Cybersecurity Information

TLP:CLEAR

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Communications Security Establishment
Canadian Centre for Cyber Security

Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

National Cyber Security Centre

Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era

Executive summary

With the rise of advanced tools that enable the rapid creation, alteration, and distribution of images, videos, and other digital content, there are many ways to manipulate what people see and believe. The ability to manipulate media is not new, but the accessibility, speed, and quality of these modifications today, powered by artificial intelligence (AI) and machine learning tools, have reached unprecedented levels and may not be caught by traditional verification methods. While many people are using generative AI to create (or modify) and distribute useful creative content faster, it is important to acknowledge the risks and potential harms of the technology when used for malign purposes. For example, malicious actors can use manipulated or fully synthetic media in cyber threat, criminal, or other malign activity against organizations and individuals to impersonate and misinform. In addition, there are broader societal risks around loss of trust online that can impact not just individuals and businesses, but whole communities. Due to this widespread ability to convincingly create or modify media, verifiable media is becoming critical for ensuring transparency by providing context about the media's provenance and integrity with an effective, secure, and robust technical standard.

Content provenance solutions aim to establish the lineage of media, including its source and editing history over time. Of these solutions, Content Credentials™ are a provenance solution that uses cryptographically signed metadata describing the provenance of media. This metadata can be attached to the media content during export from software or even at creation on hardware. [1] In addition, Durable Content Credentials add two additional layers of preservation for the retrieval of Content Credentials by adding a digital watermark to the media and implementing a robust media fingerprint matching system. [2]

This cybersecurity information sheet, authored by the National Security Agency (NSA), Australian Signals Directorate's Australian Cyber Security Centre (ASD's ACSC), Canadian Centre for Cyber Security (CCCS), and United Kingdom National Cyber Security Centre (NCSC-UK), discusses how Content Credentials (especially Durable ones) can provide transparency for the provenance of media, raises awareness of the state of this solution, introduces recommended practices to preserve provenance information, and emphasizes the importance of widespread adoption across the information ecosystem.

Introduction

The widespread availability of AI and machine learning tools, including generative models and deepfake technologies, makes it possible for anyone to convincingly create and/or modify media with minimal effort, low cost, and increased realism. This rapid evolution poses a significant challenge for traditional verification methods, which may struggle to keep up with the growing sophistication and scale of these technologies. As a result, the accuracy and effectiveness of verification methods are increasingly under strain, leaving consumers more vulnerable to misinformation and influence operations. The abuse of AI-generated media[1] also represents a significant cyber threat to organizations, including through impersonation of corporate officers and the use of fraudulent communications to enable access to an organization's networks, communications, and sensitive information. Some of these threats were described in the previous joint cybersecurity information sheet (CSI): Contextualizing Deepfake Threats to Organizations. [3] In addition to these specific threats, the inherent general trust in multimedia content is quickly eroding. As a result, the need to bolster information integrity has never been more urgent. [4] Although other technologies, such as watermarking, can be used for media provenance, Content Credentials (especially Durable Content Credentials) are the focus of this report[2].

Success in increasing trust through transparency will rely on the secure and widespread adoption of standard practices across the information ecosystem, including the Defense Industrial Base (DIB) and National Security Systems (NSS). Content Credentials are an evolving technical standard that can significantly increase the transparency of media provenance. As is typical with new and maturing standards, it often takes time to implement them securely across various modalities and address new concerns and

[1] In this report, AI-generated media will refer to media that has been created and/or edited with generative AI technology.

[2] For a more comprehensive overview of provenance solutions, refer to NIST AI 100-4.

edge cases as they arise. However, that should not deter organizations from preparing and getting started now, especially for common use cases.

A common misconception is that Content Credentials are only relevant to photojournalism. In fact, Content Credentials can be used broadly across the information ecosystem for the provenance of a variety of media, including images, video, audio, and text.[3] Established and emerging provenance technologies, such as Content Credentials, provide an avenue to accomplish the goal of preventing further erosion of trust in media, and information more generally, through secure and widespread adoption.

Although the field of provenance standards is nascent and rapidly evolving, the standards developed by groups, such as the Coalition for Content Provenance and Authenticity (C2PA) [5] and implemented by the Content Authenticity Initiative® (CAI), [6] are gaining traction. These standards are intended to guide how software and hardware products related to the creation, editing, and distribution of media content record, verify, and manage provenance information. The term Content Credentials throughout this report refers to the implementation of the C2PA technical specification. This specification is currently being fast-tracked to become ISO® standard 22144, meaning that it will likely be officially recognized soon as a global standard for content provenance and authentication. [7], [8]

The previous joint CSI on Contextualizing Deepfake Threats to Organizations offers recommendations for security professionals focused on protecting organizations from the evolving threats of AI-generated media and deepfakes through advice on defensive and mitigation strategies. This CSI expands on one mitigation strategy to:

- provide an overview of media provenance and Content Credentials,
- discuss the importance of early and widespread adoption,
- describe potential use cases on the adoption of Content Credentials beyond photojournalism,
- highlight that Content Credentials are an evolving standard that will continue to improve through community engagement,
- introduce recommended practices to ensure Content Credentials and provenance are preserved, and

[3] Currently, Content Credentials are most mature for images. However, other modalities are following rapidly, with text being the least tested.

- emphasize that awareness remains critical as generative AI and deepfake technologies mature.

The guidance in this CSI can be used in support of the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF). In particular, as described in Action GV-6.1-008 recommended by the AI RMF Generative Artificial Intelligence Profile, policies to maintain provenance information using methods like Content Credentials can help to manage AI risks associated with third parties' behaviors.

This CSI only offers information and guidance based on the current landscape of techniques and threats. That landscape is changing rapidly, so updated information and guidance will be needed as new developments arise. The intent in publishing this information now is to help raise awareness of the importance of content provenance, the state of one current solution, and steps that can be taken to get started. The technical details for secure implementation of Content Credentials are out of scope for this introductory guidance.

Content Credentials and Durable Content Credentials

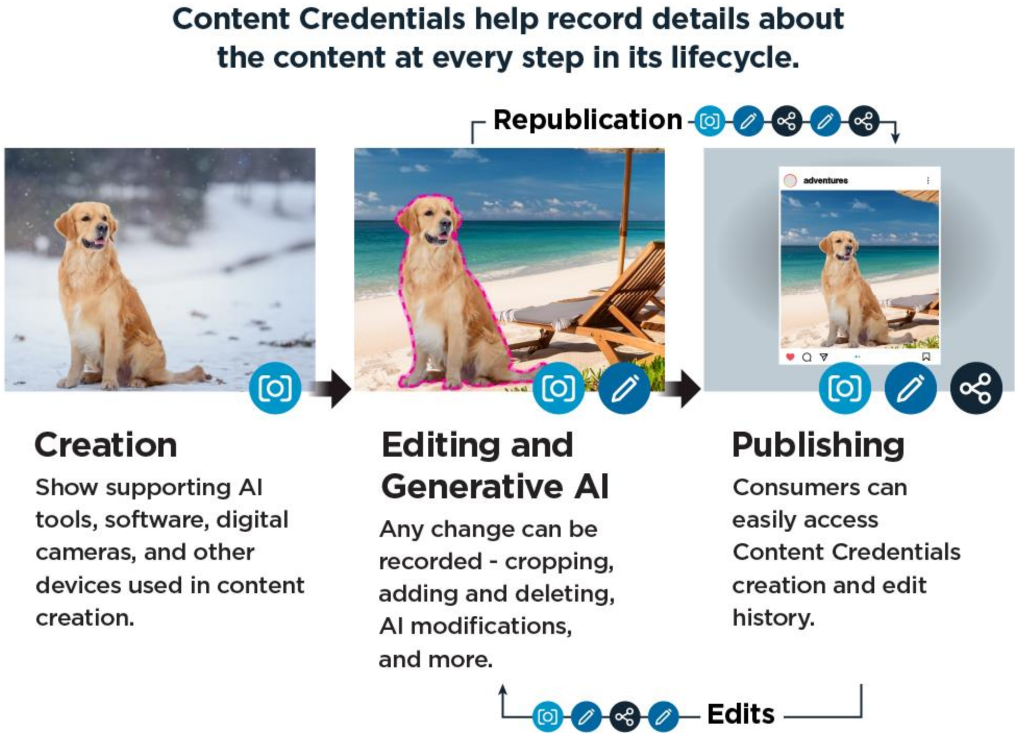
Content Credentials are defined as metadata that are secured cryptographically and allow creators the ability to add information about themselves or their creative process, or both, directly to media content. Content Credentials can be added to media at export from software or at creation on hardware. [1] Content Credentials securely bind essential metadata to a media file that can track its origin(s),[4] any edits made, and/or what was used to create or modify the content. To facilitate global accessibility, Content Credentials have built-in functionality to work offline, such as the ability to copy certificates to an enclave. The C2PA technical specification describes the format of Content Credentials and the assertions that they contain about a media item's provenance.

The overall goal of including this information in media content is to equip online audiences with information about the source and/or editing history of digital content, so that they can make informed decisions about the content, similar to the purpose of a nutrition label on food. [10] These details can also facilitate trust through transparency among consumers of the content. This metadata alone does not allow a consumer to

[4] Origins in the case of media created from multiple sources.

determine whether a piece of content is “true,” but rather provides contextual information that assists in determining the **authenticity**[5] of the content. In other words, the metadata provides claims about the content by some entity who verifiably signs those claims. For this technology to be effective, consumers must have confidence in the reliability and accuracy of the information itself.

To make the credentials more robust against stripping or modification of metadata, adding a digital watermark to the media and implementing a robust media fingerprint matching system allows two additional levels of preservation and retrieval of Content Credentials. Content Credentials with these additional protections are known as **Durable Content Credentials**.[6] [2] While Content Credentials focus on metadata, Durable Content Credentials specify measures to make the metadata durable in the manner described above and provide a more comprehensive provenance solution. [11] Durable Content Credentials have also been shown in recent experiments to be useful for recovery of original video content and Content Credentials in physical (non-digital) media, such as printed copies of credentialed images. [12]


Content Credentials help record details about the content at every step in its lifecycle.
Republication
adventures
Creation
Editing and Generative AI
Publishing
Edits

Show supporting AI tools, software, digital cameras, and other devices used in content creation.

Any change can be recorded - cropping, adding and deleting, AI modifications, and more.

Consumers can easily access Content Credentials creation and edit history.

Figure 1: Content Credentials lifecycle depicted. This can be from the point of capture (creation), to editing and generative AI use, to publishing (based on https://contentauthenticity.org/how-it-works). [13]

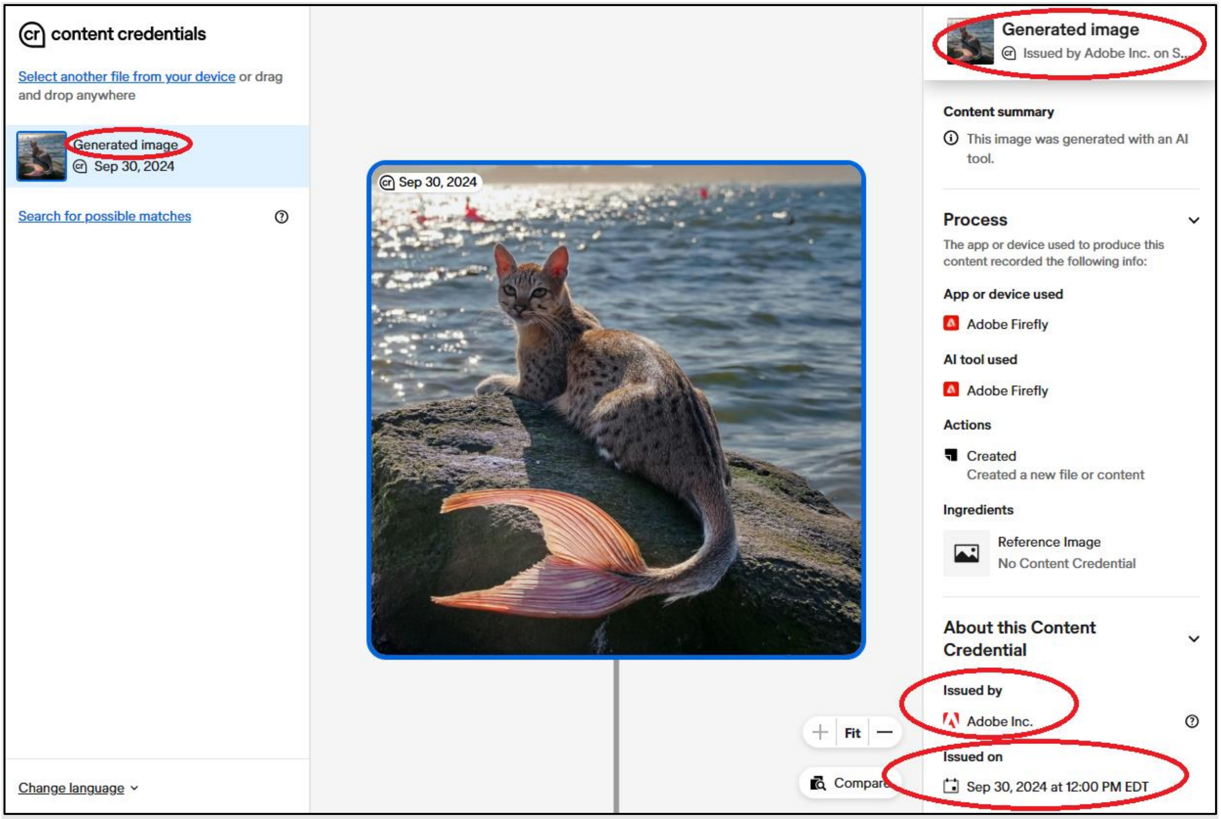
[5] The word “authenticity” is used here to describe media content—including images, videos, audio, and text—whose origin, integrity, and history are verifiable and have not been manipulated in a deceptive manner.

[6] Note that the term Content Credentials is sometimes confused with digital watermarking. Although digital watermarking is different, it can be used as a mechanism to secure Content Credentials.

Content Credentials can either be attached by hardware, like cameras, or included by software used to create or edit media, like generative AI technologies or traditional editing programs. Today, most provenance solutions pertain to image and video content, although capabilities in audio and text are emerging (with text being particularly challenging). [14] While provenance solutions for image, video, and audio content can rely on metadata at the file level, text often cannot and therefore requires more nuanced approaches.

Lack of provenance information should not automatically make media less trustworthy.

At this time, provenance technologies such as Content Credentials are generally opt-in. Creators or distributors of content decide whether to include provenance information and how much information to include. [15] Lack of provenance information should not automatically make media less trustworthy, as information may be withheld for legitimate reasons like privacy and/or proprietary concerns, or due to technological limitations. However, large-scale implementation of Content Credentials may result in the general public considering media without Content Credentials as suspect without a reasonable justification for their absence.


content credentials
Select another file from your device or drag and drop anywhere
Generated image
Sep 30, 2024
Search for possible matches
Change language
Sep 30, 2024
Fit
Compare

Generated image
Issued by Adobe Inc. on S...

Content summary
This image was generated with an AI tool.

Process
The app or device used to produce this content recorded the following info:

App or device used
Adobe Firefly

AI tool used
Adobe Firefly

Actions
Created
Created a new file or content

Ingredients
Reference Image
No Content Credential

About this Content Credential

Issued by
Adobe Inc.

Issued on
Sep 30, 2024 at 12:00 PM EDT

Figure 2: Content Credentials can include details about whether an image was created using generative AI (screenshot from https://contentcredentials.org/verify). [16]

Background on C2PA

One project developing enhanced provenance technology is the C2PA. [5] The project was named in both the EU's 2022 Strengthened Code of Practice on Disinformation and the Partnership on AI's framework for Responsible Practice for Synthetic Media as a possible way to increase transparency and authenticity in digital content. [17], [18] The C2PA is a coalition of technology companies and media organizations that aims to combat the spread of misleading information online by developing open technical standards for verifying the origin and/or history of digital content. The C2PA was formed in 2021 to unify the efforts of two similar initiatives:

- The Content Authenticity Initiative (CAI), led by Adobe®, and
- Project Origin, led by Microsoft® and the BBC®.

The C2PA has over 200 members, headed by a steering committee consisting of Adobe, Amazon®, BBC, Google®, Intel®, Meta®, Microsoft, OpenAI™, Publicis Groupe, Sony®, and Truepic®. [19] These organizations are at various stages of implementing Content Credentials, with additional implementations being reported regularly, such as from Truepic, Google, BBC, and Microsoft, to name a few. [20], [21], [22], [23], [24] The Appendix provides a table with a current snapshot of the landscape. The closely affiliated CAI, which focuses on the development of systems to provide provenance information for digital media, currently has over 4,000 members. [25]

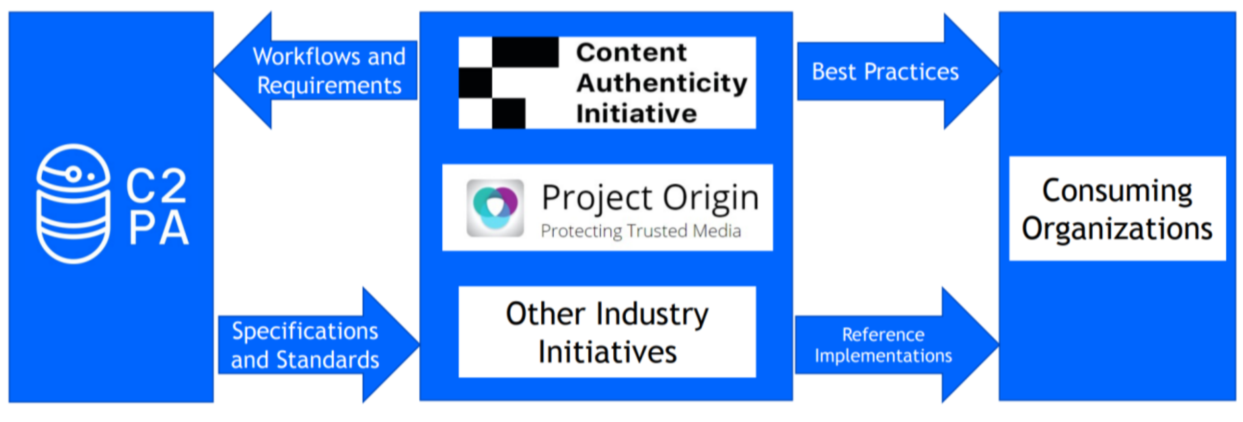

C2 PA
Workflows and Requirements
Specifications and Standards
Content Authenticity Initiative
Project Origin
Protecting Trusted Media
Other Industry Initiatives
Best Practices
Reference Implementations
Consuming Organizations

Figure 3: Diagram depicting relationship between C2PA, CAI, Project Origin, other industry initiatives, and consuming organizations (from https://c2pa.org/files/C2PA_Introduction_Deck.pdf) [26]

The C2PA has developed a freely available specification for providing digital content provenance through Content Credentials. The specification is both technical and normative in scope and is designed "to enable global, opt-in, adoption of digital provenance techniques through the creation of a rich ecosystem of digital provenance enabled applications for a wide range of individuals and organizations while meeting appropriate security requirements.” [27] Given its freely available nature, this standard could be used by anyone—including governments and their service providers—to implement digital provenance information in their products and processes.

Dangers of digital content misuse

As high-quality synthetic[7] media becomes cheaper and easier to produce, it has become increasingly difficult for content consumers to evaluate the authenticity of digital content. [28], [29] In 2023, Merriam-Webster® deemed “authentic” the word of the year, in part due to increased interest in AI and synthetic media. [30] In 2024, TIME® magazine listed Content Credentials in its “Best Innovations” list, demonstrating their understanding of the importance and need for content provenance solutions. [31]

Currently, substantial threats to organizations from the abuse of AI-generated content include impersonation of leaders and financial officers and the use of fraudulent communications to enable access to an organization’s networks, communications, and sensitive information. [3] In response, cybersecurity guidance recommends evaluating and verifying content and sources. [32]

When the information environment is full of questionable content, a general erosion of trust in media follows. The erosion of trust also can be purposely abused to discredit legitimate content in what is called the Liar’s Dividend[8], sowing even further distrust of content. [33], [34] All of this leads to the need to increase transparency in media, to bolster trust, which Content Credentials can help provide. Success relies on well-defined, accessible, and usable transparency mechanisms by which individuals and institutions can understand the information they consume. Prerequisites for this include the secure and widespread adoption of standardized transparency practices across the information ecosystem, including for the Defense Industrial Base (DIB) and National Security Systems (NSS).

However, Content Credentials by themselves will not solve the problem of transparency entirely. Instead, a multi-faceted approach that includes provenance, education, policy, and detection (Figure 4) is recommended. This report primarily focuses on provenance and awareness (education) but will now briefly address detection and policy.

[7] Produced or altered by generative AI.

[8] When continuing to spread false information makes it unclear what is true.

The detection of manipulated media and identification of generative AI content will likely continue to be necessary, given that not all individuals who produce or share media will disclose its origins. However, detection is a passive approach and will always be a cat and mouse game as technology evolves. This is why it is important for legislation (policy) to establish laws for multimedia transparency and use of generative AI. For example, the EU AI Act places obligations on providers and users of AI systems to enable the detection and tracing of AI-generated content. [35] Continuing to advance all aspects of this multi-faceted approach can help mitigate the dangers of digital content misuse.

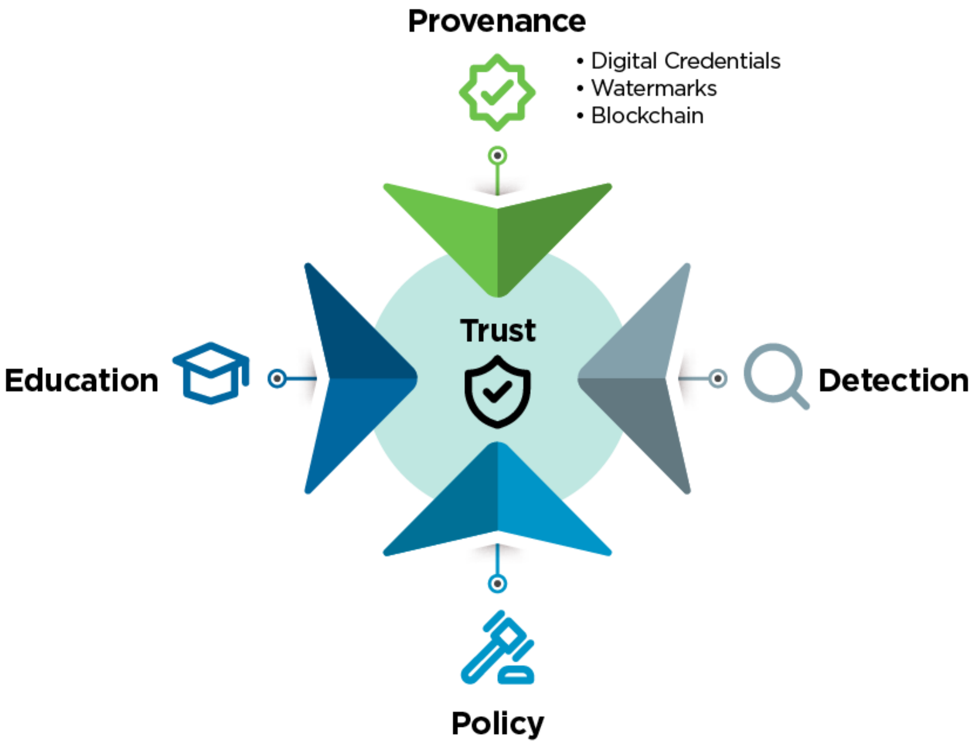


Figure 4: All of these areas play an important role in determining whether to trust media content.

Motivating factors for providing media provenance information with Content Credentials

As described in the introduction, provenance information can help to mitigate cybersecurity risks from synthetic content, as well as risks to information integrity. Additional motivators for providing media provenance information with Content Credentials are described below.

1. Protection of person identity and reputation

As described in the previous joint CSI on Contextualizing Deepfake Threats to Organizations, the authors recommended that organizations consider using active authentication techniques, especially for high-level officials' media. [3] However, this recommendation is increasingly relevant to everyone as generative AI technologies now have the ability to impersonate anyone using only a few (or in some cases only one) examples. These active authentication techniques provide additional layers of complexity an adversary must overcome to claim that a fake media asset is real. This includes watermarking and fingerprinting combined with secure metadata for increased durability. For example, if someone consistently uses Content Credentials to sign their media, it becomes more difficult for anyone to credibly claim that a fake image of them is legitimate, given the established practice.

2. Protection of content

Adding Content Credentials to digital content can enhance the creator's ability to detect unauthorized use of their content to train generative AI models. Several research papers discuss the application of Content Credentials for this purpose. [36] Additionally, Content Credentials now allow for the addition of a "do not train" tag to be added if the content creator does not want their public data to be used to train generative AI models. [37]

3. Navigating the model collapse problem

Model collapse is a documented degenerative process that involves generative AI data polluting the next generation's training set, affecting generations of the model up to collapse. [38], [39], [40], [41] To help combat this problem, organizations need to pay particular attention to the source of the data used to train models and encourage content creators to use a provenance solution, such as Durable Content Credentials, for identification of multimedia generated by AI. This also applies to models from external vendors; organizations should pay attention to how vendors acquire training data and ensure provenance. It is important to note that labelling of AI-generated content may be required in the future. As an example, the EU AI Act (Article 50: Transparency

Obligations for Providers and Deployers of Certain AI Systems) requires AI content to be labeled by August 2026. [35]

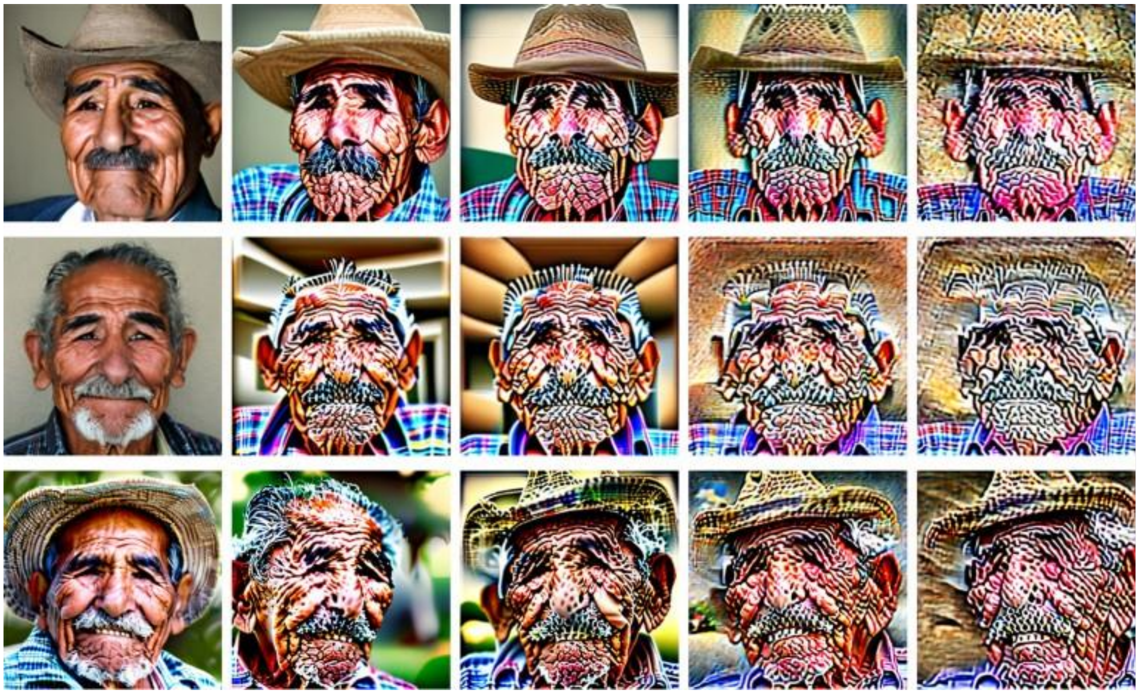


Figure 5: Figure from 'Nepotistically Trained Generative-AI Models Collapse' that shows the rapid degradation of a generative AI model trained on data generated by a previous version of the model [39]

4. Emerging trends

The amount and quality of synthetic data online will only continue to grow. Thought leaders and experts in generative AI predict up to 90% of online content will, at least in part, be synthetic by 2027. [42] In addition, this synthetic content is becoming virtually indistinguishable from real content. [43], [44] Being able to identify provenance through a solution such as Content Credentials for this staggering amount of content will be imperative for safeguarding the broader information environment.

Use cases for organizations

The authoring agencies identified the following use cases in which Content Credentials could benefit a variety of sectors and organizations, especially national security, defense, and law enforcement. In addition, as stated before, all organizations can use Content Credentials for the added protection of a company’s people, identities, and reputation.

Forensics

Examples: National security, law enforcement

Content Credentials could be used in crime scene photography, evidence collection, and forensic analysis. Having Content Credentials stamped at the sensor level would help establish provenance of the evidence.

In addition, Content Credentials could play an important role in supporting “proof of compromise” in cases of ransomware following a cyber breach. Organizations facing ransomware claims often struggle to validate the claims made by threat actors. By incorporating Content Credentials with their securely timestamped and cryptographically signed metadata into their content, organizations could have a reliable method for assessing the provenance of potentially compromised content and better assess the extent of their exposure.

Archival collections

Examples: National archives and records, libraries, museums

Different groups maintain a nation’s history, whether in analog or digital form, that includes billions of scanned images, video, and audio recordings. Content Credentials could help preserve the provenance of this data. However, special considerations will be needed for preserving media with known manipulations, such as Abraham Lincoln’s portrait or some of Joseph Stalin’s “disappearing people.” [45], [46]

In addition, Content Credentials could help establish the authenticity and provenance of important government and legal records, such as patents, titles, contracts, judicial proceedings, and official documents and guidance.

Scientific report integrity

Examples: National standards groups, national science foundations, research journals

Scientific reporting has also been increasingly impacted with uses of manipulated or generated media. [47], [48] This results in time-consuming investigations to ensure proper information is being presented in scientific papers. Using Content Credentials can add transparency to information before use in research and provide an avenue for investigating authenticity, though this may involve additional processes with associated costs. For example, adding Content Credentials directly to scientific images from the device that captured them, such as a microscope, would be ideal. However, this may require upgrading to a compatible device or a future iteration of hardware, which could involve additional costs.

Artificial Intelligence training data and data analytics

Examples: National archives, space, standards, government

Various organizations provide public access to data collections for universities, research institutions, or just the interested hobbyist to use. Content Credentials could help limit misuse of these data collections by indicating their provenance and their usage limitations. For example, if these data collections can be used to train AI models (as specified in Content Credentials), downstream data users will want to know what samples were AI-generated due to the model collapse issue noted previously. In addition, provenance on these large data collections also makes it more difficult for groups consuming this information to manipulate the data to produce desired results.

Overhead imagery

Examples: Government and commercial imagery providers (e.g., intelligence, weather, space, geology, geography)

Domestic and foreign governments and commercial organizations provide satellite imagery for many purposes. These organizations should consider implementing Content Credentials for media under their control to provide a chain of provenance with their analysis. This capability might be possible to implement as firmware updates on deployed systems that capture the information in the future.

How to prepare for implementing Content Credentials

Content Credentials and Durable Content Credentials are evolving and will continue to do so for the foreseeable future. Several organizations have begun the process of implementing Durable Content Credentials as described earlier in the Content Credentials and Durable Content Credentials section. In addition, the Appendix provides a table with a current snapshot of the landscape. As more groups implement Content Credentials, the community can begin to draw on lessons learned to aid in future implementations. In the meantime, organizations should take steps to prepare for implementing Content Credentials, including **maintaining unaltered metadata** throughout the media lifecycle and engaging with the open source standards community.

Since Content Credentials are provenance information stored in metadata, the first step for organizations looking to implement and/or consume Content Credentials is committing to maintaining metadata unaltered throughout internal processes, such as ingest, storage, and dissemination, if not done already. This may be a significant challenge for some entities. This also includes maintaining metadata throughout interactions with suppliers and subcontractors. Any processes that detach or strip metadata degrade the prospects of implementing and/or consuming Content Credentials.

To read and validate Content Credentials, organizations will need to detect the presence of special metadata tags and verify any signatures. For specialized deployments that are not connected to the Internet, this may require obtaining the signing verification certificates for prominent publishers. This is done routinely for disconnected networks to support software installation and verify document signatures.

The second step is to engage with the open source community, such as the C2PA and the Creator Assertions Working Group (CAWG), to learn about the technology space, be promptly informed of changes, and ask questions or provide recommendations on standards. [5], [49] Copious documentation is available to read about the evolving specification and how to use it, including “Getting started with Content Credentials.” [50] Organizations should engage with the CAI for questions about the specification or suggestions for improvement at https://discord.com/invite/CAI. [51]

To get started with implementing Content Credentials, organizations should consider the following questions upfront.

1. Where should we incorporate Content Credentials?

There are three stages at which incorporating Content Credentials can be attached to a piece of content:

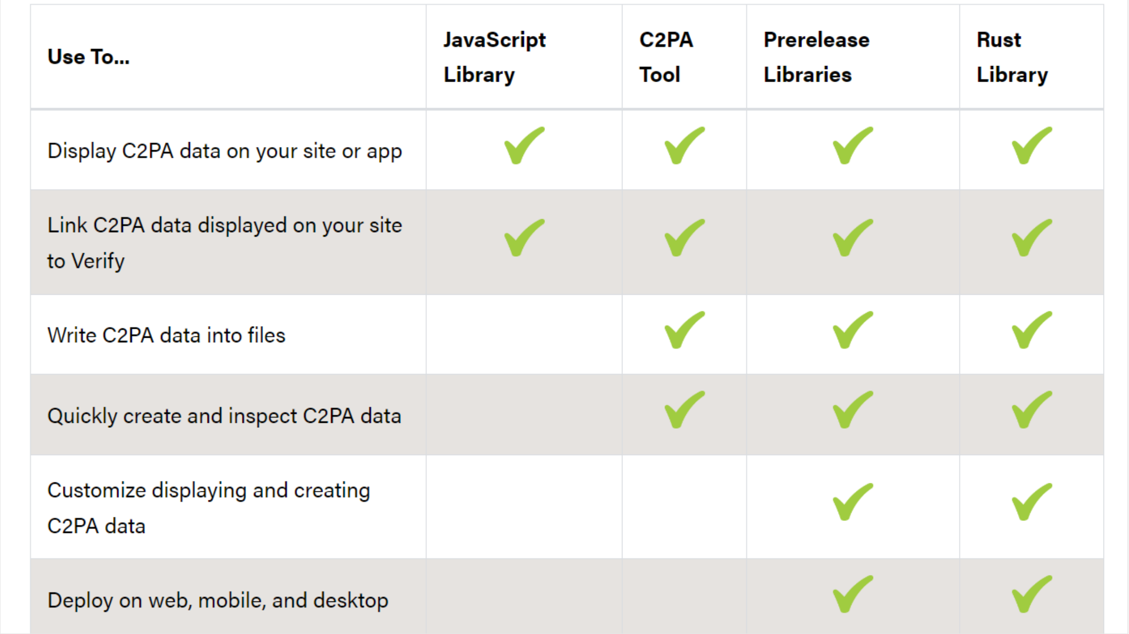
- At the sensor/upon capture
- During the editing phase
- Directly before publishing

What stage Content Credentials can be best incorporated will sometimes be beyond an organization’s control and will depend on the use case and context, among other things. For example, some organizations will not be able to purchase the cameras that are beginning to embed Content Credentials or the software that is used to edit may not yet support the addition of Content Credentials. Making the effort to provide some form of transparency to an organization’s media is a great first step regardless of the avenue.

Organizations should consider how to handle adding Content Credentials for previously captured multimedia. Since this would be done to a stored file, each organization should decide which media should have Content Credentials added. Some organizations may opt to only add credentials to media that has been previously published, whereas others may opt to add credentials to every stored media asset for preservation.

2. How can we effectively display that media has Content Credentials?

Content Credentials on their own only store the provenance information. To have an impact, that information should be presented somehow to end users. At this time, if organizations plan on displaying Content Credentials for media on websites, they may choose to implement the C2PA JavaScript Software Development Kit (SDK) (Figure 6) to display the Content Credentials logo (Figure 7). [52], [53] If unable to use the SDK, organizations can direct consumers to the Chrome® plug-ins accessible in the Chrome store. [54] This will allow viewers using the Chrome browser to see if the media uploaded to the website has credentials. In addition, Drupal® (a content management system used by around 1.7 million websites) can now process and display Content Credentials. [55]



Use To...	JavaScript Library	C2PA Tool	Prerelease Libraries	Rust Library
Display C2PA data on your site or app	✓	✓	✓	✓
Link C2PA data displayed on your site to Verify	✓	✓	✓	✓
Write C2PA data into files		✓	✓	✓
Quickly create and inspect C2PA data		✓	✓	✓
Customize displaying and creating C2PA data			✓	✓
Deploy on web, mobile, and desktop			✓	✓

Figure 6: Summary of tools offered by C2PA and their capabilities (from https://opensource.contentauthenticity.org/docs/introduction). [56]

An additional step consumers can take is to use the Content Credentials verify site; however, this tool is also evolving as the standard and trusted sources do. [16] The authoring agencies encourage consumers to share any issues encountered with the community, such as through the community discord channel at https://discord.com/invite/CAI. [51]

Figure 7: Content Credentials Logo (from https://contentcredentials.org/) [53]

Advertising the use of Content Credentials can also be powerful. Some organizations have started blogging about their efforts and the path taken to improve consumer trust through multimedia transparency. [57]

3. Where do we store media?

It is important for an organization to securely store its media for verification purposes, preferably in read-only format. This will enable the ability to look up past media if needed. For example, if the media is copied from the original location and re-purposed somewhere else with the metadata stripped, having an accessible database to find the original media with its original credentials is important for verification and provenance. Adding the additional security layers offered by Durable Content Credentials to this database can provide enhanced security as well.

4. What additional measures can be taken to make Content Credentials durable?

Beyond content metadata, the C2PA specification now supports the implementation of Durable Content Credentials in the latest specification, which enables creators to add a watermark that links the content's provenance and a reference to a fingerprint that allows for accurate retrieval of Content Credentials from an organization's database using a visual check. [2] There are both open source watermarks as well as proprietary watermarks that can be used for this purpose in multimedia. [58], [59], [60] More research and case studies from organizations implementing Durable Content Credentials will continue to arise as this solution evolves.

Current Limitations

Establishing trust in a multimedia object is a hard problem. Trust is multi-faceted and includes the parameters of who created the content, when and where it was created, how it was created, why it was created, and how it is presented. Within this larger problem, people’s general instinct often is to believe what they see or hear, making media a prime candidate for sowing distrust. [61]

Even though the ability to manipulate media is not new, with the increased accessibility and speed of media manipulation and generation tools, the barrier to entry for entities seeking to deceive is now lower than ever. Content Credentials are an important component in providing more information about media and its provenance that can answer the who, when, where, and how parameters of trust, but not the why. Content Credentials can help end users make informed decisions about the media they consume but will require continual iteration and improvement for this complex socio-technical problem. As discussed, Content Credentials, by themselves, will not solve the problem of transparency entirely. But, as they are more widely adopted, including in hardware, and as the specification evolves, they will enable consumers of information to better evaluate the authenticity of content, potentially helping to foster trust in media content.

As C2PA has discovered or was made aware of security imperfections, they have promptly corrected them in subsequent releases, as is expected of an evolving standard. An example of an early specification flaw that was identified was that people could sign media with self-signed certificates and those certificates were verified as a trusted source. This is no longer the case. A self-signed certificate is now identified as an untrusted source as C2PA now uses a list of known certificates to determine if a credential was issued by a known certificate. [62]

Some media challenges cannot be addressed by Content Credentials or other provenance solutions. For example, Content Credentials do not protect against media being removed from an archive and that provenance fact being lost. Additional measures would be required to address this concern.

Another challenge related to Content Credentials and provenance presents itself when the metadata is modified or removed. Content Credentials are cryptographically signed, so any modifications to that signed portion of the media invalidates the signature. This concern is why the concept of Durable Content Credentials has been introduced to help retrieve Content Credentials metadata if it has been removed.

Community involvement is crucial to mature this evolving specification.

Conclusion

In today's digital age, it is necessary to strengthen integrity in media with provenance solutions like Content Credentials. Advanced tools that allow the easy creation, alteration, and dissemination of digital content are now more accessible and sophisticated than ever before. This escalation threatens organizations' security, with AI-generated media being used for impersonations, fraudulent communications, and brand damage. Therefore, restoring transparency has never been more urgent.

To achieve increased transparency in digital media, secure and widespread adoption of standard practices is crucial. Emerging standards, such as Content Credentials developed by the C2PA and implemented by the CAI, are gaining traction and guiding how media provenance should be recorded and managed across the ecosystem. These are evolving standards that are constantly being updated and adapted to new security threats and vulnerabilities as they arise. However, this should not deter organizations from taking the first steps mentioned in this document to preserve media provenance.

Understanding this technology space, including the importance of early adoption and current use cases, as well as considerations for internal implementation, will help organizations mitigate the various risks of AI-generated media and deepfakes and better prepare for a future that contains mostly synthetic media. Experimenting with or adopting content provenance technologies early can help to identify issues and enable content consumers determine authenticity, which will become increasingly difficult as AI-generated online content proliferates. Content Credentials are still evolving, so organizations should stay engaged in the community to maintain awareness of new security issues and best practices. Ultimately, continuing to advance all aspects of this multifaceted topic—provenance, education, policy, and detection—will be crucial to strengthening multimedia integrity in the generative AI era.

Further resources

- Audiovisual Generative AI and Conflict Resolution: Trends, Threats and Mitigation Strategies

Works cited

[1] Adobe. Content Credentials. 2024. https://helpx.adobe.com/creative-cloud/help/content-credentials.html

[2] Content Authenticity Initiative (CAI). Three pillars of provenance that make up durable Content Credentials. 2024. https://contentauthenticity.org/blog/three-pillars-of-provenance

[3] National Security Agency (NSA), Federal Bureau of Investigation (FBI), Cybersecurity and Infrastructure Security Agency (CISA). Contextualizing Deepfake Threats to Organizations. 2023. https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF

[4] WITNESS. Audiovisual Generative AI and Conflict Resolution: Trends, Threats, and Mitigation Strategies. 2024. https://www.gen-ai.witness.org/wp-content/uploads/2024/08/WITNESS-Report_Audiovisual_Generative_AI_and_Conflict-1.pdf

[5] Coalition for Content Provenance and Authenticity (C2PA). Overview – C2PA. 2024. https://c2pa.org/

[6] CAI. Restoring trust and transparency in the age of AI. 2024. https://contentauthenticity.org/

[7] National Institute of Standards and Technology (NIST). NIST AI 100-4: Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf

[8] Jenks, Andrew. Re: NIST AI 100-4, Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. 2024. https://downloads.regulations.gov/NIST-2024-0001-0030/attachment_1.pdf

[9] NIST. NIST AI RMF Playbook: Manage. 2023. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook/Manage

[10] MIT Technology Review. The inside scoop on watermarking and content authentication. 2023. https://www.technologyreview.com/2023/11/06/1082996/the-inside-scoop-on-watermarking-and-content-authentication/

[11] CAI. Durable Content Credentials. 2024. https://contentauthenticity.org/blog/durable-content-credentials

[12] Agarwal, Shruti. Project Know How - GS3-8. 2024. https://www.adobe.com/max/2024/sessions/project-know-how-gs3-8.html

[13] CAI. How it works. 2024. https://contentauthenticity.org/how-it-works

[14] Research & Development at The New York Times. Readers Have a Right to Know the True Origin of What They See, Read and Hear Online. 2021. https://rd.nytimes.com/projects/media-provenance-vision

[15] C2PA. C2PA User Experience Guidance for Implementers: Levels of information disclosure. 2024. https://c2pa.org/specifications/specifications/1.0/ux/UX_Recommendations.html#_levels_of_information_disclosure

[16] C2PA. Content Credentials: Inspect content to dig deeper. 2024. https://contentcredentials.org/verify

[17] European Commission. The Strengthened Code of Practice on Disinformation 2022. 2022. https://ec.europa.eu/newsroom/dae/redirection/document/87585

[18] Partnership on AI. PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action. 2023. https://syntheticmedia.partnershiponai.org/#read_the_framework

[19] C2PA. Membership – C2PA. 2024. https://c2pa.org/membership

[20] Truepic. Sign Content Credentials at scale. 2024. https://www.truepic.com/c2pa/signing

[21] The Verge. YouTube is rolling out a new 'captured with a camera' label that works with C2PA Content Credentials. 2024. https://www.theverge.com/2024/10/15/24271083/youtube-c2pa-captured-camera-label-content-credentials

[22] Google. How we're increasing transparency for gen AI content with the C2PA. 2024. https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/

[23] BBC. New technology to show why images and video are genuine launches on BBC News. 2024. https://www.bbc.com/mediacentre/2024/content-credentials-bbc-verify

[24] Microsoft. Content Credentials. 2024. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-credentials

[25] CAI. Our members. 2024. https://contentauthenticity.org/our-members

[26] C2PA. Coalition for Content Provenance and Authenticity (C2PA). 2024. https://c2pa.org/files/C2PA_Introduction_Deck.pdf

[27] C2PA. Content Credentials: C2PA Technical Specification. 2024. https://c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html

[28] PetaPixel. AI Images are Causing Havoc for People Affected by Hurricane Helene. 2024. https://petapixel.com/2024/10/07/ai-images-are-causing-havoc-for-people-affected-by-hurricane-helene/

[29] The Associated Press. FACT FOCUS: Fake image of Pentagon explosion briefly sends jitters through stock market. 2023. https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4

[30] Merriam-Webster. Word of the Year 2023. 2023. https://www.merriam-webster.com/wordplay/word-of-the-year-2023

[31] TIME. The Best Inventions of 2024: Battling Fake Photos - Content Credentials. 2024. https://time.com/7094554/content-credentials/

[32] CISA. Tactics of Disinformation. https://www.cisa.gov/sites/default/files/publications/tactics-of-disinformation_508.pdf

[33] Department of Homeland Security Public-Private Analytic Exchange Program. Increasing Threat of Deepfake Identities. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

[34] Congressional Research Service. Deep Fakes and National Security. 2023. https://crsreports.congress.gov/product/pdf/IF/IF11333

[35] European Union. EU Artificial Intelligence Act: Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems. 2024. https://artificialintelligenceact.eu/article/50/

[36] Balan, Kar et al. EKILA: Synthetic Media Provenance and Attribution for Generative Art. 2023. https://arxiv.org/pdf/2304.04639

[37] Adobe. Adobe Unveils Firefly, a Family of new Creative Generative AI. 2023. https://news.adobe.com/news/news-details/2023/adobe-unveils-firefly-a-family-of-new-creative-generative-ai

[38] Shumailov, Ilia et al. AI models collapse when trained on recursively generated data. 2024. https://www.nature.com/articles/s41586-024-07566-y

[39] Bohacek, Matyas and Farid, Hany. Nepotistically Trained Generative-AI Models Collapse. 2023. https://arxiv.org/pdf/2311.12202

[40] Gibney, Elizabeth. AI models fed AI-generated data quickly spew nonsense. 2024. https://www.nature.com/articles/d41586-024-02420-7

[41] The New York Times. When A.I.’s Output Is a Threat to A.I. Itself. 2024. https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html

[42] Interpol. Beyond Illusions: Unmasking the Threat of Synthetic Media for Law Enforcement. 2024. https://www.interpol.int/content/download/21179/file/BEYOND%20ILLUSIONS_Report_2024.pdf

[43] CAI. June 2024 | This Month in Generative AI: Moving Through the Uncanny Valley (Pt. 1 of 2). 2024. https://contentauthenticity.org/blog/june-2024-this-month-in-generative-AI-the-uncanny-valley

[44] CAI. July 2024 | This Month in Generative AI: Moving Through the Uncanny Valley (Pt. 2 of 2). 2024. https://contentauthenticity.org/blog/july-2024-this-month-in-generative-AI-the-uncanny-valley-part-2

[45] Digital Camera World. Abraham Lincoln vs John Calhoun: the original deepfake photo of a US president. 2024. https://www.digitalcameraworld.com/features/abraham-lincoln-vs-john-calhoun-the-original-deepfake-photo

[46] Blakemore, Erin. How Photos Became a Weapon in Stalin’s Great Purge. 2022. https://www.history.com/news/josef-stalin-great-purge-photo-retouching

[47] Society for Scholarly Publishing. Guest Post: The Future of Image Integrity in Scientific Research. 2024. https://scholarlykitchen.sspnet.org/2024/04/23/guest-post-the-future-of-image-integrity-in-scientific-research/

[48] Moreira et al. SILA: a system for scientific image analysis. 2022. https://pubmed.ncbi.nlm.nih.gov/36316363/

[49] Creator Assertions Working Group. Creator Assertions Working Group. 2024. https://cawg.io/

[50] CAI. Getting started with Content Credentials. 2024. https://opensource.contentauthenticity.org/docs/getting-started/

[51] CAI. Content Authenticity Initiative Discord Channel. 2024. https://discord.com/invite/CAI

[52] CAI. JavaScript library. 2024. https://opensource.contentauthenticity.org/docs/js-sdk/getting-started/overview/

[53] C2PA. Content Credentials. 2024. https://contentcredentials.org/

[54] Chrome Web Store. Adobe Content Authenticity. 2024. https://chromewebstore.google.com/detail/adobe-content-authenticit/dmfbmenkapmaoldfgacgkoaoiblkimel

[55] CAI. c2pa-drupal. 2024. https://github.com/contentauth/c2pa-drupal

[56] CAI. CAI open source SDK. 2024. https://opensource.contentauthenticity.org/docs/introduction/

[57] CAI. Community Story Pixelstream. 2022. https://contentauthenticity.org/blog/community-story-pixelstream

[58] Adobe. Trustmark – Universal Watermarking for Arbitrary Resolution Images. 2024. https://github.com/adobe/trustmark

[59] Google. SynthID: Identifying AI-generated content with SynthID. 2024. https://deepmind.google/technologies/synthid/

[60] Digimarc, Guinard. C2PA 2.1—Strengthening Content Credentials with Digital Watermarks. 2024. https://www.linkedin.com/pulse/c2pa-21-strengthening-content-credentials-digital-dominique-guinard-sozie

[61] Psychology Today. Misjudgement: Why We Trust What We See vs. What We Hear. 2023. https://www.psychologytoday.com/us/blog/decisions-that-matter/202310/misjudgement-why-we-trust-what-we-see-vs-what-we-hear

[62] CAI. Verify tool known certificate list. 2024. https://opensource.contentauthenticity.org/docs/verify-known-cert-list/

Disclaimer of endorsement

The information and opinions contained in this document are provided "as is" and without any warranties or guarantees. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the United States Government, and this guidance shall not be used for advertising or product endorsement purposes.

Trademark recognition

Adobe, Content Authenticity Initiative, Lightroom, and Photoshop are registered trademarks, and Adobe Firefly and Content Credentials are trademarks of Adobe Inc. • Amazon is a registered trademark of Amazon Technologies, Inc. • BBC is a registered trademark of The British Broadcasting Corporation. • Bing and Microsoft are registered trademarks of Microsoft Corporation. • Camera Bits and Photo Mechanic are registered trademarks of Camera Bits, Inc. • Canon is a registered trademark of Canon Kabushiki Kaisha. • Chrome and Google are registered trademarks of Google LLC. • Drupal is a registered trademark of Dries Buytaert. • Facebook and Meta are registered trademarks of Meta Platforms, Inc. • Fujifilm is a registered trademark of Fujifilm Corporation. • Getty Images is a registered trademark of Getty Images (US), Inc. • Intel is a registered trademark of Intel Corporation. • ISO is a registered trademark of the International Organization for Standardization. • Leica is a registered trademark of Leica Microsystems IR GmbH. • LinkedIn is a registered trademark of LinkedIn Corporation. • Merriam-Webster is a registered trademark of Merriam-Webster, Incorporated. • Nikon and Nikon Z cameras are registered trademarks of Nikon Corporation. • OpenAI is a trademark of OpenAI, Inc. • Samsung and Samsung Galaxy are registered trademarks of Samsung Electronics Co., Ltd. • ShutterStock is a registered trademark of Shutterstock, Inc. • Sony is a registered trademark of Sony Group Corporation. • DALL·E is a registered trademark and Sora is a trademark of OpenAI OpCo, LLC. • Time is a registered trademark of TIME USA, LLC. • Truepic is a registered trademark of TruePic Inc.

Purpose

This document was developed in furtherance of the authoring agencies' cybersecurity missions, including their responsibilities to identify and disseminate threats, and to develop and issue cybersecurity specifications and mitigations. This information may be shared broadly to reach all appropriate stakeholders.

Contact

NSA: Cybersecurity Report Feedback: CybersecurityReports@nsa.gov
Defense Industrial Base Inquiries and Cybersecurity Services: DIB_Defense@cyber.nsa.gov
Media Inquiries / Press Desk: NSA Media Relations: 443-634-0721, MediaRelations@nsa.gov

Australian organizations visit cyber.gov.au/report or call 1300 292 371 (1300 CYBER1) to report cybersecurity incidents and vulnerabilities.

Canadian organizations report incidents by emailing the Cyber Centre at contact@cyber.gc.ca.

United Kingdom organizations report a significant cyber security incident: ncsc.gov.uk/report-an-incident (monitored 24 hours) or, for urgent assistance, call 03000 200 973.

Appendix: Current Content Credentials implementation

The information presented in this appendix is expected to change rapidly and was current as of the time of preparation. Monitor individual companies' press releases for the most up-to-date information. This list is presented here for a sense of current support for implementation and as references for NSS, DoD, DIB, or other organizations to review to be able to start testing and using Content Credentials now. There is no implied endorsement of any company or product listed. In addition, if there are updates that should be made to this information, please contact the email address listed in the contact information for this CSI.

Hardware	Announced Intent to Implement Content Credentials	Implemented Content Credentials	General Availability	Reference
Leica® Cameras	Yes	Yes/Hardware	M11-P (26-Oct-2023) M11-D (12-Sep-2024) SL3-S (16-Jan-2025)	https://leica-camera.com/en-US/photography/content-credentials https://contentauthenticity.org/blog/content-credentials-arrives-in-the-leica-sl3-s-camera
Sony Cameras	Yes	Yes/Firmware (27-Mar-2024)	No, requires license, available for photojournalists	https://www.prnewswire.com/news-releases/sony-electronics-delivers-firmware-updates-including-c2pa-compliancy-as-a-next-step-to-ensure-authenticity-of-images-302101879.htm https://authenticity.sony.net/camera/en-us/index.html
Fujifilm®	Yes, 25-May-2024	No	No	https://www.fujirumors.com/fujifilm-to-introduce-c2pa-content-authenticity-to-x-and-gfx-cameras/
Canon®	Proof of Concept and C2PA Member	No	No	https://www.thomsonreuters.com/en/press-releases/2023/august/reuters-new-proof-of-concept-employs-authentication-system-to-securely-capture-store-and-verify-photographs.html
Nikon®	Yes, 14-Oct-2024, Z6III	No	No	https://www.nikon.com/company/news/2024/0109_imaging_02.html
Truepic Software on Smartphones	Yes	Yes	Not General, Requires Token from Truepic Customer, i.e. Insurance Company*	https://www.truepic.com/c2pa/capture https://www.truepic.com/c2pa/signing
Samsung®	Yes, 24-Jan-2025	Yes	Samsung Galaxy® S25 (Release, 7-Feb-2025)	https://news.samsung.com/global/galaxy-unpacked-2025-highlights-from-galaxy-unpacked-a-new-era-of-ai-integration

Software	Announced Intent to Implement Content Credentials	Implemented Content Credentials	General Availability	Reference
Adobe LightRoom®, PhotoShop®, Camera Raw Plugin	Yes	Yes	Yes, 7-Jul-2024 (Out of Beta)	https://www.dpreview.com/news/8360583669/adobe-adds-cai-content-credentials-option-to-camera-raw
Photo Mechanic® by Camera Bits®	Yes 15-May-2024	Yes, Demo	No	https://petapixel.com/2024/05/15/photo-mechanic-integrates-content-credentials-to-fight-fake-imagery/
ExifTool (MetaData Analysis Command Line)	Yes	Yes	Yes, Detection not Verification*	https://exiftool.org/exiftool_pod.html Best with 13.0 versions or higher, minimum 12.70
LibExiv2 (MetaData Analysis Library)	Feature Request	No	No	https://github.com/Exiv2/exiv2/issues/2221 Used by GIMP, DarkTable, other Open Source

Generative AI	Announced Intent to Implement Content Credentials	Implemented Content Credentials	General Availability	Reference
DALL·E®, Sora from OpenAI (GenAI)	Yes	Yes	Yes	https://help.openai.com/en/articles/8912793-c2pa-in-dall-e-3
Bing® from Microsoft (GenAI)	Yes	Yes	Yes	https://www.bing.com/images/create/help
Adobe Firefly™	Yes	Yes	Yes	https://www.adobe.com/products/firefly.html

Social Media	Announced Intent to Implement Content Credentials	Implemented Content Credentials	General Availability	Reference
Facebook®/Meta	Yes	No	No	https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/
Google	Yes	No	No	https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/
LinkedIn® (Microsoft)	Yes	Yes	Yes	https://news.linkedin.com/2024/May/linkedin-rolls-out-c2pa-ai-generated-content-standard

Image Brokers	Announced Intent to Implement Content Credentials	Implemented Content Credentials	General Availability	Reference
Adobe Stock	Yes	Yes	Yes	https://blog.adobe.com/en/publish/2024/01/26/seizing-moment-content-credentials-in-2024

Image Brokers	Announced Intent to Implement Content Credentials	Implemented Content Credentials	General Availability	Reference
ShutterStock®	Yes	No	No	https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-integrates-creative-ai-library-700m-images-offer
Getty Images®	Will Use C2PA to filter generative AI content			https://www.theverge.com/2022/9/21/23364696/getty-images-ai-ban-generated-artwork-illustration-copyright