Content authenticity and provenance

Resource guide

Ana Garza Ochoa,
Digital Media Lead

Last updated: 8 May 2025

Introduction

The rise of information disorder in the digital space drives societal polarisation, eroding trust in media, institutions, and democratic processes. Disinformation campaigns, often driven by malicious actors, are deliberately crafted to exploit social fault lines such as ethnicity, religion, gender and political ideologies. Harmful narratives amplify existing divisions leaving little room for constructive discourse.

AI-powered algorithms play a critical role in amplifying these divisive narratives by prioritizing content designed to maximize engagement, often at the expense of accuracy. AI-generated content, such as deepfakes, exponentially increased the scale and sophistication of information disorder, making it harder for audiences to discern credible information. The lack of transparency and accountability in algorithmic decision-making exacerbates this problem, creating echo chambers and reinforcing biases that deepen ideological divides.

At RNW Media we will leverage our experience in upholding information integrity by combating information disorder. This includes generating evidence-based content, defining a framework for authenticity within independent media outlets, and promoting provenance standards for advocacy efforts. These measures do not only help in mitigating false information but also support in rebuilding public trust in the media.

Content authenticity is a key aspect for us and independent media organisations to contribute to fostering a safe, inclusive, and reliable media landscape.

Description

This resource guide is designed to provide a comprehensive framework for defining digital content authenticity goals and incorporating key provenance aspects to an editorial workflow. It aims to guide organisations in defining their needs and identifying different tools that fit their expertise and resources.

This framework is based on the **Adobe Content Authenticity Initiative (CAI)** (of which RNW Media is a member) joining the movement for providing content transparency in a way that it allows users to have a better evaluation of online content. The initiative offers open-source tools through the **Coalition for Content Provenance and Authenticity (C2PA),** with a set of "technical standards for embedding provenance data into digital content.

To guide users in this whitepaper is important to clarify the main concepts of content authenticity and provenance:

Content authenticity refers to the ability to verify the origin and provenance of digital content – such as images, videos and text – to ensure it has not been altered, misrepresented or manipulated in a deceptive way. It details transparency in who created the content, when and where it was created, if its AI generated or human made and whether it remains unaltered from its original state.

Content provenance is the record of a digital asset's creation, ownership, modifications and distribution over time. It functions as a digital fingerprint, providing a traceable history of content creation and who handled it (creators, editors and publishers). Provenance data is typically embedded through metadata, cryptographic signatures or blockchain-based tracking, helping users determine whether content is authentic and trustworthy.

The importance of content transparency is helping everyone understand who created the content, what has been done with the content, when/where was it created to give context, how has it been created (i.e. camera or AI software) and why is it being shared and its narrative.

Benefits

Combats misinformation & restores trust in media

The use of artificial intelligence (AI) for content creation has proliferated fake media, from low quality content to sophisticated deepfakes. The increasing degree of unrecognisable synthetic media is making it harder for audiences to discern credible information. Content authenticity and provenance play a crucial role in combating misinformation by providing transparent records of a digital asset's origin, edits and distribution. For journalism and public interest media, this is particularly vital, as the erosion of trust can lead to disengagement, skepticism and the spread of false narratives. By ensuring that users can identify manipulated content, this strengthens the credibility of trustworthy sources while mitigating the harmful effects of disinformation.

Protects journalists & media creators from content manipulation

Journalists and media creators are increasingly vulnerable to content manipulation, which can misrepresent their work, distort facts or even put them at risk. Content authenticity safeguards their credibility by ensuring proper credit attribution, preventing unauthorised alterations and protecting against the misuse of their work. Additionally, provenance verification helps media organisations comply with digital rights and privacy regulations reducing legal risks while upholding ethical journalism practices. By securing their content, journalists can continue their vital work without fear of distortion or exploitation.

Empowers audiences to verify content authenticity

Empowering audiences with tools to verify content authenticity is essential in strengthening digital media literacy. By making provenance information accessible, users can independently assess the legitimacy of images, videos and articles before sharing or acting on them. This not only reduces the spread of misinformation but also encourages responsible media consumption. When audiences can easily verify the authenticity of content, they become more discerning participants in the digital ecosystem, contributing to a more informed and resilient society.

Target Group

Public interest and independent digital media organisations, individual journalists or content creators who are interested in integrating content authenticity aspects into their editorial workflow.

Our Offerings

1. **Masterclass:** The users learn key components of an end-to-end authenticity infrastructure for editorial workflows. Explore available tools, good practices, while reflecting on aspects of content integrity.
2. **Coaching:** After completing the masterclass, user follows-up with coaching sessions to discuss current needs and incorporate content provenance framework.
3. **Research:** To understand the users' perceptions on AI generated versus human generated content. This leads to raising awareness on how users can assess the transparency of content.

SDGs Relevance

SDG 16: Peace, Justice and Strong Institutions

Target 16.10 – Ensure public access to information and protect fundamental freedoms.

Content authenticity helps combat misinformation, disinformation, ensuring that people have access to truthful, verifiable information for informed decision-making.

SDG 9: Industry, Innovation and Infrastructure

Target 9.C – Increase access to information and communications technology.

Advances in AI-driven authentication, blockchain for provenance and digital watermarking support a more transparent and secure digital ecosystem for media and content sharing.

SDG 17: Partnerships for the Goals

Target 17.6 & 17.8 – Strengthen science, technology and innovation cooperation.

Initiatives like the Content Authenticity Initiative (CAI) and C2PA foster global partnerships between tech companies, media organisations and researchers to develop open standards for content verification.

Editorial Content Authenticity Framework

Starting with content authenticity in an editorial workflow requires a structured approach to ensure the integrity of content from creation to publication, with an emphasis on aligning this framework with policies and ethical guidelines.

Stages & Key Resources of an Editorial Content Authenticity Framework

1. Define Content Authenticity Objectives
2. Create & Implement Provenance Tools
3. Set-up Verification Workflows & Tools
4. Publish Content with Authenticity Credentials

1. Define Content Authenticity Objectives

Define the scope of implementing content authenticity within your organisation. Do you want to focus on ensuring transparency in AI or human generated content, combating misinformation, publishing content with visible markers to inform your audiences? Decide the type of content (text, images, videos) you would like to cover.

Revise your editorial guidelines and align with any legal or industry standards. Consider ethical standards within your organisation and adapt or incorporate new policies that reflect your objectives.

2. Create & Implement Provenance Tools

2.1. Integrate Provenance Mechanisms

Choose tools that allow authenticity data to be injected at content creation. You can secure capture by adopting solutions to authenticate images and videos at the point of capture recording all relevant information. Integrate into your content management systems plugins or extensions for provenance. Public ledgers, like blockchain, permanently preserve authenticity certificates as immutable records and ensure long term durability.

Ensure transparency in editing by integrating software (like Adobe tools) that is C2PA-compliant. When editing images, video or text make sure all significant edits (i.e., cropping images) are traceable and logged in the provenance chain with clear attribution.

Different content formats require tailored transparency methods. For news articles and investigative journalism, version control and edit logs ensure traceability, while C2PA metadata maintains a transparent chain of custody for edits. Editorial workflow tools document modifications, reinforcing accountability.

2.2 Track AI Generated and Edited Content

- **AI Content Log**: Implement a process where AI-generated content is logged with details on how and when AI was used for content creation and editing.
- **AI Tracking Tools**: Leverage tools that can assist with recording AI involvement during editing process. For example, within CMS platforms (WordPress, Drupal) implement AI content tracking to automatically tag and label AI-assisted content.
- **Human Oversight**: Implement workflows where AI-generated content is reviewed by human editors before publication to ensure alignment with ethical guidelines. Subject matter experts should assess content for bias, accuracy and compliance with regulations. Combining human oversight with technology ensures AI-assisted media maintains credibility and prevents misinformation.

3. Set Up Verification Workflows & Tools

To ensure your content is trustworthy, build a simple and consistent process for verifying media before you publish or share it. Set up a validation interface to verify provenance metadata at the point of use. This interface should automatically assess whether digital signatures are valid, whether the content's provenance chain is complete and untampered and whether any component of the content manifest has been revoked, expired or altered. Embedded validators — integrated into apps, websites and devices — perform these checks in real time, ensuring the authenticity and integrity of content when it is displayed, shared or repurposed.

Before using content from social media or other sources, run it through trusted tools (i.e., Google Fact Check Explorer, WeVerify or Reality Defender). These tools help detect manipulated images or AI-generated media. Verification services analyse media files to extract authenticity certificates and match data to confirm validity.

Before publishing, during fact-checking or when repurposing content, verification should be part of your workflow.

4. Publish Content with Authenticity Credentials

4.1 Attach and Display Credentials

To follow the C2PA standards, embed manifests directly into media files (where supported) or link via sidecar files/URIs. For media type that supports in-file embedding (JPEG, PNG) this is a more direct process. For instance, a photo with a manifest is a multi-layer container that allows information layers to travel with the photo when it is shared online and it can also be easily verified with verification services.

To maintain the authenticity of content after publication, ensure integrating platforms that allow content credentials to be displayed. For example, integrating visual labels, trust badges, audio notifications to alert audience when content has been altered or created by AI. These human-facing transparency methods work well for news outlets, social media posts and entertainment media, where immediate recognition of AI involvement is necessary for audience trust. While machine-readable transparency methods, including cryptographic watermarking, metadata tagging and statistical pattern analysis, are effective for automated misinformation detection systems. These methods enable search engines, regulators and content verification tools to trace the origins of synthetic content at scale.

4.2 Monitor Post-Publication Integrity

Set-up your team with adequate tools to continuously monitor published content for unauthorised alterations, misuse or disinformation campaigns. Some tools will automatically detect and notify if the content has been tampered with. Reference the original manifests as the single source of truth to distinguish between authentic and altered versions.

Educating audiences about content authenticity and integrity is essential for digital media literacy. Provide a way for audiences to validate content authenticity themselves (e.g., using browser extensions or content viewer tools) or publish regular transparency reports on AI use in editorial workflows to foster accountability.

At every stage consider cross-industry collaboration. By working with media organisations, tech companies and policymakers more efforts and focus can be dedicated to adopting authenticity standards.

Here are some key tools and platforms supporting digital content provenance:

Coalition for Content Provenance and Authenticity (C2PA)

- **What it does:** Develops open standards for tracking the origins of digital content.
- **Key Feature:** Provides a standardised framework for embedding provenance metadata in images, videos and documents.
- **Website:** c2pa.org

Adobe Content Authenticity Initiative (CAI)

- **What it does:** Embeds provenance metadata into digital media, allowing creators to prove authenticity.
- **Key Feature:** Works within Photoshop and other Adobe products to track edits and modifications.
- **Website:** contentauthenticity.org

Truepic

- **What it does:** Uses cryptographic methods to verify the authenticity of photos and videos.
- **Key Feature:** Secure camera technology that captures tamper-proof images with metadata verification.
- **Website:** truepic.com

Starling Lab (USC & Stanford Initiative)

- **What it does:** Uses blockchain and cryptographic methods to create tamper-proof digital records.
- **Key Feature:** "Capture, Store, Verify" framework ensures digital content integrity over time.
- **Website:** starlinglab.org

Project Origin (BBC, CBC, The New York Times, Microsoft)

- **What it does:** Focuses on tackling disinformation by verifying news content authenticity.
- **Key Feature:** Embeds cryptographic signatures in journalism content for secure verification.
- **Website:** originproject.info

Proactive Detection of Voice Cloning with Localised Watermarking

- **What it does**: Uses imperceptible audio watermarking to detect AI-generated speech and voice cloning.
- **Key Feature**: Fast and robust detection system that remains effective even after common audio manipulations.
- **Website**: arXiv Paper

SynthID (Google DeepMind)

- **What it does**: Embeds invisible yet detectable watermarks into AI-generated images for authenticity verification.
- **Key Feature**: Ensures watermark resilience even after modifications like resizing or compression.
- **Website**: Google DeepMind SynthID

IMATAG - Digital Watermarking

- **What it does**: Provides invisible watermarking to protect digital images and videos from unauthorised use.
- **Key Feature**: Enables tracking and authentication of content across the web.
- **Website**: IMATAG

Numbers Protocol

- **What it does**: Uses blockchain technology to create verifiable records of digital content provenance.
- **Key Feature**: Implements decentralised tracking to enhance trust and ownership in digital media.
- **Website**: https://www.numbersprotocol.io/

Verify (by The Content Authenticity Initiative)

- **What it does:** A web tool that allows users to check the provenance of digital content.
- **Key Feature:** Displays content metadata (who created it, when and if it was modified).
- **Website:** Content Credentials

Implementation

Ideally, this solution should be adopted by organisations or newsrooms actively engaged in content creation and committed to incorporating authenticity measures into their processes. Successful implementation requires dedicated resources for ongoing maintenance, updates and the adoption of new technologies. Additionally, integrating these solutions within broader industry initiatives and raising awareness among your audience will enhance their effectiveness.

1. **Decide your primary objective** – Depending on your role and priorities clearly integrate the necessary methods and policies into your editorial guidelines.
2. **Choose tools that fit your workflow** – Many of the above-mentioned tools can be integrated into existing CMS platforms. Test different open-source tools and their effectiveness throughout the lifecycle.
3. **Educate your team & audience** – Train staff and inform your readers/viewers about content provenance.

Challenges & Considerations

Technical Limitations & Effectiveness

One of the key challenges in establishing provenance and verification in AI-generated media is the inadequacy of traditional methods, which are often ineffective against evolving technologies. A report from *MIT Technology Review* highlighted the shortcomings of watermarking AI-generated content, noting that such measures can be easily removed or altered by malicious actors. While initial solutions are tested and implemented, these still raise questions around the sustainability and effectiveness of these methods to reduce harm.

Cost & Accessibility: The Need for Open-Source and Scalable Solutions

The push for open-source solutions is essential in making content verification accessible to a broader range of users, particularly independent media outlets with limited resources. Without scalable and cost-effective tools, the process of ensuring content authenticity becomes out of reach for many stakeholders.

Privacy vs. Transparency: Balancing Security with Ethical Considerations

For journalists or content creators working on sensitive topics or in restrictive contexts, adding personal information to images or other types of content could pose some risks, as stated in the research from Mozilla Foundation “while transparency ensures that users can verify the origin and integrity of content, it often requires the collection and disclosure of detailed metadata, which can raise privacy concerns. For instance, embedding extensive metadata to track content authenticity can inadvertently expose sensitive information about creators or subjects, potentially leading to privacy breaches” (*Vassei Mulavi, Udoh*, 2025). This is an aspect that requires more discussions and sharing learnings within the international initiatives.

Policy & Regulation Gaps

The rapid rise of digital content creation has outpaced the development of global regulations governing content authenticity. This gap highlights the urgent need for unified global standards and better enforcement mechanisms. Moreover, the increasing degree of unrecognisable synthetic media and the challenges that come with it calls for cross-industry cooperation to establish cohesive policies that ensure the integrity of digital content and mitigate the risks posed by evolving AI technologies.

Glossary

Deepfakes
Media—often videos, images, or audio—created or manipulated using AI to make them appear real, despite being fabricated or altered. Deepfakes can be used for a variety of purposes, including misinformation, political manipulation, or entertainment.

Source: General industry definition.

Misinformation
False or inaccurate information spread without the intent to deceive. Misinformation can be unintentional, but it can also be disseminated deliberately by malicious actors.

Source: General industry definition.

Disinformation
Deliberately fabricated or manipulated information designed to mislead, deceive, or influence public opinion, often for political or financial gain. *Source: General industry definition.*

AI-Generated Content (AIGC)
Content created by artificial intelligence systems, including text, images, videos, and audio. AI-generated content is created autonomously by algorithms, without direct human input, and is used in various industries, including media, entertainment, and marketing.

Source: General industry definition.

Blockchain for Provenance
A decentralized digital ledger that tracks and records the history of a digital asset, ensuring it cannot be altered or tampered with. Blockchain technology is used in content provenance to verify and authenticate content by creating immutable records of its origin, edits, and ownership.

Source: General industry definition.

Cryptographic Watermarks
Digital signatures embedded in content that are nearly impossible to remove without damaging the content. These invisible watermarks are used to protect digital media from tampering or unauthorized use, ensuring the content's provenance and integrity.

Source: General industry definition.

Disclosure Methods
Techniques for revealing or making known the origins, intentions, and implications of digital content, particularly content generated or influenced by AI systems. These can be classified as visible/direct or invisible/indirect based on how they engage with audiences (human or machine).

Source: General industry definition.

Transparency Labels
Visual or audio indicators placed on digital content to inform the audience about its origin or the role of AI in its creation.

Source: General industry definition.

Metadata
Data that provides information about other data. In the context of digital content, metadata includes details such as the creation date, author, device used, location, and modification history. It plays a crucial role in verifying the provenance and authenticity of digital assets.

Source: General industry definition.

Synthetic Media
Media that is generated or altered using artificial intelligence technologies. This includes AI-generated

text, images, videos, and audio, which can be created entirely by machines or heavily modified from original content.

Source: General industry definition.

Content Tampering
The unauthorized alteration, manipulation, or modification of digital media, including text, images, audio, and video, with the intent to deceive, mislead, or distort information. This can include techniques like deepfakes, AI-generated misinformation, or subtle editorial changes.

Source: General industry definition.

Documentation / Annex

Content Authenticity Initiative. (n.d.). https://contentauthenticity.org/

Digital Authenticity: Provenance and Verification in AI-Generated Media. (n.d.). https://www.numbersprotocol.io/blog/digital-authenticity-provenance-and-verification-in-ai-generated-media

The Legal Landscape of Content Authenticity: Your Guide to Emerging regulations. (2025, February 12). https://www.imatag.com/blog/the-legal-landscape-of-content-authenticity-your-guide-to-emerging-regulations?utm_source=chatgpt.com

Vassei Molavi Ramak, Udoh Gabirel (2024, February). *In Transparency We Trust? Evaluating the Effectiveness of Watermarking and Labeling AI-Generated Content.* https://foundation.mozilla.org/en/research/library/in-transparency-we-trust/research-report/

Note: *The content of this whitepaper was researched and drafted with the support of AI tool: ChatGPT*

Registered office:
Koepelplein 1C
2031 WL Haarlem
The Netherlands

Email:
info@rnw.org

Website:
https://www.rnw.media/