ENCYPHER

FIELD GUIDE 01: AI watermarking / C2PA

C2PA vs AI watermarking: what each tells you

Four signals, four different questions

An image arrives with an 'AI-generated' label. That label, an invisible watermark, a Content Credential and an AI detector result offer different evidence. Treating them as equivalent can turn a useful clue into an unsupported conclusion.

C2PA 2.4 describes signed provenance: recorded statements about a file's origin and changes, cryptographically associated with the content. It can describe AI and non-AI workflows. It does not inspect pixels and independently decide whether AI made them. [1] Encypher's provenance explainer introduces the signing, embedding and verification workflow. [5]

- Visible label: text or a badge communicates a claim to people. The visible mark alone does not establish cryptographic validity.
- Embedded watermark: a deliberate signal inside content, read by a compatible detector. SynthID embeds imperceptible signals in supported AI outputs. [2]
- Signed provenance: a C2PA manifest records statements and lets a verifier check their integrity, content association and signing credential. [1]
- Post-hoc classifier: a model estimates whether existing content resembles AI output, without requiring a deliberately inserted signal. Such classifications can be wrong. [4]

> *First ask which mechanism produced the result.*

How watermarks and provenance work together

SynthID illustrates AI watermarking. Google DeepMind describes signals embedded in images, audio, video and text; its text method adjusts token probabilities during generation. Detection looks for that system's signal, rather than a universal property of everything made with AI. [2]

A C2PA manifest can supply richer context: recorded creation, editing and source material. C2PA also supports soft bindings, including watermarks and fingerprints, to help discover a credential when embedded metadata is missing. A watermark used for discovery and an AI-origin watermark need not carry the same information. [1]

Illustrative example: a publisher releases an AI illustration with a visible disclosure, a supported generation watermark and signed provenance recording its workflow. A reader can see the disclosure, check for the watermark and inspect the credential. Each answers a different question.

- Robustness depends on the system and transformation. Google warns that repeated alterations can prevent detection. [3]
- A recovered provenance record still needs validation; a lookup match alone is not a valid signature or an exact content match. [1]

> *Complementary evidence is not interchangeable evidence.*

What a missing signal means

No credential does not mean fake. C2PA is optional, metadata can be removed, and applications differ in what they can retrieve and display. No detected watermark does not mean human-made: the generator may use another system, or alterations may defeat detection. [1][3]

Classifier confidence is not proof. OpenAI withdrew its 2023 text classifier for low accuracy; that historical example does not measure every current detector. [4]

- Identify the tool, supported content types and what its result actually covers.
- Inspect available provenance details instead of relying on a badge alone.
- Seek the publisher's original file and corroborate the claim with independent reporting.
- Record unresolved gaps rather than translating 'not found' into 'authentic' or 'AI-free'.

> *None of these signals alone establishes factual truth, legal ownership or permission to reuse the content.*

Sources and further reading

[1] C2PA 2.4 Explainer, sections 2, 6, 7.2 and 7.4

[2] Google DeepMind: SynthID, how it works

[3] Google Gemini Help: verification results and SynthID limitations

[4] OpenAI: AI text classifier, withdrawal notice and limitations (2023)

[5] Encypher: What is content provenance? (explainer)

C2PA 2.4. Sources checked September 11, 2026. App interfaces and support can change.

See a signed record for yourself

Try a non-sensitive example: Encypher signed-text demo

Check format support and upload handling before submitting files. These guides explain concepts; they do not certify content or establish legal compliance.

Content Credentials, provenance and authenticity explained

What the terms actually mean

A photograph can be genuine as a publisher's released file yet carry a misleading caption. Content Credentials help investigate the file's recorded history. They do not settle whether the caption accurately describes the world. [1]

Digital provenance means information about where content came from and what happened to it. A Content Credential is the reader-facing term for a C2PA manifest: statements about an asset, a claim and a digital signature. The C2PA 2.4 standard defines how these fit together and how applications validate them. [2] Encypher's explainer connects these concepts to a publishing workflow. [3]

- An assertion is a recorded statement, for example about a creation or editing action. [2]
- A content binding associates the credential with the content. Cryptographic hashes provide hard bindings; other identifiers can help discover related records. [2]
- Content authenticity asks whether content is what it claims to be. C2PA contributes tamper-evident provenance to that assessment; it does not guarantee factual accuracy. [1][2]

> *A signed record gives evidence to assess. It does not replace the reader's judgment about the source or story.*

Checking a publisher's signed file

Illustrative example: a newsroom publishes a photograph with a Content Credential. You obtain the released file and open it in a compatible verifier. You inspect the signing information, validation results and recorded edits, rather than accepting a logo in the image as proof. [1]

Separate three checks. Signature verification checks the signed claim against the public key. Integrity checks assess the protected assertions and the binding to the content. Trust evaluation checks the signing credential against the verifier's trust policy. A mathematically valid signature alone does not settle the other questions. [2]

The signer can be a software or hardware credential holder. Do not assume its name identifies the photographer. Human or organizational identity features can involve extensions beyond core C2PA. [1]

- Read which edits and source materials were recorded, and where the record starts.
- Distinguish invalid, untrusted, missing and unsupported results; ask what failed before judging the content. Encypher's verification guide explains its separate evidence results. [4]
- Check the publisher's own site to establish context independently of the file.

> *Trusted signing infrastructure does not certify the story's truth.*

A practical reader checklist

Provenance can be incomplete. Work done before signing or in tools that do not preserve credentials may be absent. A valid credential does not prove human origin, an unedited scene or a complete history. Missing credentials alone do not establish deception. [1]

- Source: obtain the publisher's original release when possible; keep its URL and context.
- Status: inspect signature, content integrity and signer trust results in a compatible viewer.
- Scope: read the recorded actions and ingredients, meaning source assets; note gaps and unavailable records.
- Meaning: distinguish the signer from the creator, and recorded AI involvement from a detector's estimate.
- Reality: corroborate names, dates, locations and claims using evidence outside the credential.

> *Use provenance to ask better questions about content. Keep factual verification, ownership and reuse permission as separate inquiries.*

Sources and further reading

[1] C2PA 2.4 Explainer, sections 2, 7.2, 7.3 and 7.5

[2] C2PA 2.4 Technical Specification, glossary, trust model and validation

[3] Encypher: What is content provenance? (explainer)

[4] Encypher: C2PA and watermark verification (implementation guide)

C2PA 2.4. Sources checked September 11, 2026. App interfaces and support can change.

See a signed record for yourself

Try a non-sensitive example: Encypher signed-text demo

Check format support and upload handling before submitting files. These guides explain concepts; they do not certify content or establish legal compliance.

PDF watermarking: visible labels and verifiable provenance

Add a visible label in Acrobat

A PDF watermark is text or an image placed on a page, such as DRAFT or a company logo. It gives readers a visible cue. Adobe distinguishes watermarks, which are embedded in the page, from stamps. Neither a familiar logo nor the word CONFIDENTIAL proves who supplied the document. [1]

Illustrative example: a consultant labels a proposal DRAFT before sending it for review. The label communicates its intended status; it does not stop someone from forwarding the file.

- Adobe Acrobat desktop: open the PDF, choose Edit, then Watermark > Add in the left pane. [1]
- Choose Text for a label or File for an image. Set size, opacity, rotation, and position. [1]
- Preview the watermark and check the page range. Keep it away from small text, charts, and signatures. [1]
- Select OK, save a separate copy, and reopen that copy before sharing. Check both portrait and landscape pages.

> *These are Acrobat desktop instructions, not steps for every PDF viewer. Available tools depend on your application and edition.*

Labels, signatures, and credentials differ

A visible watermark is page content. A certificate-based PDF digital signature gives software a way to check the signed content and the signer's certificate. A drawn signature image is not that cryptographic check. In Acrobat, recipients can open the Signatures panel, choose Options > Validate Signatures, and inspect Signature Properties and the signer's certificate. [2]

C2PA Content Credentials provide a different structured record: assertions about an asset, bound into digitally signed manifests. A compatible validator checks the signature, credential, assertions, and association with the asset. C2PA 2.4 describes this provenance system; it does not establish that every statement in a document is true. [3]

These mechanisms can serve different needs in the same workflow. Adding a watermark does not create either a PDF digital signature or a C2PA manifest. Support for one does not imply support for the other. For an implementation example, Encypher's PDF provenance guide describes its signing and verification workflow. [4]

- For a visible status cue, add a label.
- For signature or provenance evidence, inspect the relevant validation result.

> *CONFIDENTIAL is a notice, not access control. Manage recipients and sharing permissions separately.*

Check the shared copy

Saving, exporting, and printing are different steps. Check the final file rather than assuming its watermark, signatures, or credentials match the source. Acrobat exposes watermark appearance controls; inspect both the on-screen result and a print sample when paper distribution matters. [1]

A printed page or screenshot can show a label, but its appearance alone is not digital validation. Give recipients the digital file and instructions for the relevant verifier. A PDF signature panel and a C2PA validator answer different questions. [2][3]

- Inspect every page for missing marks, clipping, and poor contrast.
- Validate signatures or credentials on the exact file being sent.
- Ask recipients to review signer trust and warnings, not badges.

> *Keep the original and final copies distinct. A visible label cannot prevent copying or establish ownership.*

Sources and further reading

[1] Adobe: Add watermarks to PDFs in Acrobat on desktop

[2] Adobe: Validate digital signatures

[3] C2PA 2.4 Technical Specification, introduction and technical overview

[4] Encypher: PDF content provenance (implementation guide)

C2PA 2.4. Sources checked September 11, 2026. App interfaces and support can change.

See a signed record for yourself

Try a non-sensitive example: Encypher signed-text demo

Check format support and upload handling before submitting files. These guides explain concepts; they do not certify content or establish legal compliance.

PowerPoint and Keynote watermarks explained

PowerPoint: use the slide master

In a presentation, a watermark usually means faint text or an image repeated behind the content. It can communicate DRAFT or identify a brand. It does not prevent copying, prove authorship, or restrict who can open the deck. A slide master helps you place repeated elements without editing each slide separately.

PowerPoint desktop: Microsoft documents a text-watermark method using Slide Master view. PowerPoint does not offer Word's gallery of ready-made watermarks. [1]

- Choose View > Slide Master. Select the top master thumbnail in the left pane. [1]
- Choose Insert > Text Box, draw the box, and type your watermark text. [1]
- Select the text and choose a light font fill. Adjust size and use the rotation handle if needed. [1]
- Close Slide Master view. Inspect the title slide and each layout; Microsoft's instructions specifically note a title-page exception. [1]

> *Illustrative example: repeat DRAFT on review slides, then check that the title page and image-heavy slides also show the intended notice clearly.*

Keynote: edit the layouts you use

Keynote on Mac uses slide layouts for repeated elements. Text or images added as ordinary layout objects become part of the background on slides using that layout. Apple distinguishes these from placeholders, which are intended to be edited in individual slides. [2]

This approach is useful for a repeated word or small logo. Choose a position that does not cover charts, captions, or slide numbers. A quiet corner label can be easier to read than a large mark across dense material.

- Open your presentation. Click the View menu button in the toolbar and choose Edit Slide Layouts. [2]
- Select a layout used by your slides. Add your text or logo image as an ordinary object, not a placeholder. [2]
- Adjust the object's appearance and position. Repeat on other layouts used in the deck; one layout change does not cover unrelated layouts. [2]
- Click Done, then inspect slides based on each edited layout. [2]

> *A layout object is a design convenience, not a security boundary. Someone editing the layout can change the mark.*

Export, inspect, and share deliberately

Inspect the exported PDF, not only the editable deck. In Keynote on Mac, choose File > Export To > PDF. Review options for presenter notes, skipped slides, comments, and build stages before exporting. These choices affect what recipients receive. [3]

Keynote can open and edit PPTX files. That does not make a downloaded PPTX a native .key file or guarantee identical rendering. Check fonts, text wrapping, images, and watermark placement after import. [4]

- Review every PDF page for missing or obscured marks.
- Inspect a print sample if the audience will use paper.
- Manage sharing permissions separately. For a separate provenance workflow, consult the document-format coverage in Encypher's provenance guide. [5]

> *Downloading an educational presentation about watermarking does not watermark your own deck. Apply the steps to your file, then inspect the result before sending it to others.*

Sources and further reading

[1] Microsoft: Add a watermark to your slides

[2] Apple: Add and edit slide layouts in Keynote on Mac

[3] Apple: Export to PowerPoint or another file format in Keynote on Mac

[4] Apple: Open or close a Keynote presentation on Mac

[5] Encypher: What is content provenance? (explainer)

C2PA 2.4. Sources checked September 11, 2026. App interfaces and support can change.

See a signed record for yourself

Try a non-sensitive example: Encypher signed-text demo

Check format support and upload handling before submitting files. These guides explain concepts; they do not certify content or establish legal compliance.