August 2020

The Content Authenticity Initiative

Setting the Standard for Digital Content Attribution

Authors

- Leonard Rosenthol (Adobe)
- Andy Parsons (Adobe)
- Eric Scouten (Adobe)
- Jatin Aythora (The British Broadcasting Corporation)
- Bruce MacCormack (CBC/Radio-Canada)
- Paul England (Microsoft Corporation)
- Marc Levallee (The New York Times Company)
- Jonathan Dotan (Stanford Center for Blockchain Research)
- Sherif Hanna (Truepic)
- Hany Farid (University of California, Berkeley)
- Sam Gregory (WITNESS)

Version 1.0

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.

To view a copy of this license, visit https://creativecommons.org/licenses/by-nd/4.0/.

© Adobe 2020

Table of Contents

© Adobe 2020

© Adobe 2020

© Adobe 2020

1 Introduction

1.1 Overview

With the increasing velocity of digital content and the democratization of powerful creation and editing techniques, robust content attribution is critical to ensure transparency, understanding, and ultimately, trust.

We are witnessing extraordinary challenges to trust in media. As social platforms amplify the reach and influence of certain content via ever more complex and opaque algorithms, mis-attributed and mis-contextualized content spreads quickly. Whether inadvertent *misinformation* or deliberate deception via *disinformation*, collectively *inauthentic content* is on the rise.

Currently, creators who wish to include metadata about their work (for example authorship) cannot do so in a secure, tamper-evident and standard way across platforms. Without this attribution information, publishers and consumers lack critical context for determining the authenticity of media. This is especially true for users of creative tools that enable augmenting reality with AI or even authoring fully synthetic content who need to be empowered to use their tools responsibly.

Ultimately, the solution to the problem of inauthentic content and the erosion of trust it causes will rely on efforts in three distinct areas:

First is *detection* of deliberately deceptive media. Through a combination of algorithmic identification and human-centered verification of intentionally misleading content the amount of inauthentic content can be reduced. However, as techniques for creating misleading content become more sophisticated and accessible we foresee an escalating arms race impeding progress on this front. As malicious purveyors of content become faster and better, detection techniques will struggle to keep pace.

Second, *education* is essential. Well-intentioned creators and consumers will need to understand the danger of disinformation and the use of techniques to eradicate it. They must also understand ways to use sophisticated creative tools responsibly. These are skills that must be learned and passed on through media literacy campaigns and formal education. We must all understand why and when to trust what we see, hear and read. And we must be equipped with the tools and knowledge to do so.

Finally, we must consider *content attribution*, which is the focus of this paper. Often referred to as provenance, attribution empowers content creators and editors, regardless of their geographic location or degree of access to technology, to disclose information about who created or changed an asset, what was changed and how it was changed. While detection

© Adobe 2020

can help address the problem of trust in media reactively by identifying content suspected to be deceptive, attribution proactively adds a layer of transparency so consumers can be informed in their decisions. Content with attribution exposes indicators of authenticity so that consumers can have awareness of who has altered content and what exactly has been changed. This ability to provide content attribution for creators, publishers and consumers is essential to engender trust online.

At the same time, it is critically important that those same content creators be able to protect their privacy when necessary. Any solution attempting to restore trust must be globally viable across technology contexts and minimize opportunities to cause unintended harms or risks. It must also have freedom of creative expression in media production at its core.

We seek to address the issue of content authenticity at scale. To accomplish this, we propose an open, extensible approach for content attribution and have begun working toward establishing standards with broad, cross-industry collaboration.

1.2 Background

At Adobe MAX 2019, the Content Authenticity Initiative (CAI) was announced by Adobe in partnership with The New York Times Company and Twitter. Since that time, this group has collaborated with a wide set of representatives from commercial organizations (software tools, publishers, social media), human rights organizations and academic research to produce this paper and the approach it describes.

1.3 Our Mission

The initial mission of the CAI is to develop the industry standard for content attribution. By augmenting subjective judgments about authenticity with objective facts about how a piece of content came to be, the CAI aims to help content consumers make more informed decisions about what to trust.

Today, most attribution information is embedded in the metadata of assets via long-established standards such as EXIF and XMP. However, most assets appear on the Web without this information intact. Content moderators, fact-checkers and end-users alike are left to reconstruct context through imperfect and inefficient methods. We will provide a layer of robust, tamper-evident attribution and history data built upon XMP, Schema.org and other metadata standards that goes far beyond common uses today. This attribution information will be bound to the assets it describes, which will in turn reduce friction for creators sharing the attribution data and enable intuitive experiences for consumers who use the information to help them decide what to trust.

We balance simplicity in use of the system with security against tampering and strong links to identity. Identity can be that of an individual, where prudent, or that of the trusted cryptographic signing entity.

There is currently no universal approach for storing attribution data appropriate for all use cases. Depending on the systems involved, this information may be large enough to make it impractical to embed in a file containing digital content (hereafter called an asset). Conversely, some creators may have privacy concerns such that no data associated with an asset nor the asset itself can be stored on servers in the cloud. A cloud-based system may provide durability, whereas a file-based workflow optimizes for disconnected workflows and the preservation of anonymity. Therefore, the CAI imagines data storage to comprise a continuum of options ranging from file-based to cloud-based, with hybrid approaches in between. Flexibility for applications in implementing persistence and flexibility for end users to choose where their data is stored is essential for widespread adoption.

Increasing trust in media requires the ongoing engagement of diverse communities. The CAI does not prescribe a unified single platform for authenticity, but instead presents a set of standards that can be used to create and reveal attribution and history for images, documents, time-based media (video, audio) and streaming content. Although the initial implementations will focus on imagery, our intent is to specify a largely uniform method for enabling attribution from various points of view through which diverse stakeholders can build decentralized knowledge graphs about the trustworthiness of media.

2 Guiding Principles

To provide clarity on goals, methodology and purpose, CAI efforts including system design and user experience design are governed by a set of guiding principles. These have informed the work done to date and will continue to be employed to ensure consistency in future work streams.

The terms *must*, *must not*, *should*, *should not*, *required*, and *may* are used in accordance with RFC 2119.

2.1 Overarching Goals

CAI specifications **should** provide a mechanism for the producers and custodians of any given content to assert, in a verifiable manner, any information they wish to disclose about the creation of that content and any actions taken since the asset's creation. We refer to such information collectively as provenance.

© Adobe 2020

CAI specifications **should not** provide value judgments about whether a given set of provenance data is "good" or "bad," merely whether the data can be verified as associated with the underlying asset, correctly formed, and free from tampering.

2.2 Privacy

CAI specifications **must** respect the common privacy concerns of creators, publishers and consumers of content.

2.3 Global Audience / Accessibility

CAI specifications **must** take into consideration the needs of interested users throughout the world.

2.4 Interoperability

CAI specifications **should** result in an ecosystem of tools for various types of target users which inter-operate successfully to create, maintain and display provenance information about assets.

2.5 Fit with Existing Workflows

CAI specifications **must** fit into the existing workflows of each of the target users, typically through incremental additions to existing tools.

2.6 Performance

CAI specifications **should** avoid unreasonable performance characteristics for implementors.

2.7 Simplicity and Cost Burden

CAI specifications **should** avoid unreasonable technical complexity and cost burden for implementors.

2.8 Extensibility

CAI specifications **should** provide extensibility to allow for extension and evolution of authenticity data.

© Adobe 2020

2.9 Misuse

CAI specifications **must** be reviewed with a critical eye toward potential abuse and misuse of the framework. In addition, CAI specifications **must** be reviewed for the ability to be abused and cause unintended harm, threats to human rights, or disproportionate risks to vulnerable groups globally.

3 Expected Users

While not intended to limit consideration of other interested parties, CAI users can be broadly understood as:

3.1 Content creators

A content creator is someone who wishes to assert information about content they've produced in a way that can be trusted. Common examples include:

- Creative professionals
- Knowledge workers
- Journalists and news media organizations
  - Including both professional and citizen journalists
  - Including those operating in high-risk environments
- Human-rights defenders
- Amateur producers of news media content

3.2 Content publishers

Content publishers wish to have better information on which to make decisions about what content to trust. They also wish to credit the proper creator. Common examples include:

- News media organizations
- Social media platforms
- Content distribution networks

3.3 Content consumers

Content consumers, those users who interact with assets, want to access authentic content and understand the process by which the content they consume was created. Common examples include:

- Legal professionals including lawyers, investigators, and law enforcement

© Adobe 2020

- Fact-checkers
- Consumers of news media and social media content

3.4 Implementors

Implementors build software or hardware tools to create, persist, exchange, or consume CAI provenance data in a way that is interoperable with other CAI-enabled systems.

4 Workflows

Though not intended to be exhaustive, the following workflows demonstrate the essential characteristics of the system. While these workflows make use of the same fundamental features, there is distinct value derived for different personas.

4.1 Photojournalism

The news media industry is facing a growing sense of distrust in digital media and a changing distribution system as the number of people who receive news from social and non-traditional media sources grows.

The following CAI-enabled workflow is meant to provide trust and transparency for photojournalists, editors and content consumers.

1. A photojournalist uses a CAI-enabled capture device during a newsworthy event they are covering. The photojournalist will set the capture application to the preferred settings of the outlet (i.e. authorship, geolocation, time, file storage preference) and register their identity via the device. The photojournalist then captures images from the newsworthy event with CAI capture attribution details included.
2. The photojournalist then moves their files from a capture application into a photo editing application (e.g. Lightroom, Photo Mechanic, Capture One, etc.). The photojournalist will ensure that the editing application has CAI functionality enabled with the proper settings and manually add metadata about subjects and context as well as complete some light editing.
3. The photojournalist then sends their assets and captions to the appropriate photo editor of their publication. The photo editor opens the assets in a digital imaging tool (e.g. Photoshop), verifies the incoming CAI provenance data, and checks that the data meets editorial standards. The editor then makes edits in accordance with their posted

© Adobe 2020

photo editing guidelines. The photo editor ensures they are utilizing CAI-enabled applications with the appropriate settings throughout their work to ensure that their editing actions are captured and documented.

4. The photo editor works with others to finalize the article to be posted on the website of the news organization. The asset is moved into the content management system of the news organization, which has a CAI implementation so that journalistic context can be displayed and carried through to the website. The article is published.
5. The social media manager then posts links to the article on various social media platforms. While social platforms may alter the asset by, for example, compressing and cropping, the CAI metadata survives these alterations. In fact, these modifications would be added to the CAI data captured in the preprocessing pipeline of the social media platform. The resulting post is CAI-enabled and provides consumers the ability to learn more details about the asset (e.g. Who took it? For what publication? When did they take it?).
6. As other social platform users continue sharing the asset (thus disconnecting it from its original affiliation with the media outlet), CAI data will travel with the asset and any user who sees the asset posted by any other user, will be able to investigate the source and original context of the asset.
7. As required, various analysts from fact-checking organizations will verify the CAI data present in the asset, correlate it to any associated article, and then add their own labels and clarifications to it. Together this collection of CAI data creates a rich, verifiable context that amplifies confidence in the authenticity of the asset.

4.2 Creative Professional

Creatives rely on attribution details to receive recognition and compensation for their completed works. The scale of digital media today has made it relatively common to disassociate the creator and context of an asset from the asset itself. Additionally, creative professionals often make edits to assets and combine them into new works for creative purposes, without any intention to deceive. In many cases the creative will have obtained permission to use ingredient assets and that permission may bring obligations of proper attribution that need to be included in the final work. Although this example depicts a visual creative flow, the scenario applies to the creation of audio, video and document assets as well.

The following CAI-enabled workflow is meant to empower creatives to continue using innovative editing techniques without removing or losing authorship attribution.

1. A creative professional opens an authoring application (e.g. Photoshop) to create a new asset. They may start with existing assets that they would like to composite, or from a new empty document. Similar to the photojournalist use case, the creative will authenticate via the software and select the appropriate CAI settings before beginning work, to ensure that CAI data capture is enabled.
2. The creative will configure their settings to ensure no detailed edit history or work-in-progress thumbnails are captured. This is because they are not attempting to represent a news image where there would be a higher level of concern for transformational edits to the asset. The creative is also less likely to want to share their detailed edit history as it may include trade techniques.
3. Given #2, the creative may use AI-assisted techniques to transform the asset into something that does not represent a discrete event that occurred in the real world. The creative can use these tools within their authoring application and have the CAI-enabled process capture only "before" and "after" renditions to share the impact of the change without revealing their trade secrets to consumers.
4. After completing their asset, the creative will save to their preferred file storage system and post to their own distribution points, such as a marketplace or a stock asset site. The creative chooses distribution points that support the CAI to ensure their attribution information remains intact. These distributors have strong incentives to inspect and preserve attribution.
5. A consumer may then view the asset on social media and decide to engage with it. The consumer will be able to click on a CAI icon to be presented with a preview of key attribution information including thumbnails, author, date and a link to follow for more information.
6. When a consumer clicks through to see more information, they will be directed a CAI-enabled website. This website will provide access to the full set of the asset's CAI data, which have been made public in accordance with the creative Professional's preferences.

4.3 Human Rights Activist

Across the world, people are capturing digital assets as proof of human rights abuses. Many of these activists are not professional content producers. They may live in areas with high levels of surveillance or lacking high connectivity or with lower digital media literacies. Due to the difficulty of sharing assets documenting abuses without exposing themselves to potential harm, human rights activists may use the CAI standards to share their work widely without compromising their identity. There are considerations specific to this workflow for ensuring activists can protect themselves and the reputation of news outlets who publish their assets.

The following CAI-enabled workflow is meant to provide human rights activists as well as non-governmental organizations (NGOs) and media institutions with an option to capture secure and provable details of an asset without unnecessary exposure of privacy details.

1. A human rights activist will select a capture application. Given the non-professional nature of many human rights activists, some will use tools provided by NGOs (e.g. ProofMode) while the majority will use native capture apps on their phone and/or preferred applications. In the path where the user is actively thinking about tools to use beforehand, they will select their CAI preferences within the tool before capture.

2. The human rights activist then captures something of importance within their community on their smartphone. The user may choose to operate online or offline depending on connectivity and privacy concerns. If operating in an offline workflow, CAI data is written to the file itself. While some users may not act beyond this step for significant periods of time, users who do post are likely to share on messaging and social applications (e.g. WhatsApp or Twitter).

3. If connected to an organization, a user may send the asset to the organization in a similar workflow to photojournalists. If the user is not connected to an organization, it is likely that the asset will be later discovered by an NGO or media outlet on messaging and social platforms. These institutions will be able to read the associated CAI data and confirm key attribution aspects before verifying that the asset is accurate.

4. The NGOs and media outlets will then enter their own verification process. In order to verify the assets, they are likely looking at information such as author, time, location, etc. If the asset has preserved CAI data, it will significantly simplify this process. The institution will then likely try to contact the original source as part of verification and may or may not receive the details they desire to feel comfortable sharing on their platform. (If required, this would also be where an institution may decide to use redaction capabilities, e.g. blur out faces before distributing widely.)

5. Assuming the verification process is completed, in a media context, the workflow is then in the hands of a photo editor who will complete a similar processes as the one in "Photojournalism" to post the asset.

6. A key point of differentiation for the asset with CAI information is that it empowers the source, rather than the institutions publishing the asset, to be the trusted entity. As consumers view the content on social media platforms and/or the CAI website, the attribution details that they can see will start with the original capture and mitigate uncertainty for end users who may or may not trust the publishing institution.

© Adobe 2020

4.4 Non-ideal Workflows

The scenarios presented above assume wide adoption of CAI standards. We would be remiss not to acknowledge that in the early phases of adoption, many steps in many workflows will not be CAI-enabled. For example, in the photojournalist case the newsroom may not be able to enforce CAI compliant capture due to software/hardware availability and legacy systems. Here, the lack of end-to-end CAI compliance could be addressed by having the newsroom itself vouch for the legitimacy of assets and add time and location as post-hoc CAI assertions. This can be accomplished if the newsroom creates claims through compliant editing software after capture. It is essential here that the organization's identity be captured in the signing process since trust is shifting from a chain of provenance with multiple identities to a single vouching entity.

These are just some of the many factors to explore to enable partial CAI workflows. Although those details are outside the scope of this paper they must be carefully considered in the design of a pragmatic CAI system.

5 System Overview

The proposed system is based on a simple structure for storing and accessing cryptographically verifiable metadata created by an entity we refer to as an actor. An actor can be a human or non-human (hardware or software) that is participating in the CAI ecosystem. For example: a camera (capture device), image editing software, or the person using such tools.

This metadata comprises information regarding asset creation, authorship, edit actions, capture device details, software used and many other subjects. There are standardized types common to most use cases as well as custom types, supported through extensibility. They are represented as claims and assertions as described in "Claims and Assertions". In short, assertions represent distinct pieces of information and claims wrap them into verifiable units. These pieces form the provenance of a given asset.

The CAI embraces existing standards. A core philosophy is to enable rapid, wide adoption by creating only the minimum required novel technology and relying on prior, proven techniques wherever possible. This includes standards for encoding, hashing, signing, compression and metadata.

© Adobe 2020

5.1 Claims and Assertions

5.1.1 Overview

As shown in "Workflows" above, each of the actors that create or process an asset will produce one or more **assertions** about what they did, when they did it, and (if possible) on behalf of whom. An assertion is typically a JSON-based data structure which represents a declaration made by an actor about an asset at a specific time. Some of these actors will be human and add human-generated information (e.g. copyright) while others are machines (software/hardware) providing information they generated (e.g. camera type or device time). Each type of assertion is either defined in the CAI specification, defined by other metadata standards such as XMP or schema.org or can be custom data for a particular actor or workflow.

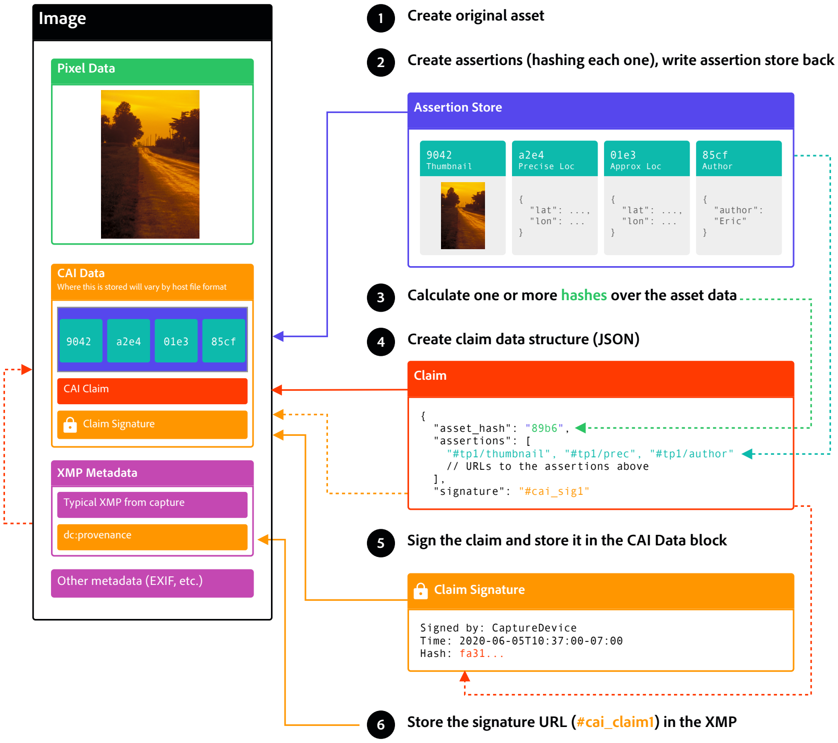
Assertions are cryptographically hashed and their hashes are gathered together into a **claim**. A claim is a digitally signed data structure that represents a set of assertions along with one or more cryptographic hashes on the data of an asset. The signature ensures the integrity of the claim and makes the system tamper-evident. A claim can be either directly or indirectly embedded into an asset as it moves through the life of the asset.

Each time the asset reaches a specific key point in its lifecycle, such as initial creation, completion of some editing operations, publication to social media, etc. a new set of assertions and a claim are created. Each new claim refers to the previous claim, thus creating a chain of provenance for the asset (see "Figure 1: Creating a Claim").

Because there are various workflows, some of which are more or less cloud-averse, assertions and claims are designed to live either embedded inside an asset or in the cloud or a combination of the two.

© Adobe 2020

5.1.2 Figure 1: Creating a Claim



1 Create original asset

2 Create assertions (hashing each one), write assertion store back to file

```
{
  "lat": ...,
  "lon": ...
}
```

```
{
  "lat": ...,
  "lon": ...
}
```

```
{
  "author":
  "Eric"
}
```

3 Calculate one or more hashes over the asset data

4 Create claim data structure (JSON)

```
{
  "asset_hash": "89b6",
  "assertions": [
    "#tp1/thumbnail", "#tp1/prec", "#tp1/author"
    // URLs to the assertions above
  ],
  "signature": "#cai_sig1"
```

5 Sign the claim and store it in the CAI Data block

```
Signed by: CaptureDevice
Time: 2020-06-05T10:37:00-07:00
Hash: fa31...
```

6 Store the signature URL (#cai_claim1) in the XMP

Diagram of claims and assertions embedded in an image file

5.1.3 Establishing Trust

One key component in establishing trust in the CAI system comes from the entities whose certificates are used for signing the claim. To ensure that only assets signed by trusted actors can be considered properly attributed, it is necessary to create a list of trusted certificates or their certification authorities (CAs).

Similar to the EU Trust List, the Adobe Approved Trust List, and similar lists used by web browsers and operating systems, the members of the CAI will establish their own Trust List of certificates that can be used to sign claims. Details on the governance of the Trust List is outside the scope of this paper.

In many cases, the holder of the certificate will not the individual who created (or edited) an asset, but instead will be the entity responsible for the hardware or software that they used. The signing certificate belongs to the actor (e.g. Truepic Camera, Adobe Photoshop, BBC, etc.) that performed the actions *on behalf of* someone else. This model allows CAI to provide

© Adobe 2020

anonymity (and/or pseudonymity) where desired. For scenarios where the certificate holder is able to reliably establish the identity of the individual, and the individual wishes their identity associated with an asset, an identity assertion is used.

5.1.4 Identity

One of the assertion types that can be present in a claim is Identity. This digital identity (also sometimes referred to as a *Subject* or an *Entity*) is present when an individual (or organization) is making a clear statement about their association with this claim.

Digital identity fundamentally requires digital identifiers — strings or tokens that are unique within a given scope (globally or locally within a specific domain, community, directory, application, etc.). In order to support a variety of use cases, including those where identity might be anonymous or pseudonymous, it is important that various schemes for the identifiers are available for use. Fortunately, most common identity formats such as Decentralized Identifiers-DID, WebIDs, OpenID, ORCiD and others are all based on URIs. This enables an identity assertion to be expressed in the standard format described in RFC 3986.

5.1.4.1 Decentralized Identifiers

Decentralized Identifiers are particularly well-suited to capturing identity in the CAI attribution system because they adapt well to different authentication scenarios. Although the name implies that a DID is to be used in a decentralized environment (commonly in conjunction with a blockchain) rather than a centralized one, there is nothing in that specification that prevents it. In fact, the specification itself not only calls that out, but points out this flexibility as a benefit of DID.

5.1.4.2 Non-URL formatted identities

Other standards used for identification that are not represented as URLs must be encoded as a URL in order to store it in an identity assertion. For example, ISNI.org recommends that the ISNI - ISO 27729 identifier be added to the end of a standard URL reference to their site like this: http://www.isni.org/isni/0000000114559647.

5.1.5 ClaimReview

CAI claims can be augmented with human-generated review assertions so that fact checking professionals can provide additional evidence of authenticity. We can leverage existing standard schemas for this.

One popular schema used for this purpose is ClaimReview which can be used in conjunction with MediaReview to add fact-checking reviews to images and other assets. By having a

© Adobe 2020

standard assertion type that contains a ClaimReview instance, these fact checks can now be embedded into the asset itself, enabling additional checks related to the context where the asset appears.

By embracing ClaimReview and other standards like it, the CAI will support fact checking with rich metadata to optimize verification workflows. This will help keep the fact checking ecosystem decentralized and diverse.

5.1.6 Redaction of Assertions

In many workflows, there is a need for assertions to be removed by subsequent processes, either because publishing the assertion would be problematic (e.g. the identity of the person who captured a video) or the assertion is no longer valid (e.g. an earlier thumbnail showing something that has since been cropped out). The CAI allows for the redaction of these assertions in a verifiable way that is also part of the provenance of the asset.

In the process of redacting an assertion, a record that something was removed is added to the claim. Because each assertion's reference includes the assertion type, it is clear what type of information (eg. thumbnail, location, etc.) was removed. This enables both humans and machines to apply rules to determine if the removal is acceptable.

NOTE: Assertion redacted only applies to assertions that are part of the CAI data. It does not have anything to do with removal of other metadata (XMP, EXIF, etc.).

5.2 Use of XMP

XMP, as defined in ISO 16684-1, is the standard for embedded metadata in numerous asset types. The CAI leverages its standard rules to ensure that the CAI claim can be reliably retrieved using existing technology including a variety of open source tooling.

5.3 Composed Documents

It is very common that a content author will integrate other assets (e.g. stock art/photos) into the work that they are creating. In the world of dynamic media, such as video and audio, this is the normal mode of operation where various clips are combined to form the final production. These scenarios produce what are called *Composed Documents* as described in the "pantry and ingredient" model of XMP. (See the Partner Guide to XMP for Dynamic Media.)

To fully understand the complete history and attribution of an asset that has been created as part of a composed document, it is necessary to include or reference each ingredient along

© Adobe 2020

with any claims and assertions made on them. The CAI provides for this with a specific type of assertion that references each ingredient's claims and assertions, whether they are embedded into the new composed document or stored in the cloud.

5.4 Digital Signatures

In order to ensure the integrity of the CAI information that is embedded into an asset, including the asset's data itself, digital signature technology is employed. That same technology also provides an authentication mechanism connected to the signer, that provides a way to establish trust in the actors involved in the CAI-enabled workflow.

Because the data being signed is, in signature terms, "arbitrary message content," a standard Cryptographic Message Syntax (CMS) signature is used to sign the asset. While a simple CMS signature is sufficient, CAI recommends using a CMS Advanced Electronic Signature (CAdES) instead. Use of a CAdES-compliant signature will ensure that the CAI data complies with the European eIDAS legislation.

It is also recommended that the signer should use a trusted timestamp authority to generate a trusted time-mark or time-stamp token. This is for proving that the signature itself actually existed at a certain date and time and can be incorporated into the CAdES content-time-stamp attribute to create a CAdES-T compliant signature.

5.5 Illustrative Flow

A brief, high-level sequence of events for a typical cycle of claim creation and verification is useful to convey how the system operates. This is illustrative, not exhaustive and could apply to any of the "Workflows" above. All data capture is optional; when we refer to identity or location, for example, it is important to note that whether to record such data is a decision of the user. In addition, some of the information in this example is shown embedded in the asset while other information is in the cloud. The choices here are purely for example since any assertion or claim can be stored either in the asset or in the cloud.

1. A capture device captures an image and concurrently creates a set of assertions about the image such as capture location, equipment details, identity of the user, and perhaps a representative thumbnail. These assertions are embedded in the asset.
2. A claim, referring to the set of assertions and including a hash of the image, is created and cryptographically signed by a trusted signing authority on behalf of the user. The claim is then embedded into the image and a reference to it stored in the asset's metadata (see "Figure 1: Creating a Claim").

© Adobe 2020

3. The image, with its attached claim, is imported into an editing tool. The prior claim is verified (see "Figure 2: Validating Claims"), such that upon successful verification its assertions and claims can be carried forward as the image's history is accumulated.

4. As edits are made, additional assertions are captured and stored in a cloud service. Upon image export, these are gathered together as part of a second claim along with the hash of the updated image. A URL to the claim is stored in the asset's metadata. This new claim (which is also stored in the cloud) refers to the prior claim, therefore ensuring that all assertions from both claims are accessible via the exported image without any link to the prior version of the image.

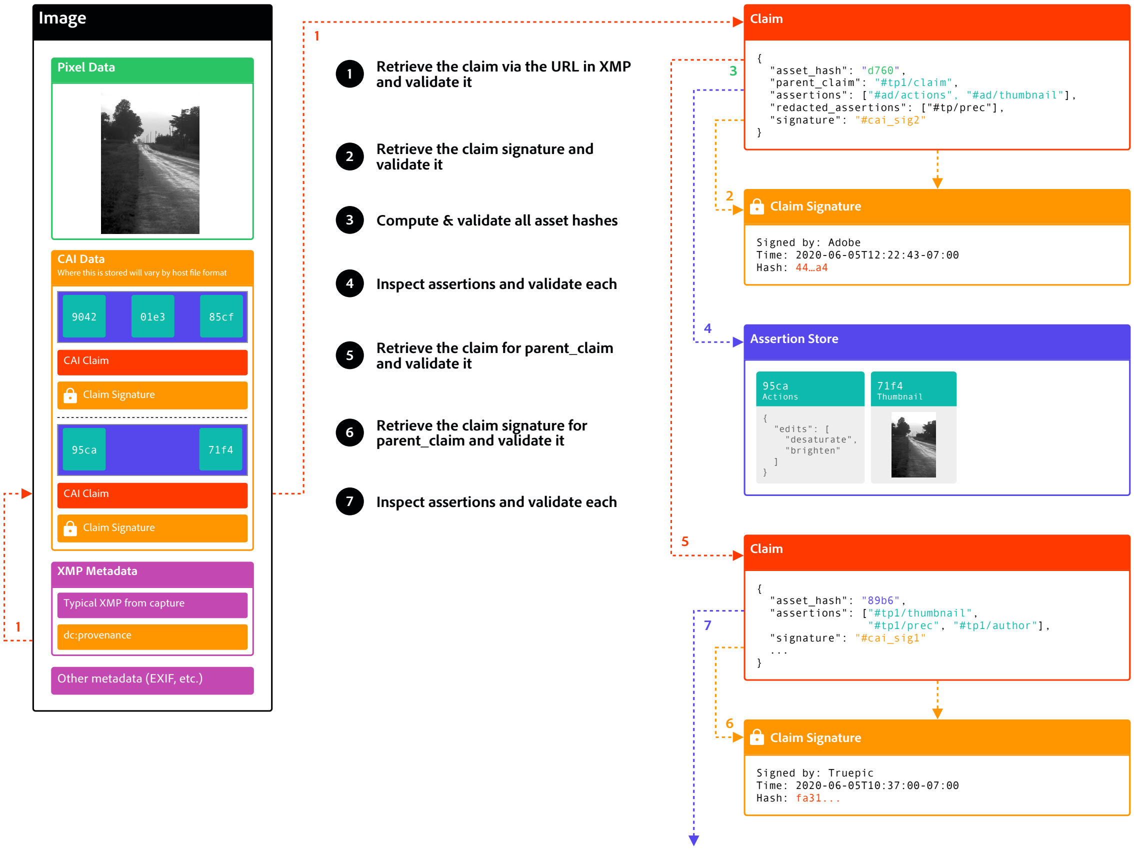
5. The edited image, with its CAI metadata attached, is shared on a platform.

6. A content consumer encounters the image on the Web via a CAI-enabled viewing experience and sees a visual indication that CAI claims are present.

7. Finally, the content consumer interacts with the visual indicator to learn about the image's history. The most recent claim is retrieved and verified along with the entire chain of claims and their component assertions (see "Figure 2: Validating Claims"). The assertions are displayed to the viewer in a clear, time-ordered user experience depicting the image's claim history from inception to display.

© Adobe 2020

5.6 Figure 2: Validating Claims



1. Retrieve the claim via the URL in XMP and validate it
2. Retrieve the claim signature and validate it
3. Compute & validate all asset hashes
4. Inspect assertions and validate each
5. Retrieve the claim for parent_claim and validate it
6. Retrieve the claim signature for parent_claim and validate it
7. Inspect assertions and validate each

```
{
  "asset_hash": "d760",
  "parent_claim": "#tp1/claim",
  "assertions": ["#ad/actions", "#ad/thumbnail"],
  "redacted_assertions": ["#tp/prec"],
  "signature": "#cai_sig2"
}
```

```
Signed by: Adobe
Time: 2020-06-05T12:22:43-07:00
Hash: 44…a4
```

```
{
  "edits": [
    "desaturate",
    "brighten"
  ]
}
```

```
{
  "asset_hash": "89b6",
  "assertions": ["#tp1/thumbnail",
                 "#tp1/prec", "#tp1/author"],
  "signature": "#cai_sig1"
  ...
}
```

```
Signed by: Truepic
Time: 2020-06-05T10:37:00-07:00
Hash: fa31...
```

Validation of Claims and Assertions

6 User Experience

There are two types of user experience (UX) each CAI implementor may need to carefully consider, depending on the goal of the implementation. First, that of creators using the implementor's tools, and second that of consumers viewing content on the implementor's platform.

Both types of experiences should optimize for clarity and provide guidance for users who may have questions about what they are seeing. Above all, experiences should default to the use of consistent terminology and iconography in order to achieve this so that over time expectations of what the CAI provides and its realized value are precisely aligned.

For content creators, it is important to ensure that data sharing via CAI claims is well understood by the user and that publishing CAI claims is not an automatic process. While flows may benefit from a simplified UX, requiring the user to make intentional decisions

© Adobe 2020

about what kind of CAI data to record will prevent unintended claim capture. It is strongly suggested that creator tools support previewing claim content before it is signed and attached.

Content consumers will be best served by a different set of UX principles to help them make decisions about what to trust through the presentation of claim evidence. For that reason, it is important that if a visual indication is associated with a valid asset, it should indicate only that CAI data is present and verified, and should *not* appear to indicate whether the content is authentic. In cases where the asset and its CAI data do not match, this should be clearly indicated as well.

For the purposes of this discussion, platforms on which content consumers view content can be native desktop or mobile applications, web sites and web browser integrations.

Since CAI data can contain a depth of detail that is not always relevant for all viewers, we envisage a UX based on the idea of progressive disclosure. This means users are presented a small amount of critical data up front, then empowered to reveal more detail by interacting with the user interface. Which information is most critical for users will have to be carefully evaluated for each situation, but in general it is recommended that the user see when an asset was modified, how it was modified and by whom as top-level information. This model also helps to provide a solution suitable for diverse levels of digital literacy.

There will often be multiple claims for an asset. Implementors should think carefully about how to display them so viewers are provided the simplest and fastest possible path to decisions about what to trust. In some cases a straightforward list of assertions in reverse chronological order might be appropriate. However, a clear visual classification of assertions based on the type of actor that created them (hardware device, software program, human fact checker, etc.) could be a very powerful way to help users decide which assertions matter most in a given context.

Not all information in CAI claims is the same. Each has different dependencies and potential vulnerabilities. For instance, on-device camera information (such as lens used) is different from creation date or location information which is dependent on an external signal (e.g. clock or GPS). An optimal UX for viewers will indicate this, through progressive disclosure, without overly complicating the experience.

7 Security, Trust & Privacy Considerations

7.1 Educating the User

The presence of the CAI information in an asset means that there is evidence about the asset that the user can use to make their own determination of trust. For a user to fully understand

© Adobe 2020

what CAI does (and does not) achieve, it is important that they understand how CAI works, what the disclosed information means, and more. This information needs to be useful but not overwhelming. These considerations apply to both the people who are creating and editing content, who will need to understand why they need to enable it for their assets *but* also how consumers will understand the information to consider it worthwhile.

One key aspect of the user experience presented to users (and perhaps the whole CAI ecosystem) that should be called out specifically is the implication of "trust" in the system. It is extremely important that a consumer of assets not interpret more trust in the presence of valid CAI information than it truly means — specifically that an asset with valid CAI information does not imply anything about the trustworthiness of the content of the asset.

This education is key in preventing social engineering attacks.

7.2 Unexpected Disclosure

In cases where adding identity (or any other form of identifying attribution) to an asset at the time of capture could lead to increased risk to content producers or others, it is important that assertions and claims can be added at some later time. The CAI assertion and claim model allows for any type of assertion to be added at any time. In addition, the redaction capability supports the inverse case where *too much* information was added and some needs to be removed.

Of course, once an asset has been released "into the wild" the information contained inside cannot be modified there, as control has been relinquished. When assertions are stored in the cloud, it may be possible to have those assertions removed or made inaccessible by the hosting provider, but it may not be possible to track down and remove all copies of that data cached by third parties.

7.3 Certificate Trust

A number of potential concerns arise from attacks against the certificates used to sign the claims of an asset. To address these concerns, the CAI ecosystem is built around its own CAI Trust List for "Establishing Trust". This provides CAI with a well-defined model for allowing only certain approved signers and managing the lifecycle of those certificates (e.g. revocation, expiration, etc.).

Another design goal of the CAI is flexibility in certificates used for signing claims. In most cases, signing certificates tied directly to an individual should not be used, but instead organization certificates (e.g. Adobe, Twitter, etc.) or even one time private keys will be used to sign *on behalf of* the actual user performing the actions on the asset. This is recommended

© Adobe 2020

due to the inherent complexity required of users to manage their own keys. Using organization certificates also has the important advantage that the domain of the certificate can be matched to URLs in the claim data, thus preventing link manipulation attacks.

Implementors should take appropriate precautions to ensure that their signing keys cannot be stolen or compromised. Use of a cloud service, hardware enclave and use of a One Time Private Key are some possible approaches that will be taken.

7.4 Distributed Ledger Technology

The use of a Trust List is the proposed model for early CAI implementations, but it is not the only possible approach to underwriting trust.

Distributed Ledger Technology (DLT) offers a consensus model for replicated, shared and synchronized data.

While other approaches to secure content attribution have investigated the use of such technology, it was felt that mandating a single ledger for all authored content may not be globally scalable for the CAI and may be at odds with the spirit of decentralization. Implementors in the CAI ecosystem might opt to use a DLT to federate their storage of assertions and claims to achieve an additional level of integrity and transparency.

Having a distributed ledger as a secondary model for storing information about issued claims (e.g. their hashes) could serve as proof that data stored in a given provider's cloud has not been modified or tampered with (either intentionally or unintentionally).

7.5 Intentional Misattribution

The CAI attribution model does not prevent a malicious user from stripping all of the CAI data (claims and assertions) from an asset and then adding new claims representing themselves as the originator. Similarly, the "analog hole" or "rebroadcast attack," which are common terms for subverting provenance systems by capturing an image of a photograph or computer screen, are not addressed directly by the model.

However, there are some solutions that can be implemented in concert with the CAI model to achieve resilience against intentional misuse.

- An actor could use watermarking technology to durably embed information (either perceptibly or imperceptibly) about the asset's current claim. The watermark could be subsequently used to recover provenance data.
- A camera device or software could utilize depth mapping to capture scene information (as CAI assertions) which would indicate whether a photograph depicts a 3D scene or a rebroadcast photo of a photo.

© Adobe 2020

- It is possible for systems using “similarity detection” to help users find additional information about an asset. For example, users could be shown whether the asset was published at some prior date or the asset could be contextualized by surfacing substantially similar assets.
- Trusted timestamps can be used to cast doubt on assets with deliberately altered histories. For example, when identical assets with different claim histories are encountered, the earlier CAI claims are likely to be trustworthy while the later ones may represent an attempt to alter history.
- Identity, actions and other assertions that include information about specific domains or organizations could be compared against the domain associated with the signing certificate. This would, for example, prevent the use of a signing certificate from “badsoftware.com” claiming that the user was using “Adobe Photoshop” to edit the image.

While not part of the core CAI infrastructure, such solutions easily integrate with CAI and provide great utility. These techniques support the application of judgment and reason — they are not technological guarantees.

8 Future Work Streams

Current and future CAI collaborators will focus on several work streams with the goal of incorporating diverse points of view while stewarding the ideas expressed here toward a unified, pragmatic, adoptable standard.

8.1 Establishment of Working Groups

To move from high-level system concepts to detailed specifications, several Working Groups have been created. Working Groups are open to any interested organization or individual. When appropriate, a Working Group will produce one or more specifications for peer review and publication.

8.2 Extension of Design for Additional Formats

Many details have yet to be proposed for ensuring the system can embrace time-based media like audio, video, and streaming formats. While early work has been done to ensure the design does not preclude these, it is important for the collaborators to focus on them next. This work would be done in conjunction with experts in the field — not only about the formats but also about common workflows involving them.

© Adobe 2020

8.3 Prototype Exploration

With the goal of understanding the practical considerations for implementing CAI-compliant devices and applications, we expect that many prototypes will be created and tested in parallel with Working Group efforts. We define success as wide adoption of carefully vetted specifications. To achieve this, exploratory proofs of CAI concepts will be built by CAI collaborators and learnings shared with the community.

9 Conclusion

The collaborators on this paper have explored the challenges of inauthentic media through problem definition, system design and use case research. The results of the exploration are expressed in the design of the CAI provenance system. To achieve widespread adoption we have based the design on existing standards and established techniques, and acknowledge that the system will need to include simple and intuitive user experiences.

However, even an optimally designed system cannot ultimately succeed in a vacuum. We now begin the important work of deeper, more expansive collaboration with leaders in technology, media, academia, advocacy and other disciplines.

With this first step towards an industry standard for digital content attribution, we look optimistically to a future with more trust and transparency in media.

10 References

- JSON
- eXtensible Metadata Platform (XMP)
- Partner Guide to XMP for Dynamic Media
- Decentralized Identifiers-DID
- WebIDs
- OpenID
- Cryptographic Message Syntax (CMS)
- European eIDAS legislation

© Adobe 2020

11 Contributors

This paper would not have been possible without the substantial contributions of these individuals.

- Will Allen (Adobe)
- Pia Blumenthal (Adobe)
- John Collomosse (Adobe)
- Oliver Goldman (Adobe)
- Andrew Kaback (Adobe)
- Gavin Peacock (Adobe)
- Charlie Halford (The British Broadcasting Corporation)
- Scott Lowenstein (The New York Times Company)
- Thomas Zeng (Truepic)
- Fabiana Meira Pires De Azevedo (Twitter)
- Corin Faife (WITNESS)

© Adobe 2020