WP-021 | Synthetic Content Provenance | June 2026

Beyond C2PA: Zero-Knowledge Content Provenance That Survives Metadata Stripping

AffixIO Research June 2026 WP-021 Download PDF

C2PA was meant to label synthetic media. Most uploads strip the credentials anyway. We bind provenance to content hashes with zero-knowledge proofs so labels survive compression, re-encoding, and deliberate metadata removal. Platforms verify origin without seeing your generation pipeline.

CONTENTS

1. The Synthetic Content Crisis

Generative AI has produced an unprecedented expansion in synthetic content: images, video, audio, and text created entirely by AI systems without direct human authorship. By 2026, the majority of content uploaded to major platforms on some media categories is estimated to have significant AI involvement in its creation. The ability of ordinary users, journalists, fact-checkers, and automated systems to distinguish authentic content from synthetic or manipulated content has become a critical infrastructure problem for democratic discourse, public safety, and commercial integrity.

The problem is not merely volume. Synthetic content quality has reached the point where visual or audio inspection cannot reliably distinguish AI-generated content from authentic content. High-quality deepfake video of public figures, synthetic audio of political candidates, and AI-generated news articles are regularly circulated as authentic. The harms extend from election interference to financial fraud to non-consensual intimate imagery.

Technical provenance solutions, systems that label content at the point of generation and verify that label at the point of consumption, are the primary proposed mechanism for addressing the synthetic content crisis. C2PA is the leading implementation of this approach. But C2PA's deployment at scale has revealed a critical gap: the provenance information exists at generation but routinely disappears before consumption, precisely the distribution pathway where the labelling is most needed.

---

2. EU AI Act Article 50 and DSA Obligations

EU AI Act Article 50 establishes transparency obligations for AI-generated content that become enforceable from 2 August 2026. The Article requires that providers of AI systems that generate synthetic content ensure outputs are marked in a machine-readable format and recognisable as artificially generated or manipulated, to the extent technically feasible. The obligation applies to video, audio, image, and text content generated by AI systems, with limited exceptions for systems used for

legitimate security testing, authorised public interest purposes, and content that has been authorised by the natural person depicted.

The Digital Services Act (DSA) adds complementary obligations for very large online platforms (VLOPs) and very large online search engines (VLOSEs): they must implement risk mitigation measures for systemic risks including the dissemination of illegal content and negative effects on civic discourse, which encompasses undetected synthetic content. The Online Safety Act in the UK imposes similar obligations on regulated services with respect to AI-generated content that is illegal or harmful.

Both the EU AI Act and the DSA specify that AI-generated content must be labelled in a machine-readable format. This is the specification that C2PA satisfies through its Content Credentials metadata standard. The gap that C2PA does not address is what happens when that machine-readable metadata is removed before the content reaches the platform's detection systems or the end consumer.

---

3. C2PA: What It Does and How It Works

The Coalition for Content Provenance and Authenticity (C2PA) is an industry consortium whose members include Adobe, Microsoft, Google, OpenAI, Meta, the BBC, and numerous camera manufacturers and news organisations. The C2PA standard specifies Content Credentials: structured metadata attached to digital content that records the content's provenance, including who created it, what tools were used, what AI systems were involved, and what edits were made.

The C2PA Technical Architecture

A C2PA Content Credential is a signed manifest containing provenance claims. The manifest is cryptographically signed using X.509 certificates, ensuring that the claims are attributable to an identified party and have not been tampered with. The SHA-256 hash of the content is included in the manifest, binding the credential to the specific content at the time of signing. Multiple manifests can be chained, recording the

provenance history of content that has passed through multiple editing or publishing stages.

C2PA 2.2, released in 2025, introduced durable content credentials that address some limitations of the original specification by combining the manifest with invisible watermarking. The watermark survives some processing operations that would strip metadata, and the watermark's presence indicates that credentials existed, even if the credentials themselves have been removed. SynthID, Google DeepMind's watermarking system, operates on similar principles for AI-generated content.

C2PA Adoption

C2PA adoption has been substantial among content creation platforms. OpenAI's image generation products, Adobe Firefly, Google's Imagen, and Meta's AI image tools all include C2PA Content Credentials in their outputs. The EU's Code of Practice for AI Content Marking explicitly cites C2PA as a recommended technology.

---

4. The Metadata Stripping Problem

C2PA credentials live in the file's metadata, specifically in JPEG EXIF data, PNG metadata chunks, or file format-specific containers. This is the fundamental architectural limitation. Digital content routinely passes through transformations that strip metadata:

- Screenshots: capturing a screenshot of an image removes all file metadata; the resulting image file contains only pixel data
- Social media uploads: most social media platforms strip or rewrite metadata during upload processing; a C2PA-credentialled image uploaded to Twitter, Instagram, or TikTok will have its credentials removed
- Format conversion: converting a JPEG to a PNG, compressing a video to a different codec, or re-encoding audio strips the original metadata
- Deliberate stripping: tools that remove metadata from files, commonly used for privacy reasons, also remove C2PA credentials

The Credential Absence Problem

Absence of C2PA credentials proves nothing. Content that predates C2PA, content that was processed by a non-C2PA-supporting pipeline, and content that had its credentials deliberately stripped are all indistinguishable from content that never had credentials. Missing credentials cannot be used as evidence of inauthenticity; they can only be used as evidence that the content was not produced or last processed by a C2PA-supporting pipeline.

---

5. Watermarking as a Complement, Not a Solution

Imperceptible watermarking, as implemented by SynthID and similar systems, survives some metadata-stripping operations because the watermark is encoded in the content itself, not in metadata. A watermarked image retains its watermark after being screenshotted, re-compressed, or uploaded to a platform that strips metadata, as long as the transformation does not significantly alter the content's pixel values.

Watermarking addresses some of C2PA's weaknesses but introduces new limitations. Watermarks are detectable only by parties who hold the watermark detection key, typically the AI system provider. A regulator, journalist, or independent fact-checker cannot detect a watermark without access to the detection key, creating a centralised trust dependency. Watermarks can be removed by sufficiently adversarial processing: adding noise, performing aggressive compression, or using purpose-built watermark removal tools can reliably defeat current watermarking schemes. And watermarks cannot carry rich provenance information: they signal the presence of AI generation but cannot record the specific model, generation parameters, or editing history.

The appropriate role for watermarking in the content provenance ecosystem is as a signalling mechanism: a watermark's presence indicates that more detailed

provenance information may be available in an associated credential. It is not a standalone provenance solution.

---

6. ZK Content Provenance: Binding Proof to Hash

AffixIO's content provenance architecture addresses the metadata stripping problem by separating the provenance proof from the content file. Rather than storing the provenance credential in the content's metadata, the proof is stored in a publicly accessible content provenance registry indexed by the content's cryptographic hash.

Hash-Bound Provenance

When AI-generated content is produced, AffixIO's system generates a ZK proof asserting: this content has hash H, H was produced by AI system S (committed without identifying S specifically), at timestamp T, under generation parameters P (committed without disclosing specific parameters). The proof is stored in the content provenance registry under the key H.

When a verifier encounters content and wants to check its provenance, they compute the content's hash, query the registry for that hash, and receive the provenance proof if one exists. The proof can be verified without accessing the original content generation system. This mechanism survives metadata stripping because the content's pixel data, which determines its hash, is not affected by metadata removal.

Hash Stability Across Transformations

One complication is that content transformations, such as JPEG re-compression or format conversion, change the content's hash. AffixIO addresses this through a perceptual hash supplement: alongside the cryptographic hash of the original content, the registry also stores a perceptual hash (using industry-standard algorithms such as pHash or dHash) that is stable across minor transformations. A verifier can query by

exact hash for definitive provenance verification or by perceptual hash for approximate verification that tolerates minor processing.

---

7. The AffixIO Content Provenance Architecture

The AffixIO content provenance system comprises three components: the generation-time proof generator, the content provenance registry, and the verification API.

Generation-Time Proof Generator

At content generation time, the AI system invokes the AffixIO proof generator. The proof generator receives the content hash, the generator identity commitment, the generation timestamp, and the generation context (model version, prompt type category, output format). It produces a ZK proof asserting these properties and stores the proof in the registry under the content hash. The proof generator is designed to integrate as a post-processing step in existing AI content generation pipelines.

Content Provenance Registry

The registry is an append-only hash-indexed store. Entries cannot be deleted or modified after insertion, ensuring that provenance records are permanent. The registry supports two query modes: exact hash lookup for content whose hash matches a stored entry, and perceptual hash lookup for content that has been transformed since provenance registration. The registry is designed to scale to billions of entries using distributed hash table storage with O(1) lookup latency.

Verification API

The verification API accepts a content file, computes its exact and perceptual hashes, queries the registry, and returns the provenance proof if one exists. The response includes the proof, the proof's verification key, and a human-readable summary of the provenance claims. The API is available as a REST endpoint compatible with C2PA's

verification flow, enabling platforms already implementing C2PA to add hash-bound provenance verification alongside their existing credential checking.

---

8. Creator Privacy: ZK Selective Disclosure of Identity

Content provenance creates a privacy tension: comprehensive provenance records associate specific content with specific creators, creating a surveillance risk for legitimate uses of AI generation. A journalist using AI to generate supporting visuals, a privacy-conscious artist, or a whistleblower using AI tools should not be obligated to disclose their identity in order for their content to carry provenance information.

AffixIO's ZK content provenance architecture addresses this through ZK selective disclosure of generator identity. The provenance proof commits to the generator's identity without revealing it. The generator can choose to reveal their identity to specific parties, such as a regulatory authority under compulsion, by providing a selective disclosure proof that opens the identity commitment. Parties who do not receive the selective disclosure see only that the content was generated by an identified AI system, with a committed but undisclosed generator identity.

This enables a regulatory framework that is both privacy-respecting and accountable: content provenance is universal, generator identity is private by default, and identity disclosure is available to authorities when legally required. This architecture is directly compatible with the EU AI Act's Article 50 transparency requirements, which focus on machine-readable AI-generation disclosure rather than mandatory generator identity disclosure.

---

9. Platform Integration and Distribution

For the content provenance ecosystem to function, platforms must check provenance at upload time and present provenance information to consumers at consumption

time. AffixIO's architecture is designed to integrate into platform content moderation and labelling pipelines with minimal overhead.

Upload-Time Checking

When content is uploaded to a platform, the platform's content pipeline computes the content hash and queries the AffixIO provenance registry. If a provenance proof exists, the platform labels the content as AI-generated, includes the provenance information in the content's metadata, and presents a provenance indicator to viewers. If no proof exists, the platform may apply additional detection checks, such as AI content classifiers or watermark detection, before deciding on labelling.

Distribution-Stable Labelling

Once a platform has verified and labelled content, the label is maintained in the platform's own content management system, independent of the original content file's metadata. If the content is shared on the platform, the label travels with the platform's content record, not with the file. This addresses a second distribution problem beyond metadata stripping: even if a platform preserves C2PA credentials at upload, the label does not follow the content if it is downloaded and re-uploaded to a different platform. With hash-bound provenance, the receiving platform can independently verify provenance by querying the registry.

---

10. Regulatory Compliance: EU AI Act, DSA, Online Safety Act

AffixIO's ZK content provenance architecture satisfies the synthetic content labelling requirements of the three major regulatory frameworks applicable in 2026.

EU AI Act Article 50

Article 50 requires that AI-generated content be marked in a machine-readable format. The AffixIO provenance proof, stored in a registry queryable by content hash, is a machine-readable format: any automated system can retrieve and verify the provenance proof for any piece of content by computing its hash. The ZK proof

records the content's AI-generated origin, satisfying the Article 50 disclosure requirement without requiring the generator to be identified.

DSA Risk Mitigation

The DSA requires VLOPs to implement risk mitigation measures for systemic risks from synthetic content. Hash-bound provenance checking at upload provides a systematic risk mitigation measure: the platform can identify AI-generated content even after metadata stripping, enabling appropriate labelling and content policy enforcement.

Online Safety Act

The UK Online Safety Act requires regulated services to implement measures to protect users from illegal and harmful content, which includes non-consensual intimate imagery created by AI systems. Hash-bound provenance provides a detection mechanism: the registry records that content was AI-generated at creation, enabling platforms to identify and act on AI-generated harmful content even when it has been processed to remove conventional watermarks or metadata.

---

11. Limitations and Threat Model

No content provenance system provides complete protection against determined adversaries. Understanding the threat model is essential for appropriate deployment.

Content Recreation

An adversary who recreates AI-generated content from scratch, for example by taking a screenshot of displayed content and then manipulating it substantially, produces content with a different hash that will not match any registry entry. Hash-bound provenance is not robust to content recreation; it is robust to the distribution transformations that routinely occur in legitimate content sharing workflows.

Unsigned AI Systems

AI systems that do not integrate with the AffixIO proof generator, or any content provenance system, will produce content with no provenance record. Missing provenance is not evidence of absence of AI generation; it may indicate a non-compliant or non-integrated generation system. Regulatory enforcement action against platforms for hosting AI-generated content without provenance records must account for the possibility that the content was generated by a system outside the provenance ecosystem.

Key Compromise

The security of the provenance proof depends on the integrity of the generator identity commitment. If the signing key used by the AI system's proof generator is compromised, an adversary could generate false provenance records. AffixIO mitigates this through HSM-held generation keys, key rotation policies, and a key transparency ledger that records all active generation keys, enabling detection of unauthorised key additions.

---

12. Conclusions

C2PA represents a meaningful advance in synthetic content provenance infrastructure, and AffixIO's architecture is designed to complement and extend C2PA rather than replace it. The fundamental limitation of metadata-bound credentials is architectural: solving it requires binding provenance to content rather than to metadata. Hash-bound ZK provenance proofs stored in a content registry provide exactly this property.

As EU AI Act Article 50 becomes enforceable in August 2026, content generation platforms, social media platforms, and content management systems all face the practical question of how to satisfy a synthetic content labelling obligation that survives the distribution transformations that routinely occur in content workflows. AffixIO's content provenance architecture provides a technically robust answer: provenance that is cryptographically bound to the content itself, verifiable by any party with access to the registry, privacy-respecting through ZK selective disclosure of

generator identity, and complementary to existing C2PA and watermarking deployments.

Related reading

- WP-018: Cryptographic AI-BOM: ZK Provenance for Model Supply Chains
- WP-005: Source Verification as a Zero-Knowledge Circuit Input
- WP-001: Cryptographic AI Governance: A Technical Framework

Frequently asked questions

Why is C2PA not enough?

Social platforms and CDNs routinely strip or rewrite embedded metadata. Credentials attached to files disappear before users see the content.

How does hash-bound provenance work?

A proof commits to the perceptual and exact hashes of the content. Verification queries a public registry, not fragile file metadata.

Who needs this for compliance?

Publishers, platforms, and model providers facing EU AI Act Article 50 and DSA obligations on synthetic and manipulated media.

All White Papers

▶ About

▶ Solutions

▶ Legal

▶ Trust & Security

Contact

truth layer | yes | no | proof