WITNESS
SEE IT FILM IT CHANGE IT

Harms Modelling in the C2PA

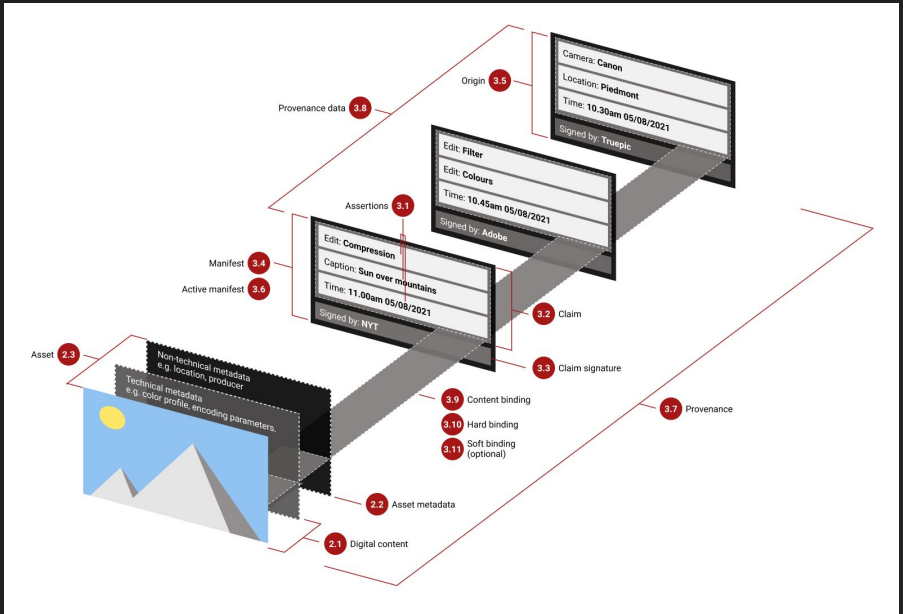
Coalition for Content Provenance and Authenticity

Jacobo Castellanos Rivadeneira
Technology Threats and Opportunities at WITNESS
jacobo@witness.org
@JacoboCas

Agenda

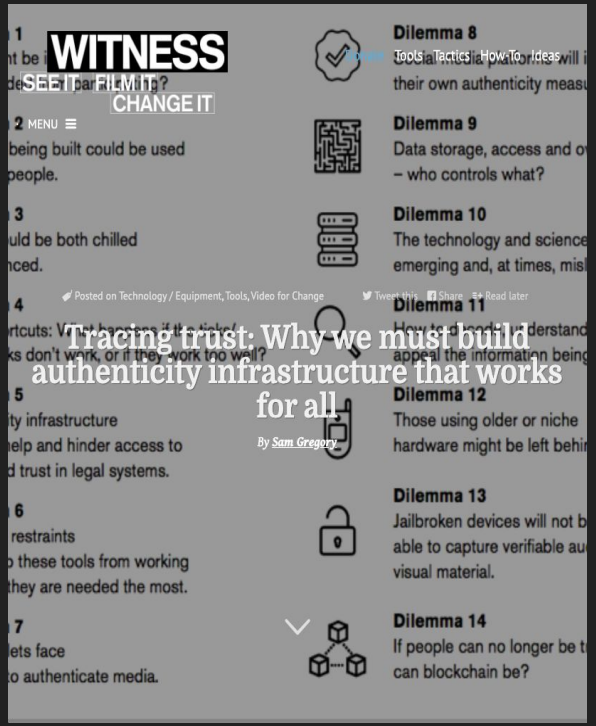
1. Background on WITNESS and our work on Provenance and Authenticity Infrastructure
2. Overview of the C2PA (Coalition for Content Provenance and Authenticity)
3. Harms Modelling in the C2PA

Provenance and Authenticity (P&A) Infrastructure



Provenance and Authenticity Infrastructure refers to the tools, services or frameworks that facilitates capturing, processing and presenting information about the source and history of digital assets in a way that is verifiable and tamper-evident.

WITNESS work on **Authenticity Infrastructure:** identifying values and importance, highlighting trade-offs



Other resources:

- Synthetic Media Lab

From niche to systemic P&A Infrastructure

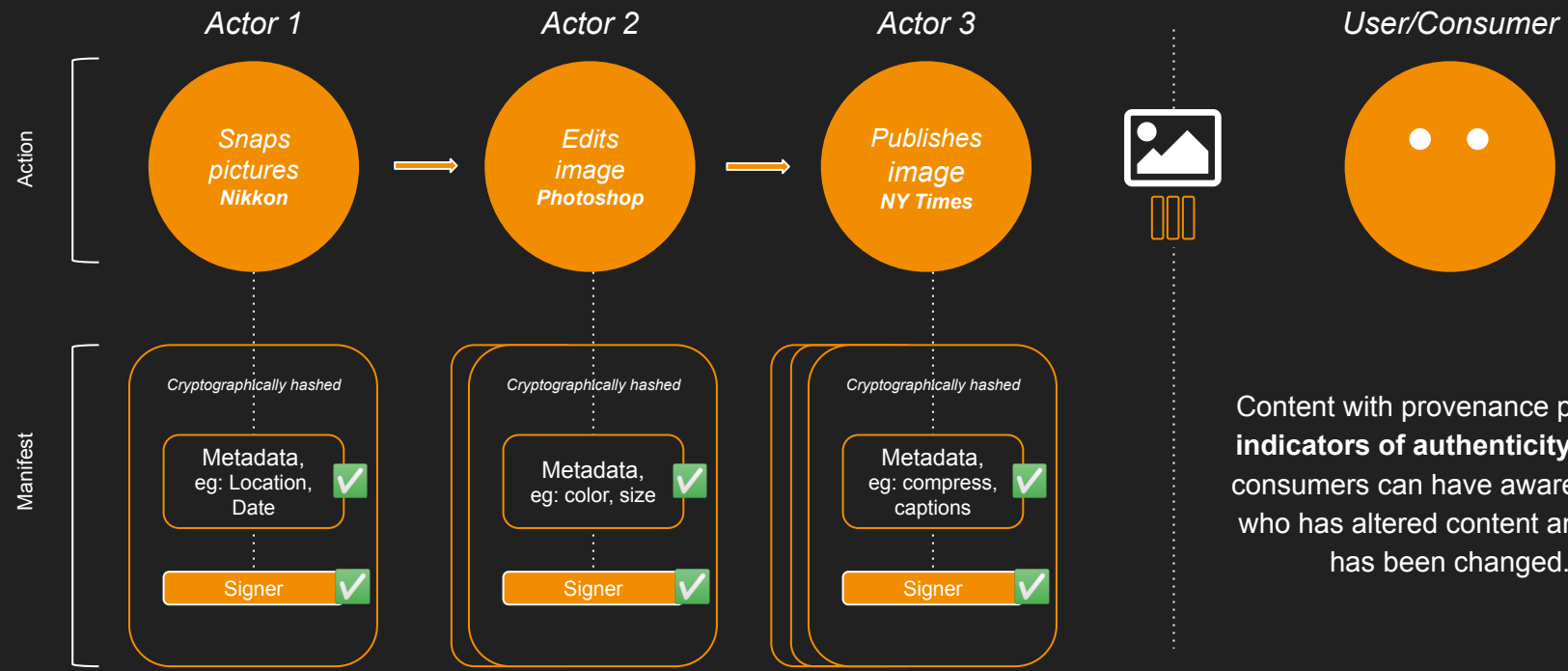
As we move towards systemic use…

How do we prevent, avert and mitigate harm?

How do we enhance freedom of expression and trust?

The Coalition for Content Provenance and Authenticity (C2PA) is an initiative that addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of digital assets.

The specifications intend to offer a secure way to establish the provenance of digital assets across platforms...



Content with provenance provides **indicators of authenticity** so that consumers can have awareness of who has altered content and what has been changed.

WITNESS
SEE IT FILM IT CHANGE IT

2 objectives within the C2PA

Prevent, avert, mitigate harms
(Bolstering human rights framework)

Promote critical usages

Steering Committee

Harms reporting / continuous assessment

Resourcing / Supporting a diverse C2PA ecosystem

Guiding Principles

Technical Working Group

Threats and Harms Taskforce

Harms Modelling

Strategies for averting, mitigating harms

The Guiding Principles of the C2PA

Guiding Principles

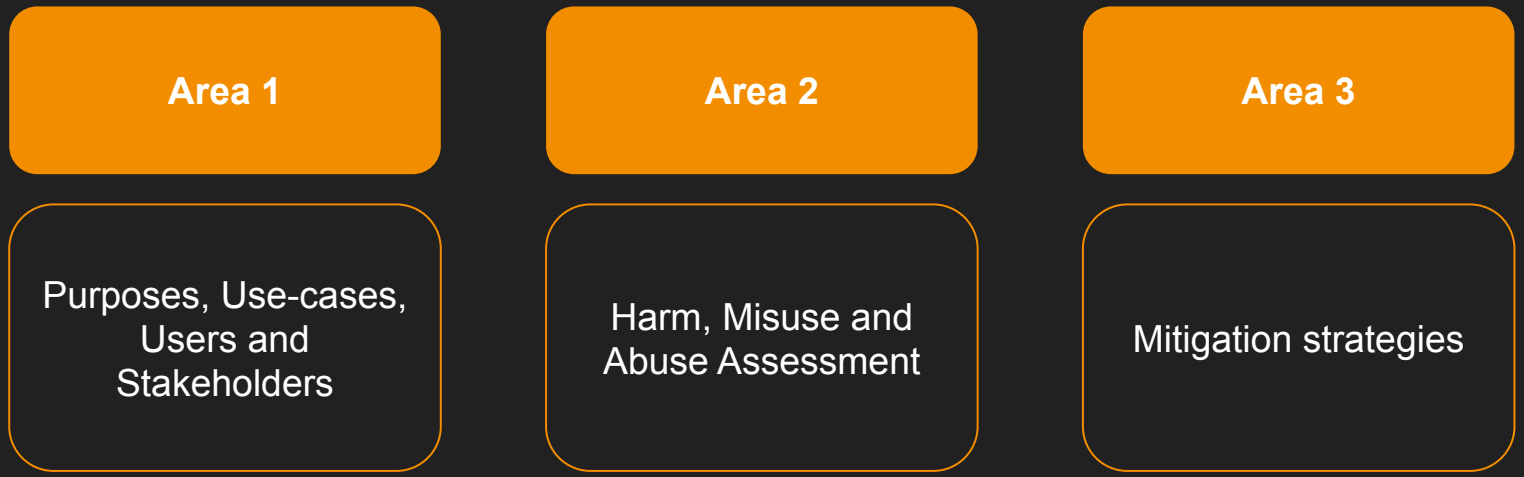
- Privacy
- Global Audience / Accessibility
- Simplicity and Cost Burden
- Misuse

- C2PA specifications MUST respect the common privacy concerns of each of the target users named earlier.
  - C2PA specifications MUST allow content creators, editors, and publishers to remove sensitive information before sharing with others. Subsequent participants must be made aware of such removal.
  - C2PA specifications MUST NOT require identity of the person or organization making any assertion or claim about an asset to be documented. The specifications MAY allow that information to be represented, provided that representation is optional.
- C2PA specifications MUST take into consideration the needs of interested users throughout the world.
  - C2PA-aware tools SHOULD be accessible to users with limited or high-cost access to Internet services.
- C2PA specifications MUST be reviewed with a critical eye toward potential abuse and misuse of the framework.

…

Harms Modelling

Harms Modelling



Microsoft's Harms Modelling Framework

The Harms, Misuse, and Abuse Assessment is an ongoing process that accompanies the design, development, implementation and use stages of the C2PA standard, and the process includes a multi-disciplinary and diverse range of stakeholders.

Assessment Methodology

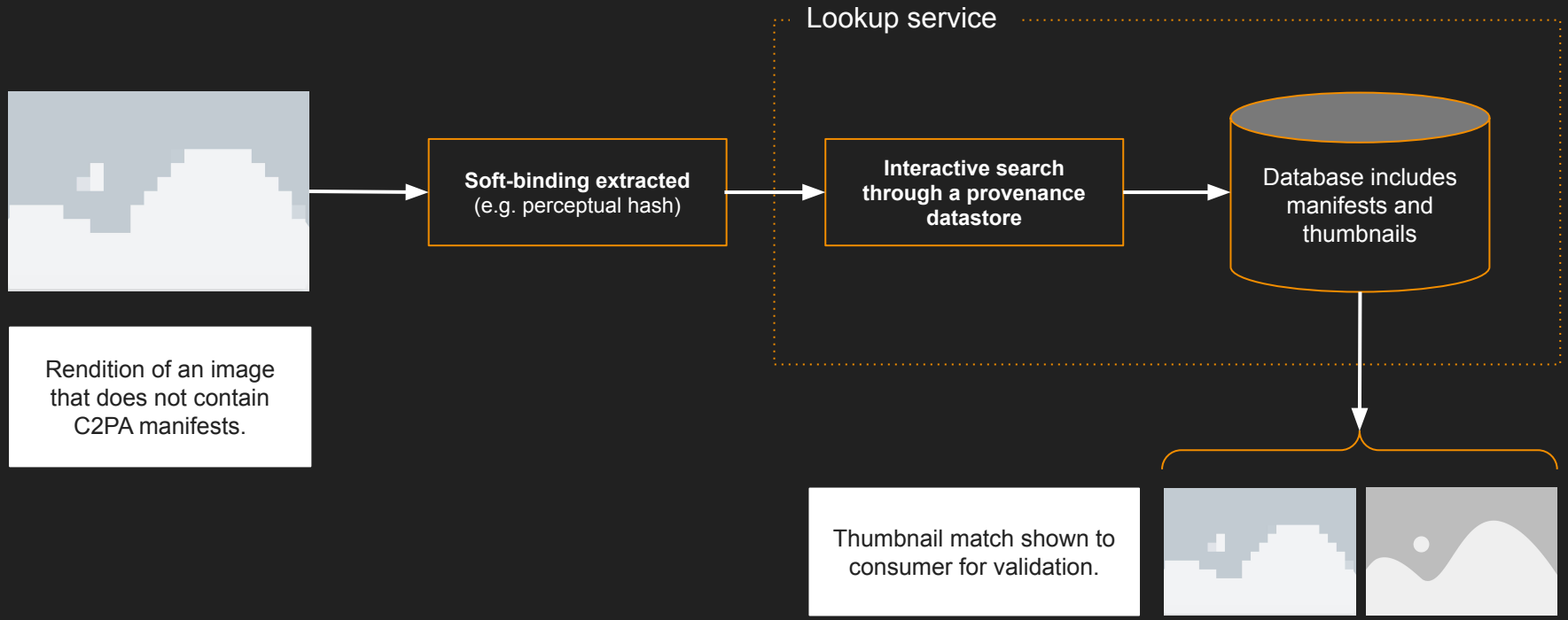
Internal consultations & discussions	**TWG** **Threats & Harms TF**
External consultations	Focus on global stakeholders, with different technical, lived, practical or professional experiences + most likely affected and marginalized from these processes

Overview of results

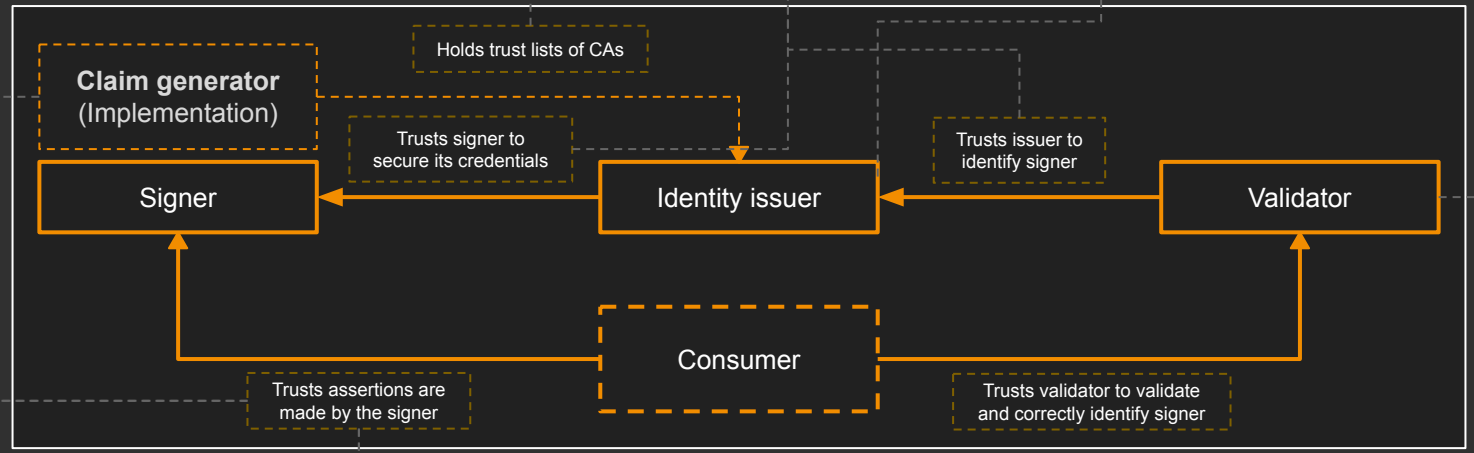
Category	Type of Harm
Denial of consequential services	Opportunity loss *(5)* Economics loss *(4)*
Infringement on human rights	Dignity loss *(1)* Liberty loss, discrimination and lack of due diligence *(6)* Privacy loss *(5)* Constraints on freedom of expression *(2)* Freedom of associations, assembly and movement *(2)* Environmental impact *(1)*
Erosion of social and democratic structures	Manipulation *(6)* Over-reliance on technical systems *(1)* Social detriment *(2)*
Risk of injury	Emotional or psychological distress or physical harm *(1)*

Reduction in options for anonymity and pseudonymity	Attacks on journalistic freedom and independence	Requiring participation in the use of technology or surveillance to take part in society
Privacy loss Human rights activist inadvertently includes location in media assertion and is subsequently targeted (c.f. existing precedents of inadvertent release of metadata, most famously John McAfee or recurring cases in human rights)	*Opportunity loss* An abuse of the C2PA system to enforce journalistic identity in laws in a jurisdiction or demand additional information on media posted on social media leads to a reduction of media diversity and suppression of speech.	*Freedom of association, assembly and movement.* For example, algorithmic ranking: content creators forced to game algorithms with particular keywords, metadata to achieve visibility/to be ranked higher in a feed.

Potential harm deriving from soft-binding and the use of manifest datastores



Trust model threats and harms



Tampering: either an update manifest, legacy media or security breach

There is no guidance for maintaining trust lists

Cost of x.509 certificates raises accessibility concerns

Technical difficulty to obtain a certificate raises accessibility concerns

Credentials compromised, e.g. stolen or revoked

Spoofing: Credential is misleading, e.g. BBC_LA

Unintended use: signed manifest is taken as trust, not trust signal.

Intended use: If you are unknown entity (e.g. La Sandía), signer is irrelevant because there is no trust in signer (no value for civic media etc.)

Credential is irrelevant to consumer

Unclear UX misrepresents trust signals

PHASE III - Existing and potential mitigations (*Version 1.0 - December 2021)*

Category	Type of Harm	Potential Harm / Misuse / Abuse	Contextual Example / Evidence	Existing and Potential Mitigations
		Language discrimination Limited language versioning on C2PA-enabled tools, despite their focus on low-cost and global accessibility, leads to more limited access for marginal markets.	C2PA-enabled tools are likely to leave out languages with marginal markets. A parallel example is that of the continued use in Myanmar of Zawgyi as the dominant typeface used to encode Burmese language characters rather than Unicode, the international text encoding standard, resulting in technical challenges for many companies that provide mobile apps and services.	**Specifications** Manifest localization specifications to be added beyond 1.0. No other limitations have been identified on language versioning of C2PA implementations. **Non-technical and multilateral harms response actions** The harms, misuse and abuse assessment should guide potential multilateral cooperation for the promotion of a diverse C2PA ecosystem, including language-inclusive implementations.
		Digital divide/technological discrimination (1) Individuals and communities using older devices or operating systems as creators/consumers or using access to the internet via Free Basics or equivalent "affordable access" approaches that limit the websites and services an actor can access.	For example, existing experiences with gated/limited access to particular websites and tools via Free Basics program for "affordable access" from mobile operators in emerging markets. See also example above on **Educational discrimination** and limited language versioning.	**Specifications** Specifications do not preclude C2PA implementations in older devices and operating systems. Specifications are open, global and opt in. The specifications use open standards for which there are existing libraries in various programming languages across a range of devices and operating systems/environments. To facilitate access for individuals or communities who do not, or cannot, have access to x.509 certificates, the specifications allow for self-signing certificates. **Accompanying documentation and guidance** Minimum viable implementations guidance to be developed as a fallback for older devices and operating systems. Guidance for implementers includes recommendations on the use of a private credential store (also known as the "address book"). **Non-technical and multilateral harms response actions** The harms, misuse and abuse assessment should guide potential multilateral cooperation for the promotion of a diverse C2PA ecosystem and encourage the development of simple products to meet claim generation and validation requirements in diverse environments. An ongoing harm asessment should inform the continuous development of the specifications to address issues that limit C2PA implementations in older devices and operating systems.
		Digital divide/technological discrimination (2) Individuals and communities without abiilty to access or use tools for compliance with system usage are excluded.	Financial costs involved in signing up to use different C2PA-enabled tools and software may exclude marginalized individuals and communities who cannot afford the cost. For example, exclusion of content creators without compliant x.509 certificates. Lack of literacy and access to education about the tool may also limit usage among marginalized populations.	
	Opportunity loss	**Journalistic Freedom and Independence** An abuse of the C2PA system to enforce journalistic identity in laws in a jurisdiction or demand additional information on media posted on social media leads to a reduction of media diversity and suppression of speech. Misuse of manifest repositories to track content or enforce restrictive laws on freedom of expression and do so with lack of effective remedy and/or exploitation of manifest repositories to track content, and curtail freedom of expression (e.g. political speech). See overlap with **Journalistic Plurality and Diversity**	An escalation of laws addressing 'fake news', misinformation/disinformation and social media globally includes laws that enforce registered identity as a journalist on social media or provide governmental right-to-reply, which are being used to suppress dissent and reduce journalistic freedom.	**Specifications** Specifications are open, global and opt in. If they are used, the C2PA provides features that can be used to protect confidentiality of personal information while still establishing the provenance of an asset, including anonymous and pseudonymous signing, redaction as an authorized action, use of update manifests with redacted information, and the use of W3C credentials. No sensitive information is required in C2PA workflows. **Accompanying documentation and guidance** User experience guidance provides recommendations to prevent inadvertent disclosure of information. Guidance for implementers highlights trusts and privacy considerations, including on the use of manifest repositories: We recommend that claim generators that add soft binding assertions to an asset's manifest do so as an opt-in addition and not make it mandatory. Guidance also recommends that content creators be informed of the trade offs involved in using manifest repositories that allow for asset link-up with soft bindings; that is, on the one hand, identifying manifests that have become 'decoupled' from their associated assets, while on the other hand, privacy risks that may result from a soft binding link-up to an earlier manifest with, for example, redacted information. **Non-technical and multilateral harms response actions** The harms, misuse and abuse assessment should inform the C2PA to proactively engage and lobby

Outputs of Harms Modelling