WITNESS

SEE IT FILM IT CHANGE IT

GETTING IT RIGHT:

PROVENANCE AND AUTHENTICITY INFRASTRUCTURE THAT WORKS FOR ALL

explainer

Table of Content

Introduction

The Coalition for Content Provenance and Authenticity (C2PA) has released version 1.0 of its technical specifications to establish **a common technical standard that would enable showing a range of information about how, where and by whom a piece of media was created, and how it was subsequently edited, changed and distributed.**

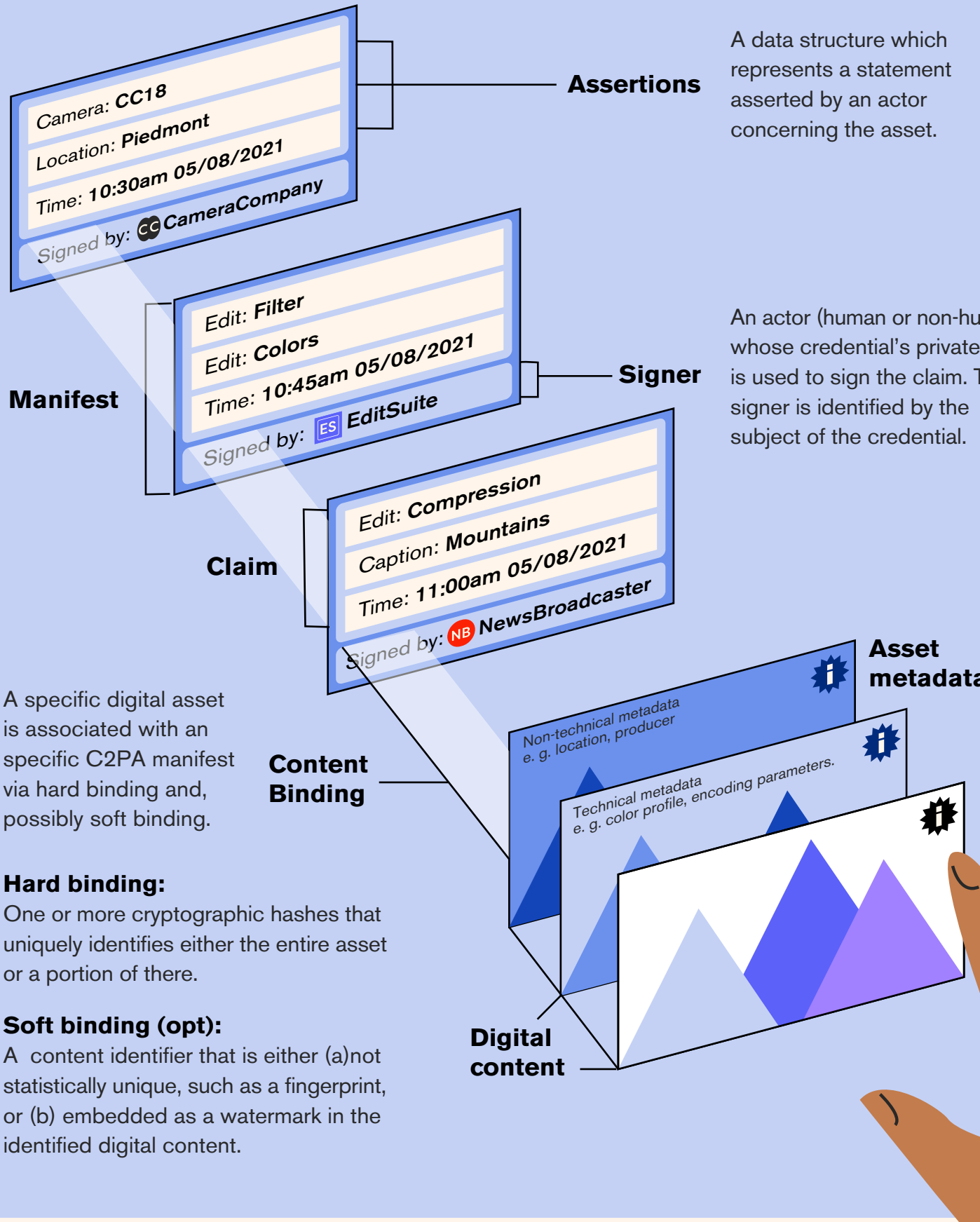
WITNESS has been part of these efforts to advocate for an approach that empowers critical voices and reflects human rights and privacy concerns. Our work has been informed by collaborating on tools like ProofMode with The Guardian Project, by our report 'Ticks or It Didn't Happen' that pinpointed fourteen key issues that need to be considered at an early stage of the development of this infrastructure, as well as by a series of global online convenings with human rights defenders, activists, community journalists, civil society and industry experts focused on identifying potential harms.

This explainer includes an overview of how the C2PA works and its potential use cases for human rights defenders, activists and civic journalists. It highlights the potential harms we must avert and mitigate in the design, development, use, oversight and regulation of the standards, its tools and the overall authenticity and provenance ecosystem.

C2PA:
How it Works

Adding verifiable indicators of authenticity

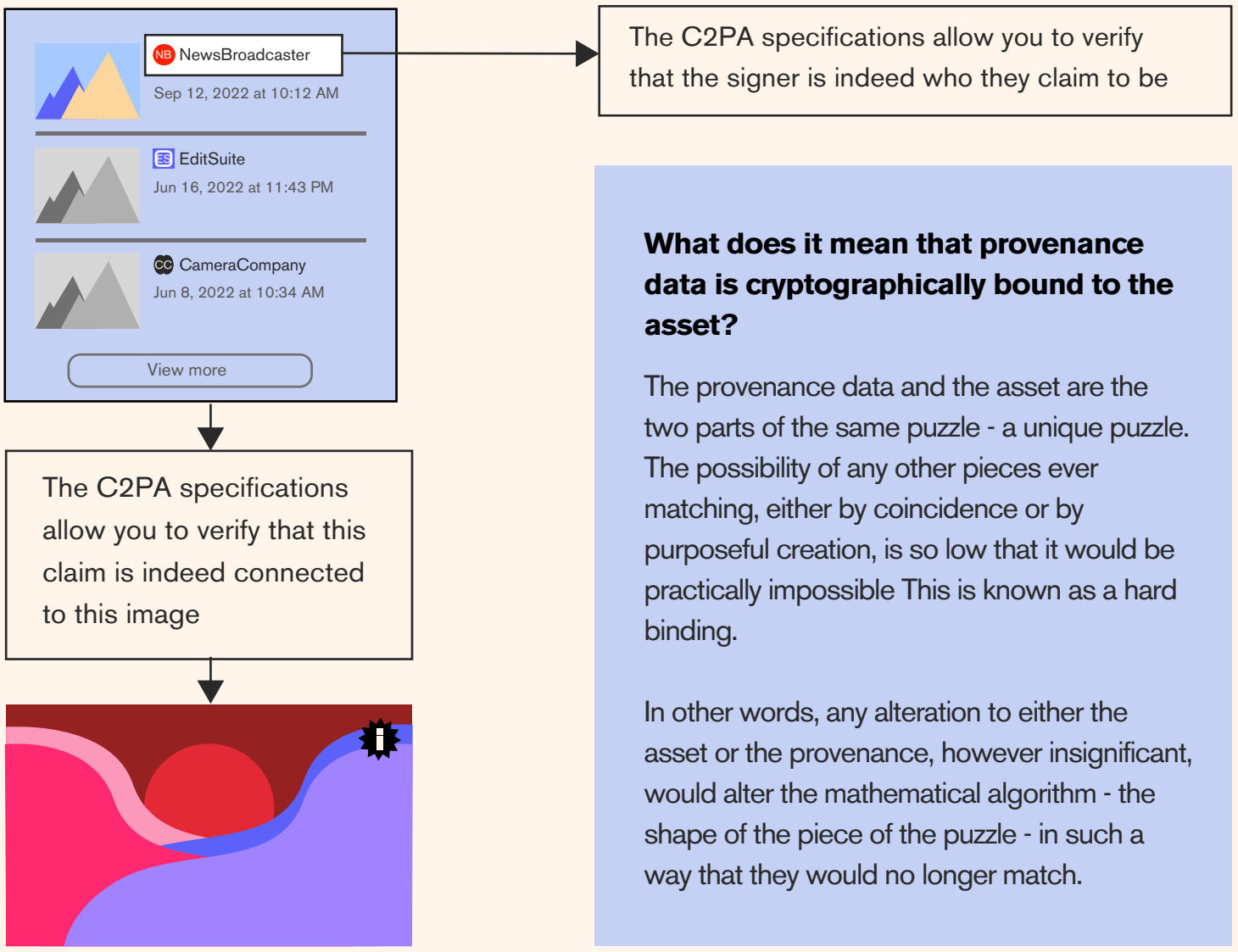
The standards have designed a mechanism to attach technically verifiable metadata (‘manifests’) to digital media. This ‘metadata’ would make up the provenance information about the digital asset, such as when and where the picture was taken, or if it was edited, and if so how and by whom.



Hard binding:
One or more cryptographic hashes that uniquely identifies either the entire asset or a portion of there.

Soft binding (opt):
A content identifier that is either (a)not statistically unique, such as a fingerprint, or (b) embedded as a watermark in the identified digital content.

Technicalities aside, this is what the C2PA specifications are designed to do



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 purposeful creation, is so low that it would be practically impossible This is known as a hard binding.

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.

Trust in the C2PA

The tools and technologies that we use do not establish trust. The C2PA specifications are designed to play into existing relationships of trust. It could be said that, **if you are a media consumer, you may be inclined to trust content if the signer of a valid manifest is a person or entity that you trust.**

Information to help you rather than a confirmation of trust or truth

Just because a digital asset has a valid C2PA manifest does not mean that it is to be trusted, or that its underlying contents are true.

Similarly, the fact that any digital asset does not have C2PA manifests does not mean that its contents are not to be trusted, or that they are false.

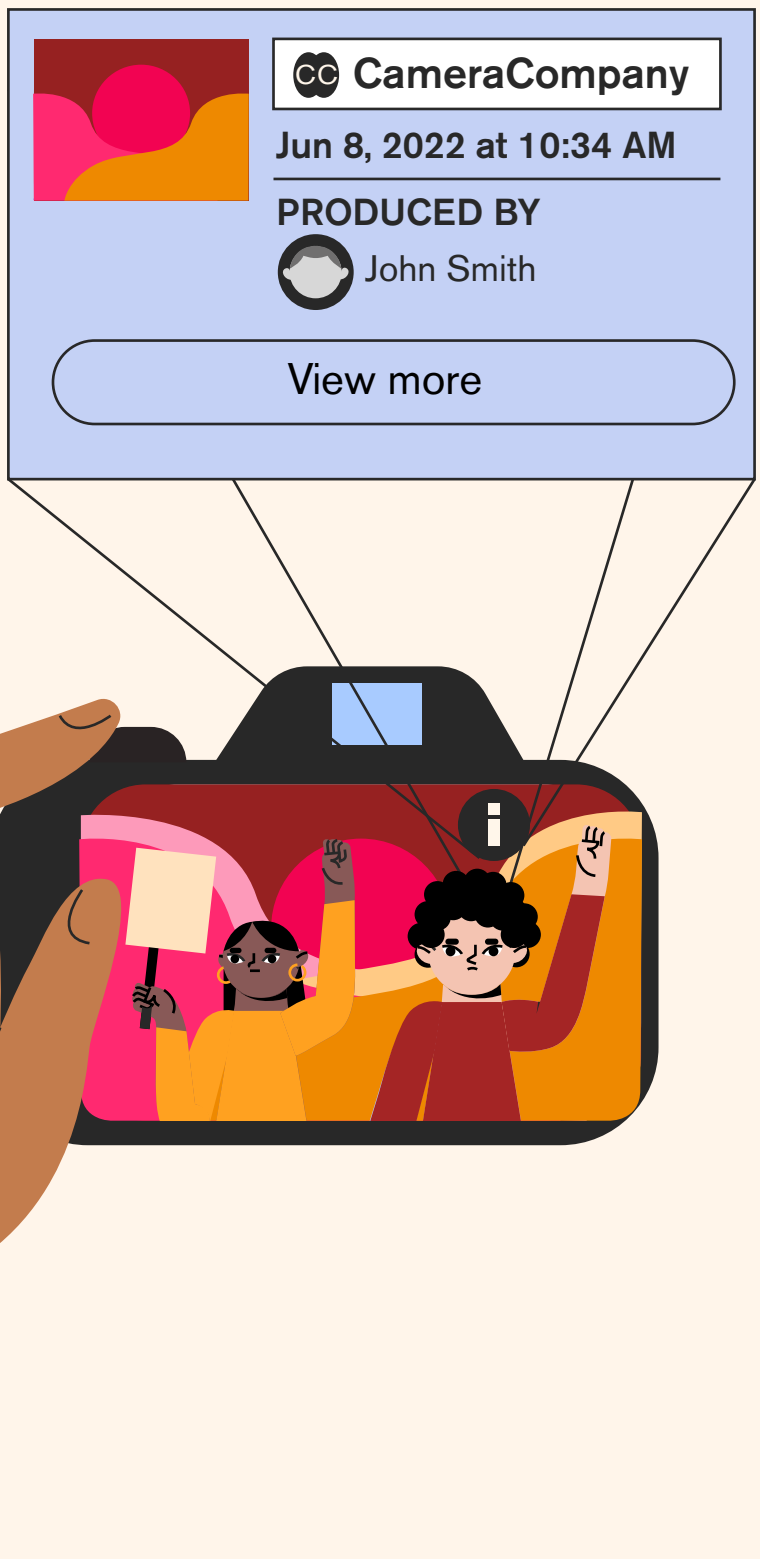
A preexisting relationship of trust between a signer and a content consumer is the basis of the C2PA trust model. This may be immediately beneficial for some mainstream news media organizations, companies and recognized individuals, but not necessarily for smaller news outlets, organizations or less recognized individuals.

Understanding some of the key actors in the C2PA ecosystem

Signer:

The signer could be an individual or an organization, and it could be human or non-human (e.g. an algorithm that automatically signs a manifest).

According to the C2PA specifications, the signer ultimately determines what information goes in a manifest and what does not. The signer could, for example, decide to strip previous manifests from a digital asset, or it could decide to add a lot (or a little) provenance information.



Content creator:

The content creator is the actor (human or non-human) that creates a digital asset. For example, it could be a photographer, or a digital artist, or an AI system such as Dall-e.

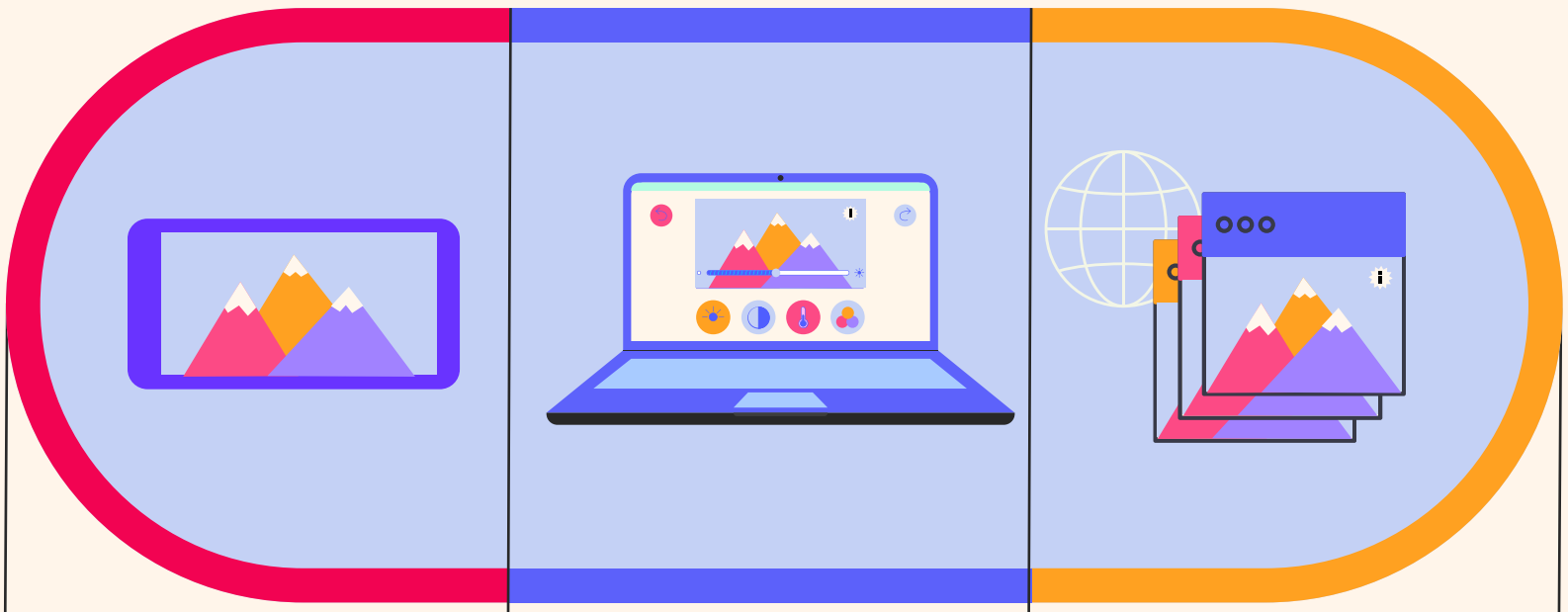
In some cases, the signer and the content creator could be the same, in others they may not. It is important to note that the C2PA recommends that the content creator have effective control over their information but certain implementations may decide to restrict this.

Validator:

The validator is the program that verifies the identity of the signer and the connection between the digital content and its provenance information. In other words, the validator verifies that the cryptographic hash is not broken. If the identity of the signer cannot be verified, or if anything has changed in the digital content or the provenance information, the validator will mark an asset as invalid.

Facilitating cross-workflow interoperability

The standards are open which means that any organization or company may implement it into their tool or service so that this provenance information can be created and tracked throughout the life cycle of digital media.

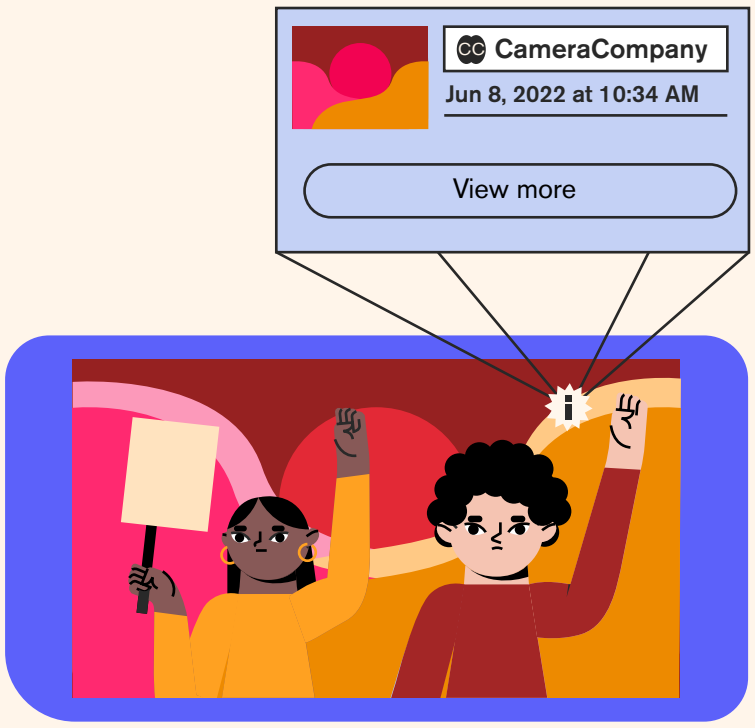


Phase 1:
Raw footage claim

Phase 2:
Editing

Phase 3:
Publishing and sharing

Phase 1
Raw footage claim



The C2PA workflow can start at the point of creation of a digital asset, when a camera clicks or when OpenAI's Dalle-E 2 generates an image, for example. In the case of the camera, the C2PA specifications could potentially be implemented into a camera's hardware or firmware, or it could be connected to a third-party app or service, as is the case with ProofMode.

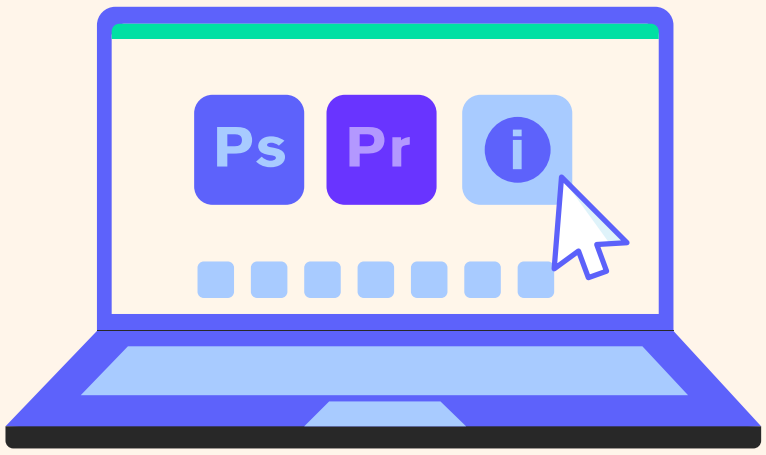
In this phase, the provenance of media could be used to claim that the content is raw - that it hasn't been edited in any way.

Enhancing credibility of on-the-ground evidence with raw footage claims

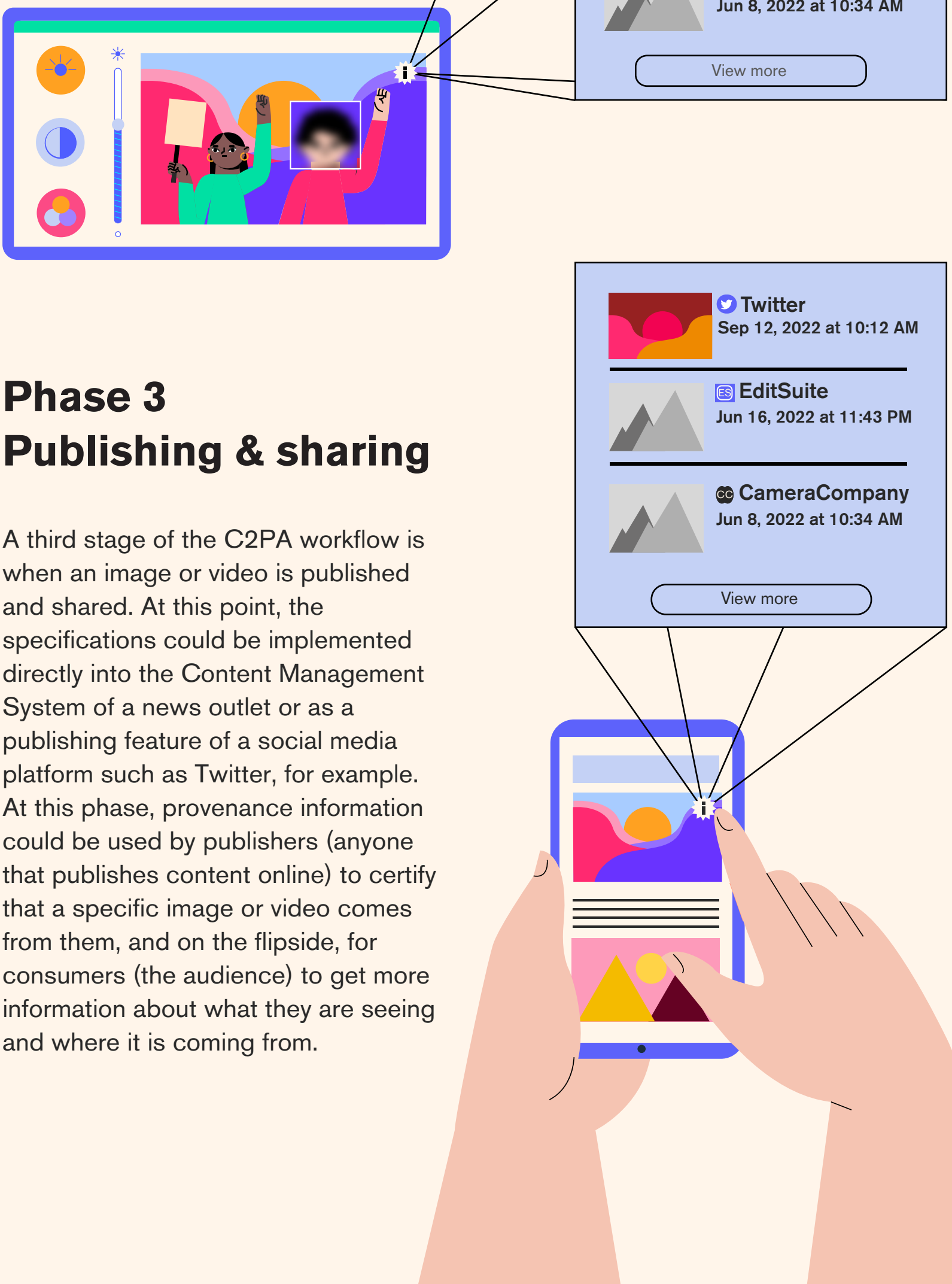
Video or image evidence can be dismissed or undermined by claims suggesting that the content is fake or misleading. With easier access to more and better tools to create synthetic media (including deepfakes), activists, civic and community activists and other grassroots defenders face added pressure to authenticate their media. Provenance and authenticity tools and infrastructure can empower but also hinder their efforts.

Phase 2
Editing

A second phase of the C2PA workflow is when digital media is edited. In this phase, the C2PA specifications could be implemented into an image and video editing software, such as Adobe's Photoshop or Premiere, or directly into a 'claim generator' that can be used to attach new manifests to the digital media.



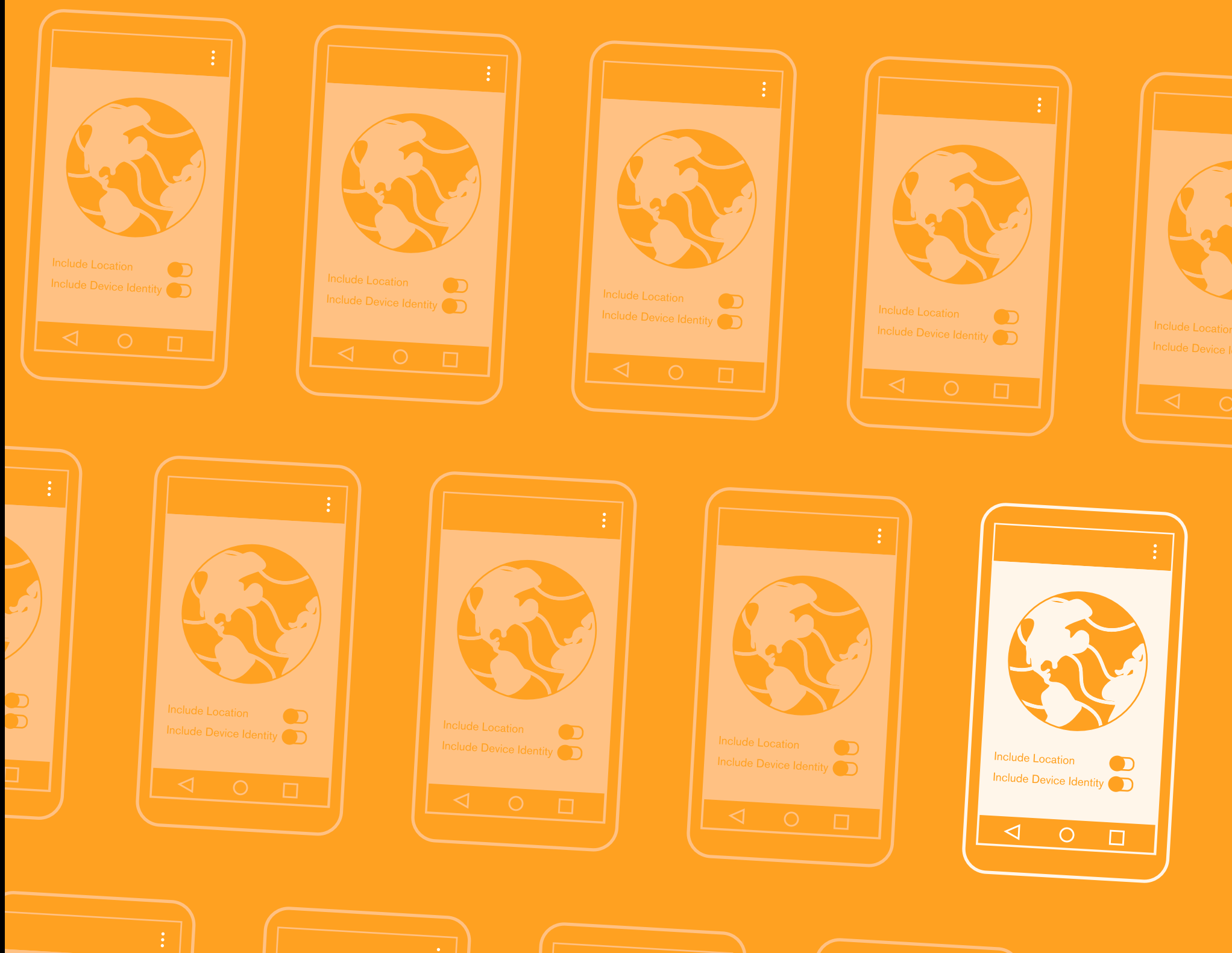
While there are many reasons to edit an image or video, in the context of human rights violations, this may be necessary to guarantee the privacy and wellbeing of those at risk. Editing can include generating a new privacy-sensitive manifest (e.g. redacting location from the metadata), or altering the actual content of the digital asset (e.g. blurring a face).



Phase 3
Publishing & sharing

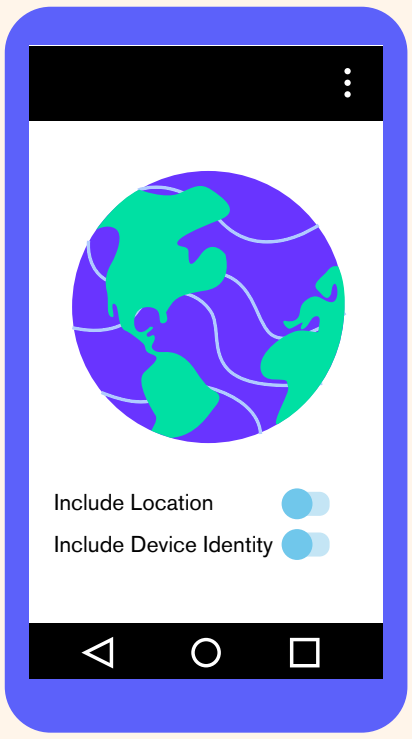
A third stage of the C2PA workflow is when an image or video is published and shared. At this point, the specifications could be implemented directly into the Content Management System of a news outlet or as a publishing feature of a social media platform such as Twitter, for example. At this phase, provenance information could be used by publishers (anyone that publishes content online) to certify that a specific image or video comes from them, and on the flipside, for consumers (the audience) to get more information about what they are seeing and where it is coming from.

From niche to widespread use of provenance and authenticity tools



Provenance and authenticity tools were pioneered by human rights organizations, such as The Guardian Project's ProofMode as a way to add credibility to the media coming from those that need it the most, those that often depend for their lives on the integrity and veracity of images they share from conflict zones, marginalized communities and other places threatened by human rights violations.

Efforts such as the C2PA, and parallel initiatives such as CAI and Project Origin are now making a push towards a more systemic use of provenance and authenticity tools. Still, it is not clear when, or if, this provenance and authenticity infrastructure will start to be widely used. At this point, we know that companies leading these efforts such as Adobe and Microsoft, are already implementing them into their tools and services, and that many other companies have at least shown interest in adopting the system.



Provenance and authenticity tools coming from and for human rights organizations are opt in, and focused on guaranteeing accessibility and privacy.

Checking interests

Although these specifications offer mechanisms to fortify truth, they are primarily developed by major technology and news media organizations who also pursue their own objectives. It is imperative that civil society globally be part of these efforts to help shape its design and deployment, and to push back against developments that may be harmful to societies.

Adobe arm BBC Microsoft intel SONY Truepic & others

Identifying and mitigating potential harms

WITNESS has led the Threats and Harms taskforce of the C2PA to assess the specifications for their potential to be misused, abused or cause unintentional harm. The details of this harms modeling exercise can be found here, and the complete list of identified potential harms along with existing and potential mitigation strategies can be found here.

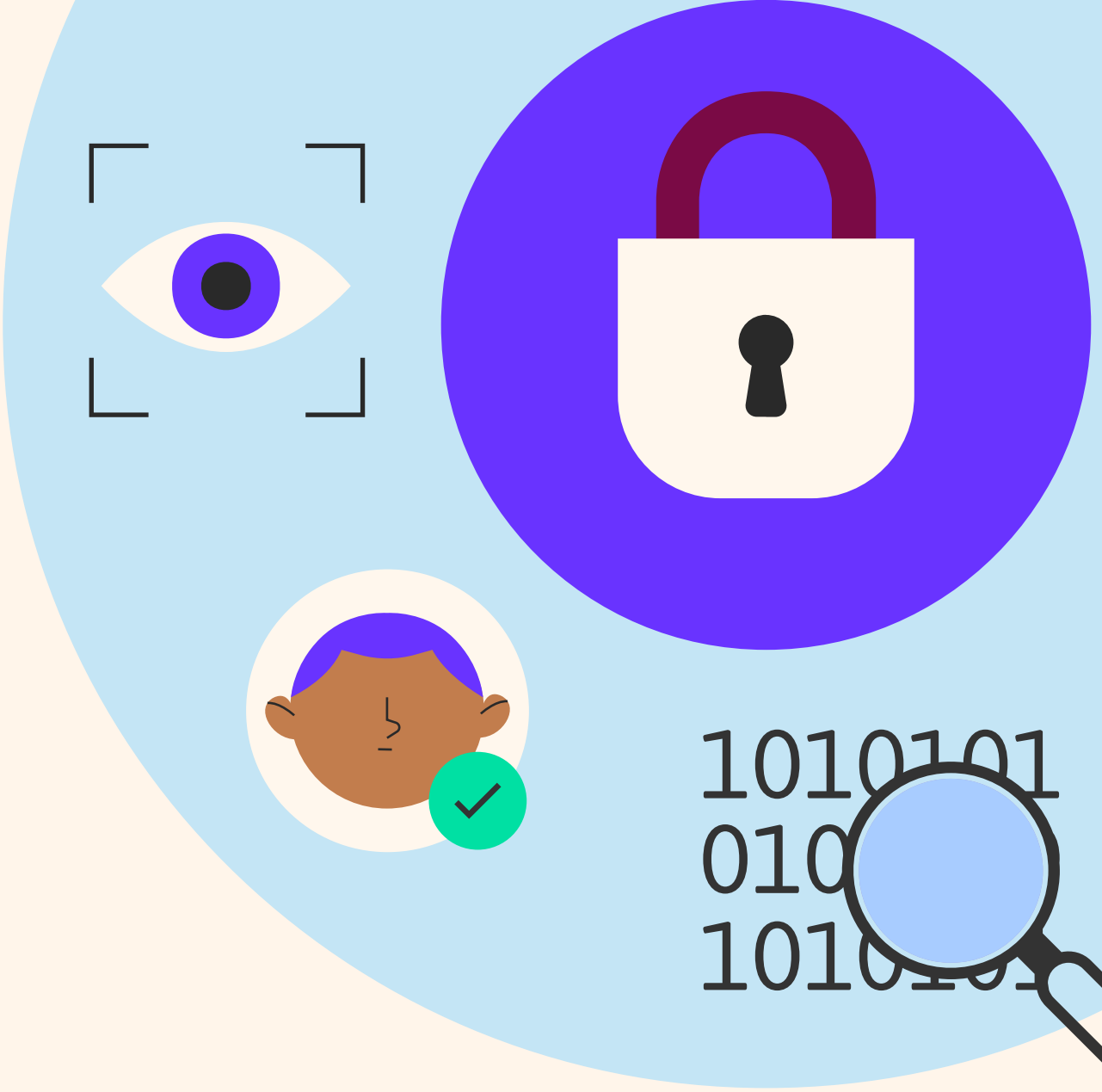
We highlight here some of the main concerns that were repeatedly raised during the feedback sessions with global stakeholders.

Government abuse and misuse

Specific tools and implementations can still opt to demand individual identifiers to be attached to provenance claims. It is technically possible, for example, for a regulation to require all journalists to use C2PA-enabled tools in order to attach their credentials to any media posted online. This is particularly concerning in countries with no checks on government surveillance and control, and where activists and journalists are already vulnerable to a lack of privacy and anonymity.

Ineffective user experience

One thing is what the C2PA is being designed to do, and another is what it actually does or what people perceive it to do. For example, what should a claim-generating platform interacting with content creators include (and how) in order to ensure that they retain effective control of their information? And on the flipside, what should consumers see when viewing media that has a provenance claim? Too little information may say nothing, too much information may be glossed over. The User Experience taskforce of the C2PA has been focusing on some of these questions, but the implications of specific designs within a broader provenance ecosystem remain uncertain.



Further privacy concerns

For those creating content, there is a possibility that they may inadvertently share sensitive information. It is therefore necessary that provenance tools include features that allow users to redact information from manifests.

Lack of access worsens social and economic inequalities

What happens if provenance and authenticity tools begin to be widely used? How will this affect those that chose to not use, or do not have access to, C2PA-enabled tools? There is a risk that content that does not include provenance information may be undermined or dismissed, further disenfranchising those that are already vulnerable due to a lack of access or technical know-how.

Getting It Right: Provenance and Authenticity Infrastructure that works for all

Moving forward, WITNESS is advocating for funding and supporting a diverse C2PA ecosystem that meets the needs of global users, particularly those of human rights defenders, civic journalists, grassroots activists and marginalized communities. We also aim to push back against legislative and political initiatives that misuse and abuse the specifications.

How you can participate

You can participate directly within the C2PA, or you can reach out to WITNESS to participate in the ongoing Harms, Misuse and Abuse Assessment, or to collaborate in the development of tools that cater to key users globally.