Coalition for Content Provenance and Authenticity

C2PA and Content Credentials Explainer

Table of Contents

Chapter 1. Introduction

In an era where digital media (e.g., images, videos, audio and documents) is easily created and manipulated, whether by humans or generative AI, it is important for users to be able to verify the authenticity and provenance of such media to prevent them from being misled or harmed. These concerns have been raised across journalism, social media, and other industries that rely on trust and authenticity. The Coalition for Content Provenance and Authenticity (C2PA) was formed to address these challenges by developing an open standard that enables creators, publishers, and consumers to verify the origins and history of digital content.

The purpose of this explainer is to provide a clear understanding of C2PA's work and its Content Credentials technology. This document attempts to use clear language to describe the C2PA architecture, motivating user needs and key components. It will also serve to provide some background and clarification on the development of the standard and its goals.

Chapter 2. Goals and Non-goals

The goal of the C2PA Specifications for Content Credentials is to tackle the extraordinary challenge of trusting media in the context of rapidly evolving technology and the democratization of powerful creation and editing techniques. To this end, the C2PA specifications are designed to enable global, opt-in, adoption of digital provenance techniques through the creation of a rich ecosystem of digital provenance enabled applications for a wide range of individuals, organizations and devices, while meeting appropriate security and privacy requirements, as well as human rights considerations.

Provenance, as C2PA defines it, refers to the facts about the history of a piece of digital content (also known as an asset) in a form such as an image, video, audio recording, or document. At the heart of the C2PA specification is the Content Credential, a cryptographically bound structure that records an asset's provenance. Content Credentials, also known as a C2PA Manifest, contain one or more assertions, which are statements about the asset, such as its origin (i.e., when and where it was created), modifications (i.e., what happened using what tools) and use of AI (i.e., how it was authored). While C2PA defines a set of assertions, it also allows others to add their own. These may include assertions about who created the content, whether they wish to allow its use for AI training or industry-specific information such as supply chain info.

It is important to highlight that Content Credentials do not provide value judgments about whether a given set of provenance data is 'true', but instead merely whether the provenance information is well-formed and free from tampering, valid and trusted (in that the signer of the Content Credential is associated with a known trust list). In addition, the provenance information can be verified as associated with the underlying asset ("a valid asset").

It is not a cure-all for misinformation, but instead seeks to mitigate against its threats in the digital domain. It complements media literacy, fact-checking, and digital forensics approaches such as deep-fake detection by providing an infrastructure to record all of that information in a tamper-evident structure, representing the provenance of any asset.

Chapter 3. Core Principles

Some of the core principles that underpin the C2PA specifications are:

- Content Credentials provides a way to establish provenance of content. There are no value judgments about whether the provenance data is good or bad.

- Implementations of Content Credentials are designed to be interoperable

- C2PA specifications respect user privacy

- C2PA specifications must not lend themselves to abuse, harm and threats to human rights

- C2PA specifications are global

NOTE The full set of principles may be found in our Guiding Principles document.

Chapter 4. Core Concepts

Assertion

A data structure which represents a statement about the asset. This data is a part of the C2PA Manifest.

Asset

A file or stream of data containing digital content, asset metadata and optionally, a C2PA Manifest.

Asset metadata

Non-technical information about the asset and its digital content.

Authenticity

A property of digital content comprising a set of facts that can be cryptographically verified as not having been tampered with.

C2PA Manifest

The set of information about the provenance of an asset consisting of one or more assertions that are digitally signed to ensure their authenticity and integrity.

Content Credential

This is the preferred non-technical term for a C2PA Manifest.

Content Credentials also refers to the overall C2PA technology, and is therefore essentially treated as a plural noun. If a C2PA Manifest is a Content Credential, then multiple C2PA Manifest or the broader, universal concept is Content Credentials.

Content binding

Information that associates digital content to a specific C2PA Manifest associated with a specific asset.

Digital content

The portion of an asset that represents the actual content, such as the pixels of an image, along with any additional technical metadata required to understand the content (e.g., a colour profile or encoding parameters).

Provenance

The logical concept of understanding the history of an asset as represented by the set of C2PA Manifests for an asset and its ingredients (if any).

Chapter 5. Motivating User Needs

What follows are some key illustrative user needs that help drive the development of this technology.

5.1. Helping consumers check the provenance of the media they are consuming

Alice sends a video to a friend, Bob. The video includes text with alarming and controversial allegations. Bob immediately seeks confirmation of its validity, starting with its provenance.

The video that Alice sent contains a Content Credential. With a Content Credential-aware application, Bob is able to establish that this video has not been modified since the Content Credential was added, was signed by a trusted implementation, and that its provenance includes (along other useful trust indicators) information about the publishing organisation. Using these indicators, Bob can make a more informed decision about the trustworthiness of the video.

5.2. Enhancing clarity around provenance and edits for journalistic work

A photojournalist uses a Content Credential-enabled capture device during a newsworthy event they are covering. The assets are then brought into a Content Credential-enabled editing application, and after editing it, they are sent to a photo editor. The editor makes additional edits also using a Content Credential-enabled application. The finalized asset is moved into the content management system of a news organization, which is also Content Credential-enabled, before posting the asset to social media.

5.3. Offering publishers opportunities to improve their brand value

A news publisher is concerned about standards of public comprehension and brand value of its publications which it makes available online through a number of social media platforms. To improve audience confidence about their content, it wishes to provide a means for the audience to verify the content that originated through its output.

For content that is consumed with C2PA provenance, the publisher provides the consumer needed details to rely upon the provenance and authenticity of the asset.

5.4. Providing quality data for indexer / platform content decisions

A news video is posted to a social media platform. By utilizing the Content Credential-provided provenance in the

video, the social media platform is able to verify that it came from the same source that posted it.

5.5. Assisting 'Intelligence' investigators to confirm provenance and integrity of media

An individual in a news/other context using open-source intelligence techniques (OSINT) can use the presence of a Content Credential in assets to better confirm the history and integrity of media. Additionally, an individual may use a durable Content Credential to re-correlate relevant media to its provenance.

5.6. Enhance the evidentiary value of critical footage

A human rights defender manages to capture footage containing a Content Credential of police violence during a protest. The human rights defender sends the footage to a human rights organization that verifies that the asset meets video-as-evidence criteria (for example, that it does not contain AI generated content). The human rights organization redacts information about the defender using a Content Credential-enabled editing application in order to protect their identity. The Content Credential is verified, which improves the chances of that footage being admissible in court proceedings.

5.7. Enforcing disclaimer laws on retouched/edited images

To prevent dangerous stereotypes of ideal bodies, a government enacts a law that requires advertisers and social media influencers to specify that their image has been edited if any aspect of a body's size, shape or skin has been altered. By having their Content Credential-enabled editing application add information about each action performed, they can easily comply and the government can easily confirm.

Chapter 6. How Content Credentials Works

6.1. An End-to-End Workflow

The C2PA specification for Content Credentials provides a standardized method for attaching Content Credentials, which can be validated, to digital assets. Here's how it works:

1. **Content Creation**: A creator uses a Content Credential-enabled tool to produce digital content, such as a photograph or video. During this process, provenance information is generated, including a cryptographic signature and possibly including additional details about the creator, the device used, and any applied edits.

2. **C2PA Manifest Generation**: The provenance information is encoded into a data structure called a C2PA Manifest, also known as a Content Credential. This C2PA Manifest can contain variety of assertions representing the provenance of the asset, including:

   ◦ A description of the asset's origin.

   ◦ Details of any modifications or edits.

   ◦ A cryptographic hash representing the content at the time of creation.

3. **Cryptographic Signing**: The C2PA Manifest is signed with the private key of the software or hardware that performed the operations, thus ensuring its authenticity and integrity. The corresponding public key is made available for verification.

4. **Embedding and/or Watermarking**: The C2PA Manifest is commonly embedded directly within the asset, though it can also be linked externally through a soft binding system (e.g., invisible watermarks). This ensures that the provenance information is always accessible.

5. **Distribution**: The asset, along with its embedded C2PA Manifest, is distributed via platforms like social media, news sites, or file-sharing services. The C2PA Manifest travels with the asset, ensuring that the provenance information remains intact.

6. **Verification**: When an end user views or interacts with an asset using a Content Credential-enabled application, the application checks the integrity of, and verifies, the C2PA Manifest and its associated asset. It can confirm that the asset has not been modified since the C2PA Manifest was created, and that the provenance information is also unmodified and authentic, added by a trusted (or not) implementation.

7. **Presentation**: When presented to a user, verified content will be marked with a clear indicator, such as a badge or icon, signalling to the user that the asset and its provenance is authentic and intact.

6.2. Core Components

Provenance

Information about content origin and modifications, and any other relevant assertion concerning the history of the asset.

Cryptographic Hashes and Signatures

Ensures integrity and authenticity of the Content Credential. If any part of the Content Credential is tampered with it will invalidate one or more hashes and signalling tampering.

Durable Credentials and Soft Binding

- So called "soft bindings" enhance the durability of Content Credentials by enabling the discovery of the manifest if it is not embedded in an asset.

- Soft bindings can either be implemented via invisible watermarking or fingerprint lookup.

Chapter 7. FAQ

7.1. General

7.1.1. Can Content Credentials alone combat misinformation?

A C2PA Manifest, the Content Credential, is a part of the solution. It complements media literacy, fact-checking, and other efforts by providing tools for establishing content provenance. Its effectiveness depends on widespread adoption and user awareness.

7.1.2. Should I distrust media without Content Credentials?

Maybe.

Adding provenance to an asset is optional, and it is not the intention of the specification and guidance to create a two-tier media ecosystem where assets without Content Credentials are universally less trusted than assets with it. The C2PA specifications for Content Credentials are open and available to everyone, and so no assumption should be made about the trustworthiness of a particular asset purely based on its usage of Content Credentials.

Since trust in media is based on many factors such as reputational source and confidence in the products and processes that create it, Content Credentials become useful to consumers of assets as they can use the information contained therein to determine their level of trust. For example, if a well-known media publisher adds provenance data to an asset, a consumer that knows that publisher can use Content Credentials to understand that the asset and its provenance data definitely came from them, and wasn't manipulated. Conversely, if a bad or unknown actor adds provenance data, a consumer may not be able to locate any useful information, and that would help them to make their own decision on whether to trust the asset.

7.1.3. Can manipulation or tampering be detected?

A C2PA Manifest, the Content Credential, along with the asset are the two parts of a unique puzzle. The possibility of any other pieces ever matching, either by coincidence or by a purposeful attempt to generate a match, is so low that it would be practically impossible. In other words, any alteration to either the asset or the provenance, however insignificant, would alter the shape of the piece of the puzzle, which is a mathematical algorithm called a cryptographic hash, in such a way that they would no longer match. This means that the provenance data can be said to be tamper-evident.

The C2PA specification combines the use of standard cryptographic hashes, such as SHA2-256, with a Merkle tree-like approach of having one hash included in a secondary hash (etc.) and then finally digitally signed by the signer using standard X.509-based credentials.

7.1.4. Can small organizations implement Content Credentials?

Yes. Organizations of all sizes can implement Content Credentials in their products and services, which can then be

utilized by themselves or their customers (i.e., individuals).

The C2PA specifications for Content Credentials are open standards that are released under a royalty free license. This enables both open and closed source implementations and thereby lowers the burden on implementers. It also enables creation of open source libraries, which further enables organizations of all sizes to implement Content Credentials.

7.2. Provenance

7.2.1. Is provenance always complete?

No. Provenance is not always complete. It may happen that an asset is modified in a way that the provenance data is not updated. For example, if an asset is cropped using a non-Content Credentials aware tool, the provenance data may not be updated to reflect that action. However, if the asset is then brought back into a Content Credentials aware tool for additional modification or preparation for publication, the signer of the new Content Credential also implicitly attests to the crop action. So even though there is missing provenance information, the asset can still be trusted based on the signer of the active Content Credential.

7.2.2. Can provenance information be used to determine whether a digital asset, such as an image or video, depicts the truth?

Provenance information can help establish the truth about the origin, history and authenticity of digital content, by providing evidence for its creation, discovery, ownership and movement over time; but provenance information alone cannot tell you whether the digital content is true, accurate or factual.

Content Credentials can include assertions about the real-world identity of the provider of those assertions and the digital asset to which they refer. This allows one to determine whether those assertions or the digital asset itself were subsequently altered. This provides consumers the means to make a more informed decision about whether they believe the digital content is true, accurate or factual, based in part on the trust relationship they have with the provider of those assertions.

For additional information, see the Trust Model section of the Technical Specifications.

7.2.3. Can the provenance metadata be removed?

Yes it can. That is why the C2PA specification includes the concept of durable Content Credentials, which combines a hard binding (aka cryptographic hashing) with a soft binding (e.g., watermarking and fingerprinting). These soft bindings allows the Content Credential to be discovered in cloud-hosted storage even if it is removed from the asset.

7.2.4. Does provenance have to include the identity of the creator of a piece of media?

The C2PA specification does not directly address the topic of human or organizational identity — that is, the identity of specific humans or organizations involved in the production of a digital asset or content. Instead, it focuses on the provenance and authenticity of the content itself. This ensures respect for user privacy.

However, there are extensions to the core C2PA specification that introduce support for such identity features which can be included in a Content Credential. When doing so, users should be aware of the privacy implications of including such information.

7.2.5. Can you explain more about human and organizational identity in Content Credentials?

In some use cases, an individual or organization may wish to describe their own identity or assert additional metadata that can not be directly represented as by the features of the core C2PA specification. In these cases, it is recommended to use an extensions to Content Credentials that supports the identification of the actor(s) who are making these statements, along with the ability for them to securely represent their relationship to the media asset and the specific assertions they wish to make.

An example of such an assertion is the Creator Assertions Working Group's identity assertion.

7.3. Multiple Sources & Ingredients

7.3.1. Can Content Credentials help with assets created from multiple sources?

Yes it can. When an asset is created from a series of other assets, those additional assets are referred to as its ingredients. Each ingredient that is used in the (composed) asset is recorded in that asset's provenance, including the addition of the provenance of each individual ingredient. This process creates a tree of provenance, much like a family tree, that can stretch all the way back to each ingredient's creation.

7.3.2. Do the ingredients of an asset have verifiable provenance?

Each ingredient that is included in a Content Credential can include its own provenance data, specifically the ingredient's Content Credential is also included in the asset's full set of Content Credentials. However, while the provenance data of each ingredient may be present, the ingredient's provenance cannot be verified in the same way as the provenance data of the asset in which it is contained. This is because the actual data of the ingredient is not usually included in the asset's Content Credential. Without the actual data, the ingredient's hard bindings cannot be verified. For this reason, the hard bindings of the ingredient, as well as the validity of its Content Credential are validated at the time they are added to the asset and a record of that validation is included in the asset's Content Credential. This allows the consumer to see that the ingredient was validated at the time it was added to the asset and use that information to determine their level of trust in the asset.

7.4. Relationship to other technologies

7.4.1. Does this make use of "the blockchain?"

No, Content Credentials from the C2PA, does not rely on blockchain (or other forms of distributed ledger) technology. Instead, it uses established cryptographic techniques like cryptographic hashing, Merkle trees and digital signatures

to provide tamper-evident Content Credentials. This approach avoids the complexity, environmental concerns, and scalability issues often associated with blockchain while ensuring robust provenance tracking.

However, the C2PA specifications are designed to be compatible with blockchain technology, and it is possible to use blockchain technology to store and distribute Content Credentials. In fact, some members of C2PA have deployed blockchain-based solutions for this purpose.

7.4.2. How does the data in Content Credentials relate to standard metadata formats such as IPTC, XMP, and EXIF?

Content Credentials are used in conjunction with standard metadata formats such as IPTC, XMP, and EXIF. The C2PA specification defines a set of metadata assertion along with a mapping from these existing metadata formats. This allows for interoperability between different systems and applications that use these formats, and also enables users to make their standard metadata available in a Content Credential and therefore tamper-evident.

Additionally, the C2PA specification allows for the inclusion of other metadata formats - both open as well as proprietary - via custom assertions. This enables users to include additional information about the asset and its provenance using a common model.

7.4.3. How do Content Credentials differ from Digital Rights Management (DRM)?

Unlike DRM, Content Credentials does not restrict user access or impose usage limitations. It focuses on transparency and trust by providing verifiable provenance information. Its open standard nature ensures adaptability and public accountability.

However, Content Credentials can be used in conjunction with DRM systems that wish to incorporate provenance information into their content protection mechanisms. Doing so is out of scope for the Content Credentials specification and the C2PA.

7.4.4. How do Content Credentials address the use of AI/ML in the creation and editing of assets?

Each action that is performed on an asset can be recorded in the asset's Content Credential. These actions can be performed by a human, organization or by an AI/ML system. When an action was performed by an AI/ML system, it is clearly identified as such through it's `digitalSourceType` field.

> An example of `c2pa.created` action that might appear in an asset that was produced by a Generative AI system appears in the specification's parameters clause of Actions.

7.4.5. Is post-quantum cryptography supported by the C2PA standard?

Future support for ML-DSA algorithms (ML-DSA-44, ML-DSA-65, and ML-DSA-87) with approved parameters, as defined in NIST's FIPS 204, is planned for the C2PA standard. ML-DSA will be added to the list of supported signature

algorithms when stable ML-DSA support is broadly available in commonly used cryptographic libraries.

7.5. Trust Model

7.5.1. Is the C2PA building their own trust infrastructure?

Yes and no.

The C2PA defines an open international standard around the creation and verification of Content Credentials. Implementers of this standard can have their implementations accredited by the C2PA, which enables users to trust in that implementation. These implementations will be recognized as part of the C2PA Trust List - similar to how a web browser or PDF viewer includes their own set of Trust Lists (e.g., CA/Browser Forum, EUTL, etc.).

However, the trust in the contents of a Content Credential is based on the trust relationship between the creator of the Content Credential and the consumer of that Content Credential. This trust relationship is not defined by the C2PA, but by the ecosystem in which the Content Credential is used, including the foundational trust layer that C2PA is building. It may also include other trust list, such as from the IPTC or even a user's own personal trust list.

This trust model is based on the same technology behind SSL/TLS and PDF signatures and can be combined with established identity frameworks, such as one from the Creator Assertions Working Group.

7.5.2. What is the signer asserting in a claim?

The signer of a Content Credential is the credential holder of the certificate that is used by the Claim Generator to sign the Content Credential. Only the set of assertions that are listed in the `created_assertions` field of the Content Credential's Claim are attributed to the signer. Those in the `gathered_assertions` field are not.