Coalition for Content Provenance and Authenticity

C2PA Explainer

1.4, 2023-12-12: Release

Table of Contents

Chapter 1. Introduction

The development of this Explainer is ongoing. This Explainer accompanies the C2PA Specifications for Content Credentials with the intent of providing further background and clarification on the development of the standard, its goals, mechanisms and guidelines.

This Explainer is not a technical document and is directed towards the general public.

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 a context of rapidly evolving technology and the democratization of powerful creation and editing techniques. To this end, the 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 and organizations while meeting appropriate security and privacy requirements, as well as human rights considerations.

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 can be verified as associated with the underlying asset, correctly formed, and free from tampering.

Chapter 3. Fundamentals (FAQs)

3.1. Provenance

3.1.1. What does "Provenance" mean with Content Credentials?

Provenance generally refers to the facts about the history of a piece of digital content assets (image, video, audio recording, document). Content Credentials enables the authors of provenance data to securely bind statements of provenance data to instances of content using their unique credentials. These provenance statements are called assertions in a Content Credential. They may include assertions about who created the content and how, when, and where it was created. They may also include assertions about when and how it was edited throughout its life. The content author, and publisher (if authoring provenance data) always has control over whether to include provenance data as well as what assertions are included, such as whether to include identifying information (in order to allow for anonymous or pseudonymous assets). Included assertions can be removed in later edits without invalidating or removing all of the included provenance data in a process called redaction.

3.1.2. What does it mean that provenance data is cryptographically bound to the asset?

The provenance data and the asset are the two parts of the same puzzle - 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 mathematical algorithm - the shape of the piece of the puzzle - in such a way that they would no longer match.

We refer to this as a hard binding. For more technical information on this see "hard binding" in the glossary and the non-normative guidance.

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

When one asset is created from a series of other assets, those sources are referred to as the 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 chain of provenance that can stretch all the way back to each ingredient's creation.

3.1.4. What is redaction and how does it work?

Redaction is the process of permanently removing information. In the world of Content Credentials it specifically refers to removing assertions.

For example, if a human rights organization wishes to remove assertions about the photographer from an image, it can do so via the redaction process. The act of redaction (and the optional inclusion of a reason for that redaction) becomes part of the provenance of the asset.

3.1.5. 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 actor responsible for signing the new Content Credentials 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.

3.2. Trust

3.2.1. How is trust in digital assets established?

With Content Credentials, trust decisions are made by the consumer of the asset based on the identity of the actor(s) who signed the provenance data along with the information in the assertions contained in the provenance. This signing takes place at each significant moment in an asset's life (e.g., creation, editing, etc.) through the use of the actor's unique credentials and ensures that the provenance data remains cryptographically bound to the newly created or updated asset.

To enable consumers to make informed decisions about the provenance of an asset, and prevent unknown attackers from impersonating others, it is critical that each application and ecosystem reliably identify the actor to whom a signing credential (also known as a digital certificate) is issued. A certification authority (CA) performs this real-world due diligence to ensure signing credentials are only issued to actors who are whom they claim to be.

In the world wide web, CAs verify that someone requesting a certificate to operate a secure web site owns and/or controls the site's domain name before issuing such a credential. For example, before issuing a certificate for https://c2pa.org/, the CA verified the requester did in fact control C2PA's domain name before issuing a certificate for that site name. Unlike the world wide web, Content Credentials will be used in multiple different application settings, and each setting will have its own requirements. Each application will therefore provide users with one or more *trust lists*, which are lists of certification authorities that issue signing credentials for that application.

Once the signing credential is issued by a certification authority, the identity it has confirmed and placed in that credential cannot be altered by anyone else, including and especially the credential's owner. This allows the consumer or user to rely upon the identity presented in a signed asset.

Example 1. News and Media

For example, in a news and media aggregation application or web site, each newspaper, television network, or other media organization has a globally-recognized identity, and there is only one such organization that operates with that name. In this situation, because brand marks and other visual indicators can and have been reproduced for the purposes of impersonation, consumers want to be certain the media they are consuming actually comes from the source it claims to be. This application would provide at least one trust list maintained by a professional or non-profit organization for journalists and media, which endorses certification authorities

that ensure such credentials are only issued to the genuine organization through real-world due diligence.

Example 2. Insurance Company

In another example, an insurance company may employ provenance tracking for images, videos, and other media as part of underwriting policies and servicing loss claims to be used by its own employees. In this case, only one trust list is applicable: the insurance company's own certification authority operated by its Human Resources department, which may already exist to employee credentials for other purposes. Here, it is not important that every participant have a unique name, as several employees may share the same name, but it is important to be assured all participants are employees of the insurance company, which the Human Resources department is certainly able to confirm and attest.

3.2.2. Should I distrust media without Content Credentials?

No. 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 is open and available to everyone, and so no assumption should be made about the trustworthiness of a particular asset or signer purely based on their usage of Content Credentials.

Content Credentials becomes useful to users when they can use the data and the signer's identity to build a trust relationship with that signer. For example, if a well-known media publisher adds provenance data to an asset and signs it, 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 and signs it, a consumer would see the actor's identity, and then make their own decision on whether to trust the asset and the provenance data.

3.2.3. 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 its 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.

3.3. 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 signature allows one to determine whether those the assertions or the digital asset itself were subsequently altered. This provides users 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.

> Signed provenance information transits trust between a creator and a consumer, based on a trust relationship between those two that exists outside the scope of C2PA. Trust anchors operate by providing digital identities within a particular ecosystem that link to real-world identities, and perform an ecosystem-specific validation to ensure those identities are sufficiently trustworthy, and that consumers can be confident when an asset is verified as being signed by a known creator, they can rely upon their existing trust relationship with that known creator. — ***C2PA Implementation Guidance Trust Model***

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

3.4. Use of Artificial Intelligence & Machine Learning (AI/ML)

3.4.1. How does C2PA address the use of AI/ML in the creation and editing of assets?

Each action that is performed on an asset is recorded in the asset's Content Credentials. These actions can be performed by a human 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.

3.4.2. Can C2PA be used to label assets that should not be used for training or data mining?

Yes. A Content Credential can include a Training and Data Mining assertion that can be used to indicate that the asset should not be used for either training or data mining purposes. The assertion is flexible and allows the author of the asset to specify whether each type of process - data mining, general AI training, or training specific to generative AI - is permitted, or not.

Chapter 4. Use-case Examples

The following is a non-exhaustive list of potential and general use-cases of the C2PA Specifications for Content Credentials. Some of these are taken from, or built upon, the use-cases developed within the Project Origin Alliance and the Content Authenticity Initiative (CAI) frameworks. Each use-case will be described using some generic personas to help make the flow clear.

For technical use-case examples, see non-normative guidance.

4.1. 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 Content Credentials. With a Content Credential-enabled application, Bob is able to establish that this video has been validated as being published by an organisation he can trust and is held in public high regard.

4.2. Enhancing clarity around provenance and edits for journalistic work

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

4.3. 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 without any Content Credentials, the publisher hopes that the consumer will take extra steps to verify the provenance and authenticity of the asset, instead of immediately attributing it to the site on which it is published.

4.4. Providing quality data for indexer / platform content decisions

A news video is posted to a social media platform. By utilizing the Content Credentials-enabled provenance in the video, the social media platform is able to verify that it came from the same source that posted it.

4.5. 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 Content Credentials in assets to better confirm the history and integrity of media. Additionally, an individual may use a decoupled binding database to re-correlate relevant media to its Content Credentials.

4.6. Enhance the evidentiary value of critical footage

A human rights defender manages to capture footage containing Content Credentials-enabled provenance 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. The human rights organization redacts information about the defender using a Content Credentials-enabled editing application in order to protect their identity. The Content Credentials-verified asset is then used to improve the chances of that footage being admissible in court proceedings.

4.7. 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 Credentials-enabled editing application add information about each action performed, they can easily comply and the government can easily confirm.

Chapter 5. Stakeholder Feedback

NOTE Implementers and other stakeholders may already have publicly stated positions on this work. They will be listed here with links to evidence as appropriate.