could responsibly be presented as part of an Internet-scale synthetic-content assurance architecture?

2.2 Evidence Classes

REF: Externally supported claim based on standards, regulation, vendor documentation, or published research. VAL: Internally validated FieldID/CNX implementation result. PROP: Architectural proposal derived from existing components but not yet experimentally validated for this application. HYP: Research hypothesis requiring dedicated testing. ILL: Illustration used to explain architecture, not evidence that implementation exists.

A PROP claim is not silently promoted to VAL. A VAL claim inside a bounded internal test environment is not silently promoted to an Internet-scale empirical claim. A REF claim concerning provenance is not silently promoted into a claim about truth.

2.3 Method

The analysis combines four evidence layers. First, current official documentation from Google DeepMind, OpenAI, Anthropic, the European Commission, C2PA, CAWG, NIST, and IETF work on SCITT. Second, peer-reviewed watermarking research, including foundational work on statistical language-model watermarking, SynthID-Text, and watermark reliability. Third, adversarial research evaluating paraphrasing, spoofing, provenance manipulation, and cross-layer contradiction. Fourth, the frozen FieldID/CNX architecture and its Internet-facing BSM/FWI extension.

Vendor documentation is used to establish what providers say they currently deploy or support, not as independent validation. Preprints are identified as such and treated as red-team evidence rather than settled scientific consensus.

2.4 Explicit Non-Claims

This paper does not claim any of the following: universal AI-content detection; universal truth detection; ownership determination from watermarking; legal attribution from FieldID; that absence of a watermark proves human authorship; that presence of a watermark proves misinformation; that a valid signature proves factual accuracy; that repeated Internet appearance establishes truth; that FieldID currently survives unrestricted paraphrasing or generative regeneration; that FieldID currently provides a production Internet trust service; that C2PA, SynthID, or CAWG are technically obsolete; that one centralized authority should decide what Internet content is true; that seedless AI-generated content has been validated to produce structural collapse; that seed persistence through AI-assisted rendering has been demonstrated; that FieldID currently determines originating authorship; or that a valid AI-generation watermark is technically wrong when it correctly reports AI participation.

These are fixed boundaries of the present paper.

3. Evidence Map at a Glance

The evidence map is intentionally bounded.

REF: AI watermarking is deployed by major frontier providers. Deployment differs by modality and model. REF: Article 50 transparency obligations apply from August 2, 2026. Legal interpretation remains jurisdiction-specific. REF: C2PA supports durable provenance and soft bindings. Soft binding is not exact cryptographic identity. REF: Actor identity can be associated with C2PA assets through CAWG identity assertions. Identity assertion does not itself establish ownership. REF: Watermarks can be degraded or spoofed under adversarial conditions. Results depend on scheme and threat model. REF: Independently valid provenance signals can contradict within studied configurations. VAL: FieldID separates structural identity from authentication within the frozen architecture. VAL: FWI forbids semantic truth assignment within its frozen specification. VAL: The frozen FieldID implementation contains a reproduced and tested mechanism for evaluating structural-identity validity through stability, curvature, and detuning. 119/119 tests passed under internal reproduction. This result is bounded to the reproduced configuration [23]. HYP: Seedless AI-generated content may produce structural-field collapse detectable by the FieldID validity mechanism. The author reports a prior internal observation, but historical experimental custody was not recovered. Controlled comparative testing is required. PROP: FieldID could participate in cross-signal evidence composition. HYP: FieldID may improve transformed-artifact continuity. This has not been demonstrated against the required attack suite. HYP: FieldID may detect structural persistence associated with a documented human-authored seed through AI-assisted rendering. This has not been tested. HYP: Any seed-persistence signal may be detectable independently of cultural recognition of the seed. This has not been tested.

4. The Problem Is Larger Than Watermark Detection

A useful first separation is generation provenance is not factual truth.

C2PA states this directly. Provenance can provide evidence about origin, history, authenticity, and changes, but provenance alone cannot determine whether the content is true, accurate, or factual [9, 11].

The IETF SCITT architecture reaches a related conclusion from a different direction. SCITT provides identifiable signed statements, transparent registration, append-only records, receipts, and auditable history. Yet the architecture recognizes that an issuer may make a false statement, either intentionally or unintentionally. Registration establishes that the issuer produced the statement. It does not make the statement accurate [14].

This is a critical convergence. Two different provenance architectures arrive at essentially the same boundary: authenticated origin is evidence about origin, not automatic authority over truth.

That boundary should become foundational to synthetic-content infrastructure.

5. Current Industry State

5.1 Google and SynthID

Google DeepMind's SynthID embeds imperceptible machine-detectable signals into supported AI-generated images, audio, text, and video [3]. For text, SynthID modifies token sampling so that a detectable statistical signal is introduced without retraining the underlying model. The production-oriented SynthID-Text method was described in Nature in 2024, with the authors reporting high detection accuracy, preserved text quality, and low operational overhead within their evaluated settings [4]. The importance of SynthID is therefore not merely conceptual. It demonstrates that text watermarking can operate at production scale. That does not imply universal robustness against arbitrary transformation.

5.2 OpenAI

OpenAI currently uses a multilayer provenance strategy for supported media. Its 2026 provenance program combines C2PA Content Credentials and SynthID for supported generated images. A July 31 update extended SynthID watermarking to supported audio generated through OpenAI tools and expanded verification support [7]. This architecture is itself instructive. OpenAI is not relying on one provenance mechanism. It is combining open provenance standards, durable watermarking signals, and verification tooling. The direction of travel is already toward layered evidence.

5.3 Anthropic

Anthropic materially changed the frontier-provider landscape in August 2026. According to Anthropic's August 2026 documentation, supported Claude models launched on or after August 2 include machine-readable marking. Generated text receives embedded watermarking, while supported generated files can include signed provenance metadata [8]. It is therefore no longer accurate to characterize large-scale text watermarking as essentially unique to Gemini. The market is moving quickly.

6. Regulation Is Converting Provenance Into Infrastructure

Article 50 of the European Union AI Act became applicable on August 2, 2026. It introduces transparency obligations concerning AI-generated and manipulated content, including provider-side machine-readable marking and deployer-side disclosure requirements for specified forms of deepfake and AI-generated public-interest content [1, 2].

The European Commission's Code of Practice provides a voluntary mechanism for demonstrating compliance with relevant Article 50 obligations [1].

The significance is larger than one regulation. Machine-readable content provenance is moving from an optional platform feature toward a component of regulated digital infrastructure.

For model-agnostic systems, this changes the engineering problem. The question is no longer whether an architecture should recognize provenance mechanisms. The question is how multiple mechanisms should be composed without allowing one to become a false oracle.

7. What Watermarking Can Do

Watermarking remains a serious and useful technical approach.

Kirchenbauer et al. introduced a statistical watermarking framework for language models in which a pseudorandomly selected subset of tokens receives increased sampling probability. The resulting token sequence can then be evaluated statistically without access to the model's parameters [5]. Subsequent work evaluated these techniques under more realistic conditions and showed that watermark reliability depends on factors including generation length, watermark strength, sampling conditions, and transformation [6]. SynthID-Text provided additional evidence that language-model watermarking can be engineered for production deployment [4].

The correct conclusion is therefore not that watermarking fails. The correct conclusion is that watermarking provides one useful evidence channel whose reliability depends on the watermark, content, detector, transformation, and threat model.

8. What Watermarking Cannot Safely Become

Watermarking becomes dangerous when a probabilistic or recoverable signal is silently transformed into authority.

Sadasivan et al. demonstrated recursive paraphrasing attacks that substantially degraded multiple classes of AI-text detectors, including watermark-based systems. The same work also demonstrated spoofing attacks intended to make non-AI text appear AI-generated [15]. These results do not prove that all watermarks can always be removed. Other research reports meaningful persistence under some transformations and sufficiently long texts [6].

Together, the results imply something more useful: watermark persistence is an empirical property, not an assumption.

A verifier should therefore never translate no detectable watermark into human generated. Likewise, watermark detected should not automatically become malicious, false, or unauthorized.

9. C2PA Is No Longer Just Metadata

Any serious comparison with C2PA must account for the present architecture rather than early descriptions of Content Credentials.

C2PA 2.4 supports both hard and soft bindings [9, 10]. A hard binding cryptographically associates provenance information with an asset. A soft binding can use mechanisms such as invisible watermarking or content fingerprinting to recover a detached manifest after embedded provenance has been removed [10]. C2PA describes use cases involving metadata stripping, provenance substitution, manifest recovery, watermark resolution, and comparison between locally embedded and remotely recovered provenance [10]. It also cautions that soft bindings are not guaranteed to be exact. Different renditions can intentionally resolve to the same binding, and different content can potentially collide through error or attack.

This matters for FieldID. FieldID cannot claim differentiation merely because it seeks identity persistence under transformation. Modern C2PA already operates in that territory.

The distinction between a C2PA soft binding and FieldID structural identity should therefore be stated carefully. A C2PA soft binding is fundamentally a provenance association mechanism. A watermark, fingerprint, or related resolver is used to associate an observed asset or transformed rendition with provenance information that may no longer be embedded directly in the artifact. Its purpose is to recover or locate an appropriate provenance relationship.

FieldID, by contrast, was not designed as a manifest locator. Its frozen architecture produces a high-dimensional structured identity representation derived from the observed field [19]. The relevant proposed distinction is therefore not that FieldID performs a stronger version of a C2PA fingerprint. It is that the two mechanisms produce different types of evidence.

A soft binding asks, in effect: Can this observed asset be associated with a provenance record?

FieldID asks a different structural question: What stable or bounded identity structure can be extracted from the observed field itself?

This distinction is architectural, not competitive. No result presented in this paper demonstrates that FieldID is more resilient than a C2PA soft binding, produces fewer false associations, or survives transformations that defeat existing provenance-recovery systems.

The potential evidence gap is narrower. If a soft binding cannot resolve uniquely, if provenance is unavailable, or if several independently valid provenance mechanisms disagree, a FieldID representation could potentially provide an additional structural observation that is independent of the provenance claim itself.

Whether that observation provides useful discriminatory information under real Internet transformations is a research hypothesis and must be tested.

10. Identity Is Not Ownership

The Creator Assertions Working Group extends the provenance ecosystem toward human and organizational identity. CAWG identity assertions can associate individuals or organizations with C2PA assets and with particular actions or assertions in the provenance chain [12]. Its use cases include journalism, human-rights evidence, political communications, and commercial impersonation.

But the specification draws an important legal and technical boundary. An identity assertion should not itself be construed as conveying attribution or ownership [12].

This distinction corrects a common description of AI watermarking as something that can simply prove ownership. It cannot.

A watermark can provide generation evidence. A credential can provide identity evidence. A signature can establish control over a signing key. A provenance chain can record history. Ownership is a separate question.

10.1 Governed AI-Assisted Authorship, Seed, and Recognition Asymmetry

A concrete boundary case clarifies why generation evidence, authorship evidence, and ownership evidence should remain separate.

This paper was produced through Writers' Loop Engineering [21], a governed AI-assisted authorship process in which manuscript structure, claim classes, evidence boundaries, continuity constraints, and acceptance criteria were established before prose was accepted. AI-assisted tools participated materially in research support, drafting, revision, and document production. Those contributions are not erased. The human author retained authority over the research question, architectural interpretation, claim boundaries, synthesis, acceptance decisions, and publication state.

The same separation extends into the research program described by the paper. The frozen FieldID RSP V1.7 architecture [19], the FWI v1.1.0 Internet-ingestion specification [20], and related architecture, implementation, test, and freeze decisions were produced through an author-directed engineering process. Draft Request Specification Protocols were authored under the author's direction. AI tools may have refined their language, and AI code-assistance tools may have generated implementation code from revised specifications under author review. AI participation at the rendering and implementation layers does not by itself answer who originated the architectural decisions those artifacts encode.

For this work, the author identifies an originating architectural seed: the decision to frame synthetic-content trust through a separation between identity, provenance, authentication, authority, and truth, and to ask whether those relationships can be structurally preserved rather than collapsed into one detector or one scalar verdict. The seed is compact relative to the prose, specifications, code, and publication artifacts rendered from it.

This creates a proportion problem.

A useful illustration comes from legal document production. A very large production can contain a relatively small amount of decision-critical material. The important evidence is not absent, but locating it can require disproportionate effort because its volume is small relative to the surrounding material. In an adversarial legal setting, that volume may be deliberate. In a transparent governed AI-assisted authorship process, it need not be adversarial at all. The analogy is only about proportion: a compact originating structure can be surrounded by a much larger rendered surface.

The same problem appears in AI-assisted work. A detector operating on rendered prose may correctly identify AI-associated generation evidence. A code-analysis system may identify AI assistance in implementation. A specification may show evidence of AI-assisted editing. Each observation can be locally valid. None, by itself, determines who originated the architectural structure, who retained authority over the development process, or whether the rendered artifacts descend from a coherent originating seed.

Recognition introduces a second asymmetry.

Consider E = mc^2. Its association with Albert Einstein is culturally established. If an AI system rendered a long exposition organized around that equation, a generation watermark could correctly identify AI participation in the prose without displacing the recognized origin of the equation. External cultural knowledge preserves attribution of the seed even when the rendered surface is heavily AI-assisted.

Now consider an original but culturally unrecognized formulation. The structural relationship can be analogous: a human originates a compact formulation, and AI tools render a much larger exposition around it. But cultural recognition supplies no independent attribution channel. If the evaluator observes only the rendered surface and overinterprets AI participation as originating authorship, the human-originating contribution can become invisible even though it exists.

The asymmetry therefore does not show that watermarking fails at watermarking. A watermark can correctly report AI participation. The failure appears when that valid observation is promoted into a stronger conclusion about originating authorship or intellectual structure.

In that failure mode, a system can unintentionally privilege famous seeds over unknown ones. Recognized seeds retain attribution because external knowledge overrides a simplistic inference from the generated surface. Unknown seeds receive no equivalent protection. In that failure mode, the system is no longer functioning as a provenance architecture. It is functioning as a fame detector.

This is an illustrative inference, not a measured result. It identifies a hypothesis that provenance and structural-identity systems should be able to test without importing fame, author identity, or semantic recognition as a hidden variable.

The FieldID question is correspondingly narrower. The hypothesis is that an originating seed may induce a structural relationship that remains partially observable through downstream renderings. FieldID provides a framework in which that possibility can be tested. It has not yet been demonstrated for seeded AI-assisted authorship.

The counterbalancing observation comes from the opposite pole. The author reports a prior internal observation in which FieldID was applied to configurations where AI regenerated derivative or clickbait-like material from Internet sources without human architectural direction. In that prior observation, the field structure rapidly collapsed.

However, the historical experimental custody for that specific observation was not recovered. The observation is therefore classified as a prior author observation and HYP, not as VAL. It cannot be presented as a validated result. It is reported here as a hypothesis-generating observation that motivates controlled comparative testing.

What has been validated (VAL) is the mechanism itself: the frozen FieldID implementation contains a reproduced and tested structural-identity validity mechanism based on stability, curvature, and detuning. Its test suite includes both valid and invalid identity conditions and demonstrates stable acceptance after sufficient repeated observations [23]. This means the measurement mechanism exists and functions within the reproduced configuration. What it produces when applied to seedless AI content under controlled conditions remains to be tested.

Field collapse is defined prospectively as sustained post-initialization failure to maintain a valid FieldID structural identity. A transient invalid state caused solely by insufficient observations is not classified as collapse. This definition makes the proposition testable and prevents initialization artifacts from being mislabeled as collapse.

The prior observation nevertheless supplies a hypothesized pole. The open pole is persistence. Can a documented human-authored seed retain a measurable structural relationship through AI-assisted rendering? Can that relationship be distinguished from collapse behavior under a controlled seedless condition? Can it be detected without semantic recognition of the author, equation, framework, or cultural fame of the seed?

Those are HYP questions. The validation program in Section 23 now makes them explicit.

The relevant distinction is therefore not detection versus no detection. It is evidence type and inferential authority. AI-generation evidence can be completely correct about the rendered artifact and still be insufficient for attribution of originating intellectual structure.

The paper you are reading contains AI-assisted prose and was produced through a broader program in which AI tools have also assisted specification refinement and implementation. The author identifies the originating architecture and publication authority as human-directed. Those facts can coexist without contradiction.

The structural thesis can therefore be stated carefully: the seed may carry structural identity; a rendering may preserve, transform, or obscure that relationship; and seedless AI generation may produce sustained structural invalidity under the prospective collapse definition. Whether that contrast exists is a falsifiable hypothesis for controlled testing. Distinguishing persistence from collapse, without substituting cultural recognition for evidence, is the research question FieldID is being asked to resolve.

11. The Cross-Layer Composition Problem

Perhaps the most important external research result for the present FieldID/CNX discussion is not an attack that removes a watermark. It is an attack that leaves multiple mechanisms functioning.

Nemecek et al. formalized an Integrity Clash, demonstrating configurations in which an asset carries a cryptographically valid C2PA provenance representation while an independent watermark provides contradictory evidence about AI generation. The mechanisms can each pass their own verification procedure while producing an inconsistent combined story [16].

No cryptographic primitive has to fail. The failure occurs in composition.

This is a deeper architectural problem than detector accuracy. If system A says valid and system B says valid, the combined system cannot automatically conclude therefore globally valid. The two outputs may refer to different claims.

This leads to a general principle: local verification does not automatically produce global authority. That principle is already familiar within CNX.

12. FieldID's Frozen Boundary

FieldID was not designed as a watermarking system.

The frozen FieldID RSP V1.7 defines FieldID as an independent structural-identity system whose output is a high-dimensional identity representation rather than a scalar identifier [19]. Its processing path was defined structurally as Embedding -> Transforms -> Invariants -> Contraction -> Identity -> Stability -> FIP.

The integrated system also preserves an explicit separation between FieldID identity processing and CNX authentication [19]. The frozen inventory describes FieldID as responsible for canonical signal construction and identity extraction, while authentication, PUF verification, device trust, and related trust decisions remain separate CNX responsibilities.

This is important because identity is not authentication, authentication is not authority, and authority is not truth.

Those separations are not incidental to this paper. They are the reason FieldID may be relevant to the provenance problem at all.

13. The Internet Extension: BSM and FWI

The previously developed Internet-facing extension introduces a second relevant property.

The FieldID Web Ingestion Module converts browser-acquired raw content into deterministic structural representations. It explicitly prohibits semantic interpretation at the ingestion stage [20]. FWI can segment raw content, measure statistical properties, detect repetition, record position, generate hashes and signatures, and construct adjacency, similarity, and recurrence relationships.

But it may not infer meaning, identify semantic entities, classify topical content, infer source trustworthiness, infer source authority, or treat recurrence as truth.

The underlying principle is particularly relevant to misinformation: a repeated falsehood is still a falsehood. Recurrence is evidence of persistence or propagation. It is not evidence of correctness.

The implemented BSM/FWI boundary subsequently passed its defined deterministic and conformance checks, including raw preservation, exact structural matching, non-semantic treatment, and dual structural output [20]. Those tests demonstrate conformance to the frozen implementation contract. They do not demonstrate fake-news detection.

14. What FieldID May Add

The existing ecosystem already provides powerful mechanisms for preserving or recovering provenance. FieldID should therefore not be presented as another provenance locator.

Its proposed contribution is different: structural identity as an independent evidence surface.

C2PA hard binding primarily provides cryptographic asset-to-manifest association and exact provenance binding. C2PA soft binding provides recovery or association after separation or transformation and returns a candidate provenance relationship. AI watermarking provides an embedded generation or processing signal. Actor credentials provide actor or organizational identity. FieldID provides a high-dimensional structural characterization of the observed field.

The proposed value of FieldID is therefore not that it replaces or outperforms these mechanisms. It is that it may produce a different class of evidence.

That distinction becomes potentially useful when provenance is incomplete or contradictory. A provenance mechanism describes a claimed history or resolves an asset toward that history. FieldID is intended to characterize the observed signal independently of that claimed history. This creates a possible additional comparison surface: claimed provenance versus observed structural identity.

The usefulness of that comparison has not yet been validated against C2PA soft bindings, modern watermark systems, or Internet-scale transformations. It remains a PROP at the architecture level and a HYP where transformation resilience or discrimination is concerned.

The research question is therefore not whether FieldID is better than C2PA soft binding. It is whether an independently derived structural identity provides useful evidence when provenance mechanisms are absent, ambiguous, transformed, or mutually inconsistent. That question is experimentally testable.

A further observation is relevant. The author reports a prior internal observation in which FieldID was applied to configurations where AI regenerated derivative or clickbait-like material from Internet sources without human architectural direction. In that prior observation, the field structure rapidly collapsed.

However, the historical experimental custody for this observation was not recovered. The observation is classified as a prior author observation and HYP, not as VAL. It motivates controlled comparative testing but does not itself constitute a validated result.

What has been validated (VAL) is the FieldID structural-identity validity mechanism itself: a reproduced test suite of 119 passed and 0 failed demonstrating the mechanism across valid and invalid identity conditions using stability, curvature, and detuning [23]. This establishes the bounded measurement mechanism. What it observes when applied to seedless AI content under controlled, documented conditions remains to be tested.

Field collapse is defined prospectively as sustained post-initialization failure to maintain a valid FieldID structural identity. A transient invalid state caused solely by insufficient observations is not classified as collapse.

The observation suggests that structural organization may provide an evidence surface different from simple AI participation. It does not establish the complementary proposition. Whether a documented human-authored seed remains detectably persistent through AI-assisted rendering is HYP. Whether any such persistence can be detected independently of cultural recognition is also HYP.

The expanded research question is therefore: can FieldID distinguish structural persistence associated with a documented seed from sustained structural invalidity under a controlled seedless condition, while remaining independent of semantic recognition, fame, or author identity?

15. From Watermarking to Governed Evidence Composition

The architecture suggested by the combined evidence is not a detector followed by a binary REAL/FAKE verdict. A more defensible architecture preserves several independent evidence surfaces and introduces an explicit governance boundary before downstream reasoning.

HRE stands for Horn Reasoning Engine. In the frozen FieldID/CNX architecture, HRE is the downstream reasoning core that operates on inputs admitted through CNX [19]. CNX governs admissibility, trust boundaries, and authority. HRE preserves and evaluates relationships among admitted evidence, including contradiction, recurrence, temporal coherence, and unresolved proof burden. It does not authenticate a source or independently grant authority.

Within the synthetic-content architecture proposed here, standards-based provenance, generation watermarks, FieldID structural identity, and independent authentication evidence remain distinct until they reach CNX. CNX governs whether the evidence is admissible. HRE then operates on the admitted evidence without erasing its type or its contradictions. The resulting state is preserved in an evidence ledger and presented through a governed interface.

The HRE component itself is part of the frozen architecture. Its use in this particular provenance-composition profile is proposed rather than validated.

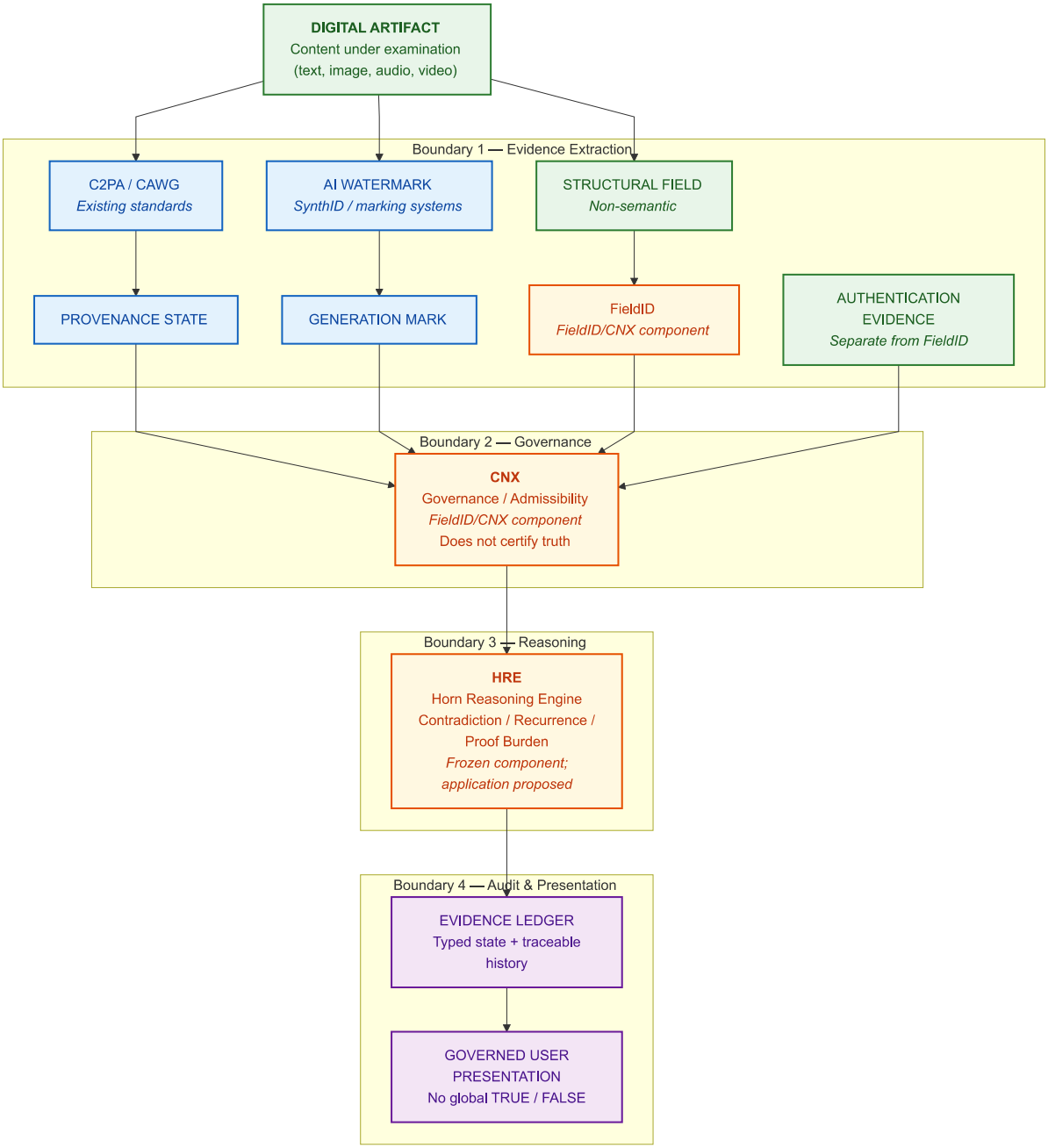

Figure 1. Governed Composition of Synthetic-Content Evidence. Illustrative architecture. The figure represents the proposed composition of existing provenance mechanisms with frozen FieldID/CNX components. It does not represent a validated production deployment. See Appendix C for the complete construction specification.

The central rule remains: local verification does not automatically produce global authority.

A valid watermark result remains watermark evidence. A valid Content Credential remains provenance evidence. A FieldID result remains structural-identity evidence. A valid credential remains authentication or actor-identity evidence. None becomes factual truth merely because it validates locally.

16. Evidence Should Remain Typed

A future implementation should not return one global confidence score. A single scalar risks collapsing different types of certainty.

The evidence vocabulary proposed here includes PROVENANCE VERIFIED, ACTOR VERIFIED, AI MARK DETECTED, STRUCTURAL RELATION DETECTED, TRANSFORMATION RECORDED, PROVENANCE

RECOVERED, EVIDENCE CONFLICT, AUTHORITY INVALID, UNVERIFIED, and EPISTEMIC REVIEW REQUIRED.

This vocabulary intentionally contains no global TRUE or FALSE state.

16.1 Illustrative Composition Example - Claim class: ILL

The following example is illustrative only.

Suppose a published video contains a valid C2PA provenance chain and a recognized AI-generation watermark, but the organizational credential associated with publication has subsequently been revoked.

The evidence state could therefore contain PROVENANCE VERIFIED, AI MARK DETECTED, and AUTHORITY INVALID. If the video's substantive claims also require factual examination, the system could separately retain EPISTEMIC REVIEW REQUIRED.

These observations compose without collapsing. The provenance history may be valid. The AI-generation signal may also be valid. The authority state may no longer be acceptable. The factual content may remain unresolved.

The system therefore does not need to choose between TRUE and FALSE merely to represent what it knows. That is not an omission in the architecture. It is the reason the evidence remains typed.

17. Application to Fake News

The term fake news combines several different failure classes. A useful architecture should separate them.

17.1 Fabricated Media A synthetic video is presented as footage of a real event. Watermarking may provide generation evidence. C2PA may provide provenance. FieldID may potentially provide structural relation evidence. None alone establishes the truth of the event being represented.

17.2 Publisher Impersonation A fabricated statement is attributed to a government agency, company, journalist, or individual. Actor identity and cryptographic provenance become more important than merely detecting AI generation. A human can impersonate a corporation without using AI.

17.3 Provenance Laundering An artifact receives a new apparently legitimate provenance chain that obscures an earlier origin. This is particularly relevant to the Integrity Clash and provenance-substitution problem [10, 16].

17.4 Selective Editing Authentic media is modified while retaining enough original material to remain persuasive. The relevant question becomes not simply whether the artifact is AI-generated, but what transformations occurred and whether the present claim remains supported.

17.5 Authentic Source, False Statement A verified organization publishes incorrect information. No provenance technology solves this problem. The artifact can be perfectly authentic and factually wrong. That is why epistemic verification must remain distinct.

17.6 Misclassification of Governed AI-Assisted Authorship

The inverse error also matters: authentic human-directed work can contain substantial AI participation and still be mischaracterized if generation evidence is promoted into a claim about originating authorship.

A paper produced through Writers' Loop Engineering [21] can define its architecture, claims, evidence boundaries, and acceptance conditions before prose rendering. AI tools can participate heavily in drafting. The same research program can include implementation code produced with AI assistance from author-directed specifications, and specifications whose language was improved with AI assistance. A detector may therefore correctly observe AI participation at several surfaces.

The false positive appears only if that locally valid result is interpreted to mean that the human-originating structure or authority did not exist.

The seed/typing proportion problem explains one mechanism. A compact originating architecture can be surrounded by a much larger rendered artifact. The legal-document-production analogy illustrates the ratio: important structure can occupy a small part of the total material. In governed authorship, the volume need not be adversarial, but the surface detector still sees the volume more readily than the originating decision.

Recognition adds another mechanism. A culturally recognized seed such as E = mc^2 retains an external attribution channel even if AI renders a large exposition around it. An unknown seed does not. If a system uses AI-generation evidence as a

proxy for authorship, the same underlying relationship can therefore receive different attribution outcomes simply because one seed is famous and the other is not.

That is not a demonstrated property of all provenance infrastructure. It is an illustrative failure mode. In that failure mode, the system is no longer functioning as a provenance architecture. It is functioning as a fame detector.

The author reports a prior internal observation providing a hypothesized counterpoint. In a prior internal test, seedless AI-regenerated derivative-content configurations appeared to produce rapid structural-field collapse. However, the historical experimental custody for that observation was not recovered. The observation is classified as a prior author observation and HYP, not as VAL. It motivates controlled comparative testing but does not itself constitute a validated result.

What has been validated is the FieldID structural-identity validity mechanism: a reproduced test suite of 119 passed and 0 failed confirming bounded evaluation through stability, curvature, and detuning [23]. What it observes under controlled seeded-versus-seedless conditions remains to be tested.

The complementary proposition remains HYP. FieldID has not yet demonstrated that a documented human-authored seed remains structurally persistent through AI-assisted rendering, nor that any such persistence can be detected independently of cultural recognition.

The complementary proposition remains HYP. FieldID has not yet demonstrated that a documented human-authored seed remains structurally persistent through AI-assisted rendering, nor that any such persistence can be detected independently of cultural recognition.

The validation target is therefore not to prove that watermark detection is impossible or technically invalid. It is to determine whether structural evidence can distinguish seed-present rendering from seedless generation while preserving a separate typed state for AI participation.

This distinction has practical consequences. A journalist using AI-assisted editing under a governed editorial process is not equivalent, in originating structure or authority, to an autonomous system generating derivative articles at scale. A researcher who originates an architecture and uses AI to render specifications, implementation, and prose is not equivalent, in authorship structure, to a model generating derivative content without that documented seed.

A mature synthetic-content architecture should be able to represent those differences without denying AI participation and without allowing cultural fame to substitute for evidence.

18. Application to AI Abuse More Broadly

The same architecture applies beyond political misinformation. Potential domains include executive impersonation, financial misinformation, fraudulent corporate communications, cloned-voice scams, fabricated evidence, medical misinformation, synthetic product endorsements, academic provenance, manipulated journalism, automated propaganda, and large-scale synthetic spam.

NIST similarly treats provenance tracking, watermarking, synthetic-content detection, testing, and auditing as related but distinct intervention classes [13]. This reinforces the conclusion that synthetic-content assurance is an ecosystem problem rather than a single-detector problem.

19. The Negative-Space Problem

An important unresolved problem is what to do when provenance evidence is absent.

Suppose there is no watermark, no C2PA manifest, no known actor credential, and no registered provenance history. That does not tell us why the evidence is absent.

Possibilities include genuinely human-authored content, unsupported AI generation, stripped metadata, removed watermark, legacy content, anonymous publication, unsupported file format, detector failure, adversarial transformation, or deliberately destroyed provenance.

Therefore absence of evidence should become an evidence state, not a verdict.

This principle is essential for avoiding a system in which unsigned or anonymous content is automatically treated as false.

20. Recoverability After Trust Failure

Digital provenance also changes over time. Keys are compromised. Certificates expire. Organizations change control. Devices are stolen. Algorithms weaken. Credential issuers fail. Trust relationships change.

SCITT explicitly addresses key rotation, key compromise, new credentials, historical statements, append-only records, and independent audit [14]. It also recognizes that transparency does not prevent dishonest or compromised issuers. It makes their actions more accountable and observable.

This aligns closely with a broader CNX principle: do not assume that a security assumption will remain unbreakable. Preserve the ability to recover authority when the assumption fails.

For synthetic-content provenance, this means a mature architecture should eventually represent valid when issued, subsequently compromised, revoked, migrated, and re-established under new authority. A permanently binary credential state cannot fully represent this history.

21. Privacy and the Risk of Turning Provenance Into Surveillance

A provenance infrastructure can itself become dangerous. Universal provenance must not silently become universal identification.

Journalists, whistleblowers, abuse victims, political dissidents, confidential sources, and ordinary citizens may have legitimate reasons to publish authentic material without revealing civil identity. CAWG's identity framework recognizes use cases in which identifying information may need to be withheld or redacted [12].

Therefore FieldID should preserve another important separation: identity continuity does not require public civil identity.

A system may determine same bounded source relationship without publishing the legal identity of that source. This distinction should be treated as a requirement, not an optional privacy enhancement.

22. Adversarial Research Must Remain Part of the Architecture

The current provenance ecosystem should not be treated as final.

A 2026 independent preprint by Golaszewski et al. presents a broad security critique of C2PA and argues that the current specification does not yet satisfy all security properties required for high-stakes deployment [17]. Because this is independent preprint research and because C2PA continues to evolve, it is treated here as red-team evidence rather than settled consensus.

Similarly, a July 2026 preprint by Tamim and Khan evaluates selected text-watermark implementations under a forensic-readiness framework and reports substantial weaknesses under paraphrasing in the tested configurations [18]. Those findings should not be generalized to every deployed watermark, but they reinforce the need to test provenance mechanisms under the actual evidentiary conditions in which they may later be used.

The architectural lesson is broader: a provenance system should assume that its own mechanisms will be attacked.

23. Required FieldID Validation Program

Before FieldID/CNX could responsibly be presented as an Internet synthetic-content assurance system, a dedicated validation program would be required.

23.1 Text: light editing, heavy paraphrasing, translation, summarization, expansion, style transfer, mixed human and AI authorship, and copy-paste transformations.

23.2 Image: recompression, scaling, cropping, screenshotting, filtering, generative regeneration, partial inpainting, and compositing.

23.3 Audio: transcoding, compression, noise, time stretching, pitch changes, speech-to-speech transformation, and partial editing.

23.4 Video: transcoding, frame-rate conversion, cropping, subtitling, partial generative replacement, soundtrack replacement, and screen capture.

23.5 Provenance Attacks: metadata stripping, C2PA manifest substitution, watermark removal, watermark spoofing, credential expiration, credential revocation, compromised signer, conflicting recovered and embedded provenance, and unsupported provenance systems.

23.6 Epistemic Controls: repeated false claims, unique true claims, authenticated false statements, anonymous genuine evidence, synthetic but authorized content, and human-generated impersonation.

Seed-present versus seedless AI-assisted content must be included as a dedicated HYP test. The benchmark should use matched conditions that separate originating seed, renderer, and recognition: (1) culturally recognized human seed with AI rendering; (2) experimentally novel or culturally unrecognized human seed with AI rendering; (3) seedless AI derivative generation; (4) AI-generated candidate seed followed by AI rendering; (5) human seed with human rendering; and (6) human seed with AI rendering under the same downstream task.

The FieldID evaluator for this experiment must not receive author names, citations, semantic labels, fame indicators, or external cultural-knowledge lookups. The purpose is to test whether structural evidence distinguishes seed state rather than whether the system can recognize a famous equation, framework, or person. All seed provenance, rendering steps, transformations, model configurations, and evaluation conditions should be preserved for reproducibility.

24. Required Metrics

The validation program should report more than detector accuracy.

Relevant metrics include false-positive rate, false-negative rate, structural identity stability, provenance recovery rate, watermark recovery rate, evidence-conflict detection rate, transformation tolerance, credential-revocation handling, uncertainty calibration, latency, computational overhead, storage overhead, privacy exposure, reproducibility, and cross-platform portability.

For the seed-persistence and recognition-independence hypothesis, the validation program should additionally report: seed-state discrimination rate; seed-persistence stability across allowed transformations; collapse-to-persistence separation between matched seeded and seedless conditions; recognition-independence differential between culturally recognized and experimentally novel seeds; seed-origin discrimination without author identity or semantic lookup; and transformation retention under editing, paraphrasing, translation, compression, regeneration, or other benchmarked transformations.

These metrics are attached to HYP claims. They do not presume that seed persistence or recognition-independent discrimination exists.

A system that reports high detection accuracy but silently converts uncertainty into authority should not be considered successful.

25. What Would Falsify the FieldID Proposition?

The proposed FieldID role should be falsifiable.

The proposition would weaken substantially if experimentation showed that: FieldID structural identity provides no useful information beyond existing C2PA soft bindings and conventional fingerprinting; structural identity becomes unstable under ordinary Internet transformations; robustness improvements require semantic inference inside FWI, violating its frozen architecture; cross-evidence composition produces unacceptable false associations; the proposed evidence layer introduces excessive latency or storage cost; privacy-preserving identity continuity cannot be achieved; or existing open standards already provide equivalent composition functionality with less complexity.

The seed-persistence proposition would also be falsified or substantially narrowed if FieldID continued to detect collapse in the bounded seedless condition but failed to detect reproducible structural persistence in the documented seed-present condition. In that case, the structural distinction would be observable only at one tested end of the proposed spectrum and would not provide the discriminatory power required for the authorship application.

Recognition independence would likewise fail if apparent seed discrimination depended materially on cultural familiarity, semantic entity recognition, author identity, citations, or external lookup rather than on the structural field itself.

These are not implementation inconveniences. They are legitimate rejection conditions.

26. Why FieldID Should Not Replace C2PA

C2PA already has something FieldID does not currently have: an international standards ecosystem.

It has deployed integrations, manifest formats, cryptographic bindings, transformation history, soft-binding recovery, interoperable verification, and an expanding identity ecosystem [9-12]. SynthID similarly has production deployment and peer-reviewed research [3, 4].

Trying to replace those systems would create an unnecessary standards competition.

A better architectural strategy is to consume external provenance as evidence without becoming dependent on any one provenance technology. If SynthID changes, the adapter changes. If C2PA evolves, the adapter changes. If another jurisdiction mandates another provenance mechanism, that becomes another evidence source. The structural identity and governance boundaries remain stable.

This is consistent with a model-agnostic design.

27. The Larger Architectural Insight

The industry began with: How do we detect AI content?

The next stage is: How do we establish provenance?

But even that is not the final problem.

The deeper problem is: How do we preserve trustworthy evidence when content, actors, credentials, platforms, transformations, models, and security assumptions continuously change?

That is an authority problem as much as a detection problem.

Watermarking helps answer whether a particular generation signal remains present. C2PA helps answer how an artifact says it was created and transformed. CAWG helps associate actors with claims and actions. SCITT provides auditable transparency for signed statements. FieldID may provide an independent structural-identity surface. CNX can govern whether any of those signals receive authority. HRE can preserve contradiction and proof burden rather than collapsing them prematurely.

The important unit therefore becomes not the watermark, but the evidence relationship.

28. Implications for Internet Trust

A mature Internet provenance system should make several states easier to distinguish: known origin and known authority; known origin and unknown authority; known origin and revoked authority; AI-generated and authorized use; AI-generated and unknown use; human-generated and verified publisher; human-generated and unverified publisher; provenance missing; provenance contradictory; artifact transformed; and claim factually unresolved.

This is far more informative than REAL or FAKE. It is also more compatible with an open society.

A trust architecture should improve the visibility of evidence. It should not claim ownership of truth.

29. Limits

This paper has several important limitations.

First, the provider landscape changes rapidly. Statements about deployed watermarking capabilities are current to August 12, 2026. Second, vendor documentation establishes provider-described behavior, not independent measurement. Third, watermark robustness varies by implementation, modality, length, transformation, detector, and adversarial capability. Fourth, C2PA is evolving, and criticisms of one specification version may not remain applicable to later versions. Fifth, CAWG and SCITT continue to evolve. Sixth, the FieldID/CNX Internet assurance profile described here has not been implemented or validated as a complete system. Seventh, internal FieldID and BSM/FWI tests demonstrate bounded implementation properties, not Internet-scale forensic validity. The FieldID structural-identity validity mechanism has been reproduced internally (119 passed, 0 failed) and is classified VAL within the tested configuration [23]. The reproduction produced 13 NumPy RuntimeWarnings during correlation normalization; none caused test failure, and this paper does not infer that the warnings are harmless beyond the reproduced test verdict. The seedless AI field-collapse proposition is a prior

author observation classified HYP; historical experimental custody was not recovered, and the proposition requires controlled comparative testing. Seed persistence and recognition independence remain untested hypotheses. Eighth, structural recurrence is not semantic truth. Ninth, provenance is not ownership. Tenth, a technically authentic source can publish false information.

These limitations constrain the paper's conclusion. They do not eliminate it.

30. Conclusion

AI watermarking is becoming infrastructure. That development is useful and necessary. It is not sufficient.

As watermarking, Content Credentials, actor identity, transparency services, and regulatory requirements mature, the central technical problem shifts from detecting isolated signals to composing evidence without confusing evidence with authority.

The current standards ecosystem already provides much of the machinery required for provenance. C2PA provides cryptographic provenance and durable recovery mechanisms. CAWG extends actor identity. SynthID demonstrates production-scale embedded marking. SCITT develops transparent, replayable histories of signed statements.

The unresolved space appears increasingly concentrated in composition. What happens when evidence is missing? What happens when two valid systems disagree? What happens when a signer is compromised? What happens when an artifact changes but remains recognizably related? What happens when the provenance is authentic but the claim is false? What happens when an anonymous artifact is genuine?

These questions are where FieldID/CNX becomes relevant.

FieldID should not become another watermark. It should not become another truth detector. It should not replace C2PA.

Its potentially valuable role is narrower and deeper: FieldID can be investigated as an independent structural-identity substrate inside a governed evidence architecture in which provenance, watermarking, authentication, actor identity, recurrence, and structural relation remain separate observations until their relationships are evaluated.

CNX contributes the corresponding authority principle: no evidence source should acquire authority merely because it is intelligent, cryptographically valid, statistically confident, structurally persistent, or widely repeated.

The resulting proposition is not that fake news can be eliminated. It is that a better Internet trust architecture may make impersonation, provenance laundering, contradictory origin claims, credential failure, and unresolved uncertainty more visible and more difficult to hide.

That is a narrower claim. It is also the more important one.

31. Declarations

31.1 Author Responsibility The author directed the research question, architectural interpretation, claim boundaries, synthesis, and final editorial position. For this work, the author identifies the originating architectural seed as human-authored: the decision to frame synthetic-content trust through separation between identity, provenance, authentication, authority, and truth, and to investigate whether those distinctions can be structurally preserved.

The FieldID RSP V1.7 frozen architecture [19], the FWI v1.1.0 frozen Internet-ingestion specification [20], and related frozen artifacts were produced through an author-directed process of architectural design, specification, implementation, testing, revision, and freezing. Draft Request Specification Protocols were authored under the author's direction. AI tools may have assisted with language refinement of specifications and may have participated in implementation-code generation from those specifications. AI participation in those activities is acknowledged rather than erased.

The author also identifies original formulations in physics and mathematics developed in the broader research program as author-originated intellectual seeds. Some remain hypothesis propositions. Their scientific validity is outside the scope of this paper and is independent of their use here as illustrations of originating intellectual structure.

The relevant authority boundary is that AI assistance did not hold final authority over the architectural decisions, boundary definitions, acceptance decisions, or freezing of the cited artifacts. AI-assisted drafting tools were also used in development of this manuscript through Writers' Loop Engineering [21], in which manuscript structure, claims, evidence boundaries, continuity requirements, and acceptance criteria were defined before final prose was accepted. Final publication authority remains with the author.

31.2 Competing Interests The author is the developer of FieldID, CNX, and related Carlonoscopen research architectures discussed in this paper. These systems should therefore be treated as author-developed technology rather than independent third-party evidence.

31.3 Evidence Boundary External standards and research are cited independently. Internal implementation results are identified separately from externally validated findings. Proposed extensions are not presented as implemented capability.

31.4 Data and Code No new public benchmark dataset was generated for this paper. Internal FieldID/CNX implementation artifacts remain subject to their existing custody and publication status.

31.5 Archival Status The Zenodo DOI 10.5281/zenodo.21909033 has been reserved for this manuscript. Zenodo archival deposit provides a citable, versioned record. It does not itself constitute external peer review.

Appendix A. Claim and Evidence Boundary

The paper can be reduced to five bounded statements.

A.1 Externally Supported: Watermarking, cryptographic provenance, actor identity, and transparency systems are becoming increasingly capable and increasingly deployed.

A.2 Externally Supported: None of those mechanisms independently establishes factual truth.

A.3 Internally Validated: FieldID preserves a separation between structural identity and authentication, and FWI preserves a separation between structural Internet ingestion and semantic authority.

A.4 Proposed: Those separations may be useful in an architecture that composes provenance evidence from multiple independent systems.

A.5 Not Yet Demonstrated: FieldID/CNX has not yet been shown to outperform existing provenance systems or provide Internet-scale misinformation assurance.

No conclusion in the paper should be interpreted more strongly than these five statements.

Appendix A verification for v1.2: All five bounded statements remain accurate. The FieldID structural-identity validity mechanism is VAL within the reproduced test configuration [23]. The seedless AI field-collapse proposition is HYP (prior author observation, historical custody not recovered). Seed persistence and recognition independence remain HYP. No claim-class escalation is authorized. The VAL mechanism result is bound to the formal custody record by the reproduction-log and evidence-record SHA-256 values.

Appendix B. Evidence Presentation Contract

A future public verifier based on this architecture should follow four rules.

Rule 1: UNVERIFIED is not FALSE. Rule 2: AI-GENERATED is not MALICIOUS. Rule 3: AUTHENTIC SOURCE is not FACTUALLY TRUE. Rule 4: VALID LOCAL EVIDENCE is not GLOBAL AUTHORITY.

These rules summarize the paper's central architecture.

Appendix B verification for v1.2: All four evidence presentation rules remain intact. The recognition-asymmetry analysis reinforces Rule 4 by separating locally valid AI-generation evidence from authority over originating authorship.

Appendix C. Figure 1 Construction Specification

The specification below documents the construction of Figure 1. The author-supplied vector graphic is embedded in this frozen publication version and preserved as a separate vector source. This textual specification is retained so the figure's nodes, flows, evidence classes, and authority boundaries remain independently documented even if the graphic representation changes.

C.1 Figure Title Figure 1. Governed Composition of Synthetic-Content Evidence

C.2 Figure Status Claim class: ILL. The figure is an architectural proposal and explanatory device. It is not evidence of a validated production deployment.

C.3 Overall Layout Construct the figure as a top-to-bottom directed flow diagram. The top layer contains the observed digital artifact. The next layer separates evidence into three parallel paths. Those paths converge at the CNX governance layer. A separate authentication-evidence input also enters CNX. CNX passes admitted evidence to HRE. HRE outputs the structured evidence state to the evidence ledger. The evidence ledger supports the final governed user presentation.

C.4 Nodes Digital Artifact: content under examination, such as text, image, audio, video, or composite digital object. Outgoing to C2PA/CAWG, AI Watermark, and Structural Field.

C2PA/CAWG: existing standards ecosystem for provenance and actor-identity evidence. Outgoing to Provenance State.

Provenance State: standards-based provenance evaluation before governance. Outgoing to CNX.

AI Watermark: external generation evidence, including SynthID or supported marking systems. Outgoing to Generation Mark.

Generation Mark: watermark or AI-generation evidence before governance. Outgoing to CNX.

Structural Field: non-semantic structural representation of the observed artifact. Outgoing to FieldID.

FieldID: FieldID/CNX component producing bounded structural-identity representation. Outgoing to CNX.

Authentication Evidence: independent authentication or trust evidence associated with source, actor, device, or authority context. Authentication remains separate from FieldID structural identity. Outgoing to CNX.

CNX: governance/admissibility layer. Incoming from Provenance State, Generation Mark, FieldID, and Authentication Evidence. Outgoing to HRE. CNX does not transform a valid local signal into factual truth.

HRE: Horn Reasoning Engine. Maintains relationships among admitted evidence, including contradiction, recurrence, and unresolved proof burden. Existing frozen reasoning component; application here is proposed. Outgoing to Evidence Ledger.

Evidence Ledger: preserves the resulting typed evidence state and traceable history. Outgoing to Governed User Presentation.

Governed User Presentation: presents typed evidence and unresolved states without silently converting them into a global TRUE/FALSE judgment.

C.5 Required Edges DIGITAL ARTIFACT -> C2PA / CAWG -> PROVENANCE STATE -> CNX DIGITAL ARTIFACT -> AI WATERMARK -> GENERATION MARK -> CNX DIGITAL ARTIFACT -> STRUCTURAL FIELD -> FieldID -> CNX AUTHENTICATION EVIDENCE -> CNX -> HRE -> EVIDENCE LEDGER -> GOVERNED USER PRESENTATION

No additional architectural edges should be introduced.

C.6 Authority Boundaries Boundary 1, Evidence Extraction: C2PA/CAWG, Provenance State, AI Watermark, Generation Mark, Structural Field, FieldID, Authentication Evidence. These elements produce or transport evidence. None receives final epistemic authority merely by producing a valid result.

Boundary 2, Governance: CNX. CNX governs admissibility and authority conditions. It does not certify factual truth.

Boundary 3, Reasoning: HRE. HRE evaluates relationships among admitted evidence and preserves contradiction and proof burden. It does not replace CNX's governance role.

Boundary 4, Audit and Presentation: Evidence Ledger and Governed User Presentation. The ledger preserves evidence state. Presentation communicates that state without collapsing uncertainty into an unsupported verdict.

C.7 Visual Classification Annotations Existing standards and external systems: C2PA, CAWG, and AI watermark systems such as SynthID. FieldID/CNX components: FieldID, CNX, HRE. Proposed composition elements: combined evidence flow, Evidence Ledger in this provenance profile, and Governed User Presentation in this provenance profile.

C.8 Required Caption Figure 1. Governed Composition of Synthetic-Content Evidence. Illustrative architecture. The figure represents the proposed composition of existing provenance mechanisms with frozen FieldID/CNX components. It does not represent a validated production deployment.

Temporal Context of Sources

This manuscript reflects the state of standards, provider deployments, regulation, and cited research available to the author as of August 12, 2026. C2PA 2.4, the cited CAWG draft material, the cited SCITT Internet-Draft, and current vendor implementations are versioned and evolving systems and may have been superseded when this paper is read.

The 2026 preprints cited as adversarial or red-team evidence are included according to their status at the time of writing. They should not be interpreted as settled scientific consensus or as statements about specification versions published after the manuscript date.

References

1. European Commission. Code of Practice on Transparency of AI-Generated Content. 2026. Article 50 transparency obligations applicable from August 2, 2026. https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content
2. European Commission. Guidelines on transparency obligations for providers and deployers of AI systems. 2026. https://digital-strategy.ec.europa.eu/en/policies/guidelines-transparency-ai-generated-content
3. Google DeepMind. SynthID. Current technical and product documentation as of August 12, 2026. https://deepmind.google/models/synthid/
4. Dathathri, S., et al. Scalable watermarking for identifying large language model outputs. Nature 634, 2024. https://www.nature.com/articles/s41586-024-08025-4
5. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A Watermark for Large Language Models. Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 2023, pp. 17061-17084. https://proceedings.mlr.press/v202/kirchenbauer23a.html
6. Kirchenbauer, J., et al. On the Reliability of Watermarks for Large Language Models. International Conference on Learning Representations, 2024. https://openreview.net/forum?id=DEJIDCmWOz
7. OpenAI. Advancing content provenance for a safer, more transparent AI ecosystem. May 19, 2026; updated July 31, 2026. https://openai.com/index/advancing-content-provenance/
8. Anthropic. How Claude marks AI-generated content. Current documentation as of August 2026. https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content
9. Coalition for Content Provenance and Authenticity. C2PA Technical Specification, Version 2.4. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html
10. Coalition for Content Provenance and Authenticity. C2PA Soft Binding API and implementation guidance, Version 2.4. https://spec.c2pa.org/specifications/specifications/2.4/softbinding/Decoupled.html
11. Coalition for Content Provenance and Authenticity. C2PA and Content Credentials Explainer, Version 2.4. https://spec.c2pa.org/specifications/specifications/2.4/explainer/Explainer.html
12. Creator Assertions Working Group. Identity Assertion, current draft specification as of August 12, 2026. https://cawg.io/identity/1.3-draft%2Bvlei/
13. Chandra, B., et al. Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. NIST AI 100-4. National Institute of Standards and Technology, 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf
14. IETF SCITT Working Group. An Architecture for Trustworthy and Transparent Digital Supply Chains. Internet-Draft version 22 as cited in this manuscript. https://datatracker.ietf.org/doc/html/draft-ietf-scitt-architecture-22
15. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can AI-Generated Text be Reliably Detected? arXiv:2303.11156, 2023. https://arxiv.org/abs/2303.11156
16. Nemecek, A., He, H., Cheng, G., and Ayday, E. Authenticated Contradictions from Desynchronized Provenance and Watermarking. CVPR Workshops, 2026. arXiv:2603.02378. https://arxiv.org/abs/2603.02378
17. Golaszewski, E., et al. Verifying Provenance of Digital Media: Why the C2PA Specifications Fall Short. arXiv:2604.24890, 2026. Preprint used as independent red-team evidence. https://arxiv.org/abs/2604.24890
18. Tamim, S. R., and Khan, A. L. AI Watermark Evidence Fails Forensic Readiness: An Empirical Evaluation. arXiv:2607.16010, 2026. Preprint used as time-bounded red-team evidence. https://arxiv.org/abs/2607.16010
19. Silva, I. FieldID RSP V1.7, System-Integrated Edition. Internal frozen architecture baseline. Carlonoscopen, LLC.
20. Silva, I. CNX FieldID Web Ingestion Module, FWI v1.1.0. Internal frozen Internet-ingestion specification. Carlonoscopen, LLC.
21. Silva, I. Writers' Loop Engineering: Precompiled Narrative Architecture for Governed AI-Assisted Authorship. Carlonoscopen Journal of Coherence Intelligence, 2026.
22. Zenodo. DOI reservation record for 10.5281/zenodo.21909033. 2026.

23. Silva, I. FieldID Structural Validity Mechanism Evidence Record, v1.0. Internal reproduced evidence record. Carlonoscopen, LLC, 2026. Reproduction log SHA-256: 13ADAC1BC211905C0E3F58FC427CCA7C61BF351736F0329F7CBAE3A3BE3A7762. Evidence-record JSON SHA-256: 1c51a6270fbd3e6de1199103fd2b451c2ad5df9c26585dd26d80c81d42ddcd59.

Suggested Citation

Silva, Ivan. "AI Watermarking, Digital Provenance, and FieldID: An Industry and Technical Assessment of Synthetic-Content Trust. From Content Marking to Governed Evidence Composition." Carlonoscopen Journal of Coherence Intelligence, Vol. 1, No. 23, 2026. Version 1.2. Reserved DOI: 10.5281/zenodo.21909033.

v1.2 Revision and Publication Record

On August 12, 2026, the author requested assembly of the initial v1.0 publication package for CJCI Volume 1, Issue 23 using the Writers' Loop Engineering publication discipline [21] and the v1i22 release precedent. The resulting v1.0 ZIP is preserved as the immutable baseline for this revision lineage.

The author subsequently initiated the v1.1 revision to incorporate the seed/typing proportion analysis, recognition asymmetry, the FieldID structural-validity mechanism evidence, and explicit seed-persistence and recognition-independence validation and falsification conditions.

The FieldID structural-identity validity mechanism evidence gate has been closed. The mechanism has been reproduced internally: 119 tests passed, 0 failed, 13 warnings (NumPy RuntimeWarnings during correlation normalization, none causing failure). Formal custody artifacts have been created, including reproduction log SHA-256 13ADAC1BC211905C0E3F58FC427CCA7C61BF351736F0329F7CBAE3A3BE3A7762 and evidence-record JSON SHA-256 1c51a6270fbd3e6de1199103fd2b451c2ad5df9c26585dd26d80c81d42ddcd59 [23].

The seedless AI field-collapse proposition was reclassified from VAL to HYP following the author's determination that historical experimental custody for the prior observation was not recovered. The mechanism (VAL) and the application (HYP) remain distinguished throughout this frozen v1.2 publication.

This frozen v1.2 publication does not overwrite or silently replace the v1.0 or v1.1 packages. Zenodo activation remains an external author-controlled action. Figure 1 has been supplied by the author, verified against Appendix C, embedded in the manuscript, and preserved as a separate vector source. Appendix C preserves the complete construction specification.

On August 12, 2026, the human author approved v1.2 for final freeze with one authorized editorial adjustment: References 1-23 are formatted as separate entries for publication readability. The VAL mechanism result remains bound to the formal custody record, mechanism VAL and application HYP remain distinguished in every passage, and the Figure 1 production and embedding gate remains complete. No substantive claim expansion was authorized by the final-freeze adjustment.

Revision status: V1_2_FINAL_FROZEN Publication freeze: AUTHOR_APPROVED_FINAL_FROZEN Zenodo activation: EXTERNAL_AUTHOR_ACTION_NOT_EXECUTED DOI: 10.5281/zenodo.21909033 (reserved) CJCI: Volume 1, Issue 23