Cryptographic Provenance and AI-generated Images

Jessica Bushey
School of Information
San José State University
San José, CA, USA
ORCID: 0000-0002-8569-5360

Nicholas Rivard
School of Computer Science
Carleton University
Ottawa, ON, Canada
ORCID: 0009-0002-3068-5496

Michel Barbeau
School of Computer Science
Carleton University
Ottawa, ON, Canada
ORCID: 0000-0003-3531-4926

***Abstract*— In a world of proliferating synthetic multimedia content, it is increasingly important to develop the ability to trace the origin of digital assets and verify their authenticity. A bold initiative is the Coalition for Content Provenance and Authenticity (C2PA), which proposes a data model for associating authenticated provenance information, known as content credentials, with multimedia assets. This paper situates C2PA within the field of Computational Archival Science (CAS), examining how cryptographic provenance and authenticity frameworks can operationalize archival trustworthiness in big data environments. With a focus on digital images (include AI-generated), this paper explores the creation of a C2PA-based pipeline for digital asset preservation informed by archival science, computer science and cybersecurity. Using an analytical framework derived from archival diplomatics and computational provenance modeling, the study maps authenticity metadata for digital images. The role of C2PA in records management and archival preservation of AI-generated images is introduced. The use of emerging blockchain technology to support provenance of multimedia content is discussed in detail. Alternative solutions are discussed. C2PA security risks are reviewed. The findings demonstrate that C2PA represents a computational model of provenance – transforming authority-based trust into trust by design. This contribution addresses a key CAS research challenge of establishing provenance for digital assets across distributed systems.**

Keywords—provenance, C2PA, blockchain, computational archival science, AI-generated Images

I. Introduction

The authority of archivists is predicated on their role as arbiters of documentary truth. Their actions are guided by a code of professional ethics and established procedures designed to manage information and records in a manner that preserves and protects both content and context. Researchers value photographic records in archival institutions for their capacity to witness the past. The ongoing use of images in news media, medical diagnostics, and law enforcement as visual evidence of real events, people, and places is being questioned once again with the rise of Artificial Intelligence (AI) technology. AI-generated images pose unprecedented risks due to their ease and speed of creation, as well as their realistic image quality. While the images themselves are not necessarily problematic, their dissemination and use without accurate labelling and/or metadata can mislead audiences and obscure the context of their creation.

As synthetic images join the visual flow circulating through news and social media platforms, it is increasingly complex to ascertain fact from fiction. "Viral" sharing and reuse of images online quickly obscure their provenance and original function, which can lead to new interpretations and applications with unforeseen consequences. The inability to determine if an image is synthetic or not impacts public understanding of current events and introduces epistemological doubt. On a more practical level, it has the potential to complicate existing recordkeeping approaches to acquiring, preserving, and providing access to trustworthy images (i.e., reliable, accurate and authentic). Within the scope of CAS, these problems are understood not only as social and evidential issues, but also computational ones. CAS seeks to explore the relationships between archival concepts of provenance and large-scale data structures. In this sense, the issues raised by AI-generated images provide an opportunity for research into application of computational provenance models and recordkeeping practices.

In response to the threats posed by AI-generated content, global media and technology companies launched the Content Authenticity Initiative (CAI) [1], and have developed a trusted, end-to-end media ecosystem, referred to as the Coalition for Content Provenance and Authenticity (C2PA) specification [2]. Identified in a prior literature review [3] on AI-generated images as an emergent record type, the C2PA approach to establishing and managing verifiable provenance metadata for born-digital assets is relevant to archival organizations that perform authentication activities through archival description and preservation. Essential to the approach is the C2PA Manifest [2] - a set of information about the provenance of an asset (i.e., the history of an asset and its interaction with actors and other assets), along with a claim (i.e., a digitally signed and tamper-evident data structure) and a claim signature (i.e., the digital signature created using a private key owned by a signer). The preferred non-technical term for C2PA Manifest(s) as well as the overall C2PA technology is "Content Credentials", promoted by CAI as a digital nutrition label, or an ingredient list for media that provides transparency about "who created the content and how, when, and where it was created… and how it was edited" [1].

This study examines the C2PA specification through a CAS lens to determine whether its architecture effectively encodes properties that establish and protect archival authenticity. By mapping C2PA's cryptographic provenance mechanisms with archival concepts, the research demonstrates how system design can instantiate archival requirements. This article responds to the call for an analysis and evaluation of the C2PA metadata

specification to better understand it as a trust framework that supports the capture of provenance data and establishes a secure metadata profile for digital images (including those generated by AI) that is tamper-evident and can be used to verify authenticity so that consumers and producers of online images can better understand the nature of digital communications and their intended uses [3]. Building on prior work [3, 4, 5], the authors present a transdisciplinary approach that utilizes computational archival science methodology to explore a C2PA pipeline for creating and preserving digital assets, particularly images generated by-AI algorithms. The guiding research question is therefore: How does C2PA serve as a computational implementation of archival authenticity, and what are the implications for archival appraisal, arrangement and description and preservation of digital and AI-generated images?

By leveraging available tools to capture digital images with content credentials, the authors present the results from a small dataset. From an archivist's perspective, the goal of this work is to determine if content credentials can provide a durable manifest that presents the totality of actions and actors that have contributed to the creation and use of digital images. Based on the rationale that in the near future, archivists will acquire fonds that contain born digital images (both real and synthetic) and it will be necessary to determine their provenance through methodologies and tools that recognize and utilize available technical, descriptive and administrative metadata, trust signatures, and cryptographic hashes that are bound to digital images because of their creation and use within C2PA compliant hardware, software and platforms. Framed in this way, the paper contributes to CAS by testing how archival theory can be rendered computationally. To do this, the authors draw on knowledge and methodologies from archival science, computer science and cybersecurity to examine content credentials and identify their potential application in recordkeeping and archives. The rapidly evolving media ecosystem requires archivists to engage and explore the consequences of emerging technologies and current trends aimed at establishing trust through transparency. This research is conducted under the auspices of the InterPARES Trust AI Project (2021-2026), a multinational, interdisciplinary team of researchers and practitioners exploring the application of AI to archives with the aim of ensuring the creation, management and preservation of trustworthy public records.

In Section II, we review related work from the transdisciplinary field of computational archival science. Section III discusses content credentials and presents the results of metadata mapping for digital images. In Section IV, we review a blockchain architecture for content credentials. Section V reviews the security of C2PA. We conclude with Section VI.

II. Related Work

An initial review of the literature on AI-generated images as an emergent record format revealed an absence in archival journals [3]. The rationale at the time was the recent availability of generative AI tools to produce photorealistic content and the omission of AI-generated images being managed and preserved by archival institutions. In contrast, a growing body of literature on the application of AI technologies and tools in recordkeeping and archives to support automation of appraisal, digitization, archival description, and enhanced access to online collections can be found in records management and archival journals [6, 7, 8], computational archival science conferences [9, 10] and digital humanities journals [11, 12]. By casting a wider net to include keyword searches in the Library and Information Science Source (LISS) database and Factiva database (global news content covering 18,000 sources from 159 countries), sixty-one sources were selected to form the basis of the literature review on AI-generated images as an emergent record format [3]. These sources addressed how AI-generated images were being approached in journalism and media communications, medical diagnostics and law enforcement. They provided many insights, including new terminology (e.g., synthetic images), an industry-led technical standard for establishing and certifying the digital provenance of media (e.g., C2PA), and public consultation on copyright policy directions regarding AI-generated works in Canada [13] and the United States [14].

Using an archival lens, Bushey [3] discussed the potential of C2PA as a technical standard for binding cryptographically secure provenance data, referred to as a claim, to digital assets. This metadata includes information about the creation device, the editing software used, and the identity of the authoring party. Such metadata is critical for establishing the authenticity (i.e., identity and integrity) of a digital image from creation through use.

Furthermore, Bushey suggested that C2PA metadata should be analyzed to determine if it fulfills the InterPARES 2 requirements for the creation and preservation of trustworthy digital records, as outlined in the Creator Guidelines [15] and Preserver Guidelines [16]. Bushey [4] draws on prior studies of the reliability and authenticity of born-digital images [17] and the archival trustworthiness of social media images [18] to demonstrate the effectiveness of archival diplomatics as a methodology for analyzing image metadata and clarifying which elements constitute supporting evidence for establishing the identity and integrity of a digital image from creation through preservation, recognizing that an ongoing challenge has been the lack of consistent industry support for writing and reading embedded metadata in image files through capture devices and imaging software. Additionally, metadata can be inadvertently removed from digital image files during online dissemination and use. These earlier findings establish a foundation for the present study's CAS-oriented inquiry, where archival diplomatics is used not only to interpret metadata conceptually, but also to evaluate its computational implementation with the C2PA framework.

In contrast, the growing global concern over misinformation has galvanized camera manufacturers (e.g., Leica, Sony, Panasonic), imaging software (e.g., Adobe), and social media platform providers (e.g., Meta, TikTok) to unite under the CAI and integrate C2PA architecture and its ecosystem into their products and services.

A recent search for literature on AI-generated images reveals ongoing concern with synthetic images circulating online and the risk they pose to information integrity and cybersecurity [19]. Discussions focus on technological approaches to identify image provenance and protect digital image integrity. Potential solutions include visible and invisible digital watermarks [20-

23] and the C2PA framework [24-28]. Emphasis is placed on the importance of integrating C2PA into existing workflows [29] and the role and responsibility of the public in verifying the authenticity of online images and making their own trust decisions based on the available context of creation and use [25, 26]. In addition to these sources, a significant body of promotional and instructional literature is being produced by CAI regarding content credentials and the Adobe-led Verify app [30]. Access to use cases is currently limited as news media organizations and individuals slowly adopt and develop hardware and software that integrate content credentials. Since 2022, the Starling Media Lab for Data Integrity (a joint initiative between Stanford University and the University of Southern California) has been using open-source tools, image standards and encryption in case studies to securely capture, store, and verify digital content in different organizational contexts [31]. The case studies focus on three communities of practice: journalism, history, and law. The Starling Lab case studies have utilized C2PA and Web3 technologies, specifically cryptography and blockchain, along with the Verify tool, to establish provenance, protect digital integrity, and authenticate images and videos disseminated online.

In the 2022 journalism case study [32], "Documenting Stockton's Homeless," key takeaways were the considerable cost, time, and tools to develop the required ecosystem to support photographers and newsroom staff, and that the work towards building software and platforms needs to continue. When discussing digital archives and the long-term preservation of digital images and video, the Starling Lab acknowledged the challenge of bit rot. When addressing the issue of gradual degradation or loss of data integrity over time in digital storage systems, their solution employs a decentralized, cryptographic approach, along with redundant cloud-based storage. The Starling Lab and Reuters case study "78 Days" [33], does not address the role of archival institutions in this evolving ecosystem of digital provenance and verification. Still, it does highlight the challenges posed by the dissemination of mobile phone images online (and utilizing centralized platforms controlled by Apple and Google) to the authentication and preservation of, and access to aggregates of digital images over the long term. These case studies serve as prototypes for how the C2PA framework could be deployed to provide transparency at scale and informed trust in the online environment. They also demonstrate how providing public access to an image's provenance and a link to cryptographic records that log authenticity metadata can allow the public to independently verify the image. In this way, C2PA embodies trust by design.

Taken together, this body of literature demonstrates both the technical maturity of C2PA and the conceptual gap that remains for archival and CAS research. What has yet to be explored is whether such cryptographic provenance frameworks can sustain the evidential, ethical and contextual requirements established by the recordkeeping community.

III. Content Credentials

Previous studies [17, 18] have demonstrated that archival diplomatics can assist archivists in identifying the persons, contexts and instruments behind digital image creation and contribute to a more informed understanding of how photographic images are made, their different uses and the circumstances of their management and preservation. The archival diplomatic framework establishes that the trustworthiness of photographic records is determined by their reliability, accuracy and authenticity (itself comprised of identity and integrity). The widespread use of smartphone photography and social media platforms for dissemination introduced new combinations of software, hardware, and exchange protocols, as well as new roles and responsibilities regarding image creation, management, and stewardship [18]. As a result, established procedural controls over the creation, use, management and preservation of photographic records were disrupted.

Professional adherence to the use of digital image metadata to provide technical and descriptive information, which contributes to the identity and integrity of images and provides a method for showing the reliability of an image and establishing its authenticity, was largely unsupported by social media platform systems. The removal of image metadata during actions of uploading and downloading digital files into and out of social media platforms presents a real threat to establishing and maintaining the trustworthiness of digital images circulating online. Within a CAS framework, this loss of metadata can be understood as a loss of computable provenance – an interruption in the chain of preservation.

The advent of content credentials and their adoption by media and communication industry leaders and implementation into the design of instruments and online systems that create, exchange and disseminate digital images is an important development, one that invites scrutiny through the transdisciplinary lens of computational archival science. This article provides a preliminary investigation into the scope and durability of the C2PA specification and provides insight into an essentially cryptographic provenance framework, one that allows a digital image to carry multiple manifests that are each individually signed, cryptographically sealed and linked to its predecessors, forming a chain of provenance events. This multi-manifest structure demonstrates that change can be documented and validated rather than treated as corruption. It also recognizes the online practice of image re-use and re-distribution.

Drawing from InterPARES [15, 16] and ISO 15489: Records Management and ISO 16263: Certification of Trustworthy Digital Repositories, in digital environments, the authenticity of a digital image (including AI-generated images) is comprised of its identity metadata and conveyed by its integrity metadata. Identity metadata comprises the properties that convey the unique identity of a digital image—who created it, when and where it was made, in what form, and for what purpose—distinguishing it from all other digital assets. Integrity metadata, by contrast, refers to the properties that convey the completeness and unalterability of the digital image over time. The following tables address identity metadata and integrity metadata for digital images (both real and synthetic) and map each element to relevant C2PA v2.2 assertions [2]. In some cases, C2PA relies on existing image metadata schemas to populate the assertion, such as Exif technical metadata, IPTC Image metadata and Dublin Core. This highlights how different photographic practices will result in more or less metadata, and that requirements for metadata will reflect both immediate needs and

future use, including accountability and evidentiary purposes. These mappings demonstrate C2PA's technical capacity for encoding metadata that establishes and protects the authenticity of digital images throughout their lifecycle.

TABLE I. IDENTITY METADATA FOR DIGITAL IMAGES

Identity Metadata Elements	**C2PA Field / Assertion**
Name of persons involved in the creation of the digital image.	claimSignature (signer identity via X.509); claim_generator_info (non-human-actor).
Name of the action or matter	c2pa.metadata (e.g., dc:title, photoshop:Headline)
Place	c2pa.metadata (Exif GPS field)
Documentary form	c2pa.actions (first action c2pa.created with digitalSourceType); c2pa.metadata (Iptc4xmpExt:DigitalSourceType); c2pa.thumbnail; claim_generator_info
Digital presentation	c2pa.metadata (dc:format)
Date(s) of creation, use, re-use	c2pa.metadata (photoshop:DateCreated, Exif timestamps); Time-stamp on claim signature (timeStampToken)
Expression of documentary context	c2pa.metadata (optional description fields)
Copyright or IPR	c2pa.metadata (IPTC rights/admin fields); claim_generator_info (tool identity)
Other forms of authentication (e.g., watermark)	Soft binding & recovery indicators within Ingredient; softBindingAlgorithmsMatched
Version number	c2pa.actions.v2; c2pa.ingredient.v3; Manifest URN version suffixes
Identification of duplicates	Ingredient.hash (dedupe); c2pa.hash.data for hard binding
Prompt (for AI-gen images)	c2pa.actions.parameters.inputs. role="prompt"

TABLE II. INTEGRITY METADATA FOR DIGITAL IMAGES

Integrity Metadata Elements	**C2PA Field / Assertion**
Name of handling office	claim signature (X.509 identity)
Name of office of primary responsibility (Issuer)	claimSignature (certificate issuer)
Indications of annotations added to the digital materials	c2pa.actions (edit operations)
Indications of technical changes to the files or the system	c2pa.actions + hard binding assertions c2pa.hash.data - SHA-256
Controls over access	Hard binding assertions; validation states (Well-Formed, Valid, Trusted)
Globally Unique Identifier	Manifest URN (urn:c2pa:<UUID> (for each manifest, version & generator)
Persistence (e.g., fingerprints)	External manifests discoverable via HTTP Link: rel=c2pa-manifest; activeManifest references; claim/signature hashes

IV. CONTENT CREDENTIALS AND BLOCKCHAIN

Past research and development focused on combining photograph content credentials and blockchain for provenance preservation. This combination aims to provide an infrastructure for the traceability of images and verification of their authenticity. Target application fields include image and video data security [34], medical imaging [35, 36], stock photo websites and peer-to-peer image sharing [37], industry [38], and social networks [39]. C2PA-based implementations exist at both the prototyping stage, such as Photrace [40], and the commercial stage, such as the Numbers Protocol and Capture infrastructure [41], which is further discussed in the sequel.

A typical C2PA blockchain pipeline for preserving digital assets involves i) a digital asset model, ii) a digital asset storage, iii) a blockchain, and iv) an asset creation procedure.

A. Digital Asset Model

A digital asset can be an image, a video, an audio file, or any other type of file. In this paper, we focus on images. A sample digital asset is pictured in Figure 1. It is a JPEG image. At the top right, there is a provenance data block. Firstly, the Nid (Numbers ID) uniquely identifies the asset. It is generated when the asset is created. It is also called the Content Identifier (CID). There is also the identity of the creator of the asset, the date it was made, and the location where it was created. In this case, the digital asset was created by uploading an image. The second block of text, located on the right side of the top, comprises an Integrity Proof used to sign and verify the authenticity of the asset. The Integrity Proof is defined in the sequel. For archivists, the Integrity Proof represents a form of computational fixity statement – evidence that the bitstream and its metadata have not been altered since certification.

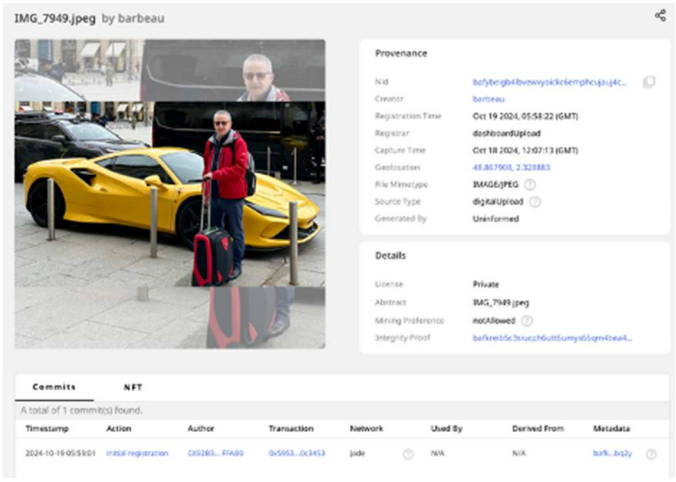

Figure 1. Sample digital asset.

B. Digital Asset Storage

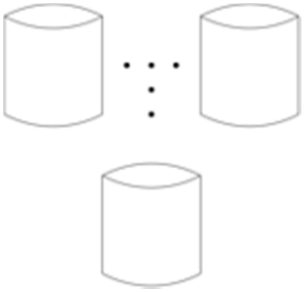

Figure 2. Distributed storage for digital assets.

As shown in Figure 2, the permanent storage of digital assets is achieved using a distributed file system. The expression *pinned in* refers to an image in a persistently stored state. For instance, the Capture provenance infrastructure utilizes the InterPlanetary File System (IPFS).

C. Blockchain

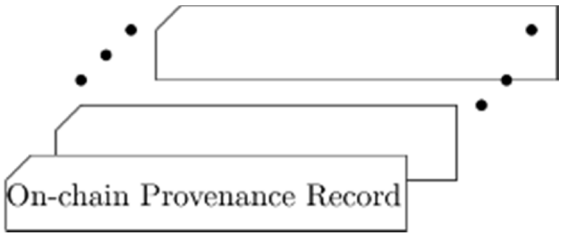

Figure 3. Blockchain of provenance records.

A blockchain stores on-chain provenance records for verifiability and traceability of digital assets. Every on-chain provenance record is linked with a digital asset. The relationship is established by storing a cryptographic hash of the digital asset in the record. A cryptographic hash is a one-way function $h()$ that derives a unique value $h(a)$ from the digital asset a. This calculation is straightforward, but generating a digital asset that matches a given value is challenging. This means that breaking the relationship between the hash value and the digital asset is a challenging problem. This cryptographic hash constitutes the CID or Nid, i.e., $Nid = h(a)$, see also Figure 1. For instance, the Capture provenance infrastructure uses the Numbers Mainnet Blockchain to store on-chain provenance records.

D. Asset Creation Procedure

Every asset creation architecture user has i) an identity (id), ii) a private key (e_{id}), and ii) a public key (d_{id}). The user must never divulge its private key, which is used to sign on-chain provenance records. The public key is used to verify the authenticity of provenance records. In the asset creation process, an entity, e.g., a photographer, plays the role of creator. They must have an identity, a private key, and a public key. There is no need to emphasize the importance of a photographer keeping their private key secret to prevent identity theft in this architecture.

For example, on Capture, the asset creation procedure consists of the following steps.

Integrity Proof Creation. The integrity proof p is a triple comprising the cryptographic hash of the asset, its type (e.g., jpeg image), and creation time, i.e., $p = (h(a), type, time)$. Figure 1 represents the numerical integrity proof as a serialized alphanumeric string.

Asset Signature Creation. The asset signature is generated using the creator's public key. A hash $h(p)$ of the integrity proof of the asset is calculated. This, together with the private key, is the input of the signature function $E()$. The asset signature is a quintuple comprising: i) the hash of the asset (links the asset signature to the asset), ii) creator identity (links the asset signature to the photographer), iii) the signature (a verifiable item), iv) the creator's public key (used for verification), and v) the hash of the integrity proof, i.e., $s = (h(a), id, E_{e_{id}}(h(p)), d_{id}, h(p))$.

Asset Registration. The asset, i.e., image, integrity proof, and asset signature, are submitted for registration and preservation in distributed storage (Figure 2) and blockchain on-chain provenance records (Figure 3).

Note that the public key used to produce the asset signature can be part of a self-signed certificate or an X.509 certificate. The Capture infrastructure uses a self-signed certificate model. While it has the advantage of being simple, it cannot verify the identity of the owner of a public key. In contrast, the Photrace blockchain can do this verification using X.509 certificates. Every camera is configured with an X.509 certificate.

The storage model of Photrace consists of storing an image and its asset signature together in the same file. Any subsequent use of the image, e.g., by a photo editing software, augments the original file (image and asset signature) with a new asset signature.

In contrast to typical blockchain systems, such as those used for cryptocurrencies, the Capture blockchain implementation distinguishes itself by focusing on the secure and collaborative capture of provenance for digital images. In contrast, other blockchain systems often focus on specific features, such as immutability, decentralization, enhanced security through encryption and consensus mechanisms, as well as public or consortium participation models. Public blockchains offer complete transparency and decentralization, while hybrid or consortium blockchains strike a balance between selective openness, efficiency, and control among trusted participants. The Capture blockchain's emphasis lies in cross-chain provenance and forensic requirements rather than just transactional integrity or energy-efficient consensus models. Unlike standard blockchains, which primarily offer decentralized ledgers for transactions, the Capture blockchain emphasizes secure extraction, reliability, and synchronization. It is more specialized than general-purpose blockchains. It is a specialized system within the broader blockchain environment, addressing multi-chain collaboration, security, and provenance

tracking. This distinguishes it from classical public, hybrid, or consortium blockchains that prioritize decentralized transaction validation and ledger maintenance.

V. Security Review

We review and discuss C2PA security issues. Note that the C2PA website lists threat models against C2PA along with vulnerability and harm taxonomies [42, 43]. Within a CAS context, the review of these vulnerabilities extends beyond cybersecurity assessment and speaks to how computational provenance frameworks distribute trust and authority across technical, institutional and social actors.

At the 2024 International Press Telecommunications Council (IPTC) Photo Metadata Conference, security researcher Krawetz highlighted several threats to C2PA, including unauthorized metadata alteration, as well as commenting on numerous other real and potential concerns and exploits [44, 45].

One such concern relates to provenance. All data (and metadata) of an image is stored in the file itself. It is therefore accessible to anyone who has access to the file. This approach may, in certain respects, be at odds with a security principle that aims to minimize the attack surface [46]. As Krawetz stated [44], it is therefore trivial to edit most metadata fields because an attacker can reliably get a service provider (e.g., Adobe or Microsoft) to sign most arbitrary files because they do not validate the metadata itself. Doing so would require activities and resources such as Open-Source Intelligence (OSINT) research, subject matter expertise, digital forensics, and elaborate trust chains. Instead, many service providers currently put a C2PA signature over any existing metadata.

A second such concern relates to watermarking. The watermarking method must be disclosed for a watermark to be validated. This also makes the method known to attackers. This is another potential source of vulnerability, as described by Harris and Norden in the case of Meta's AI watermarking scheme, which is built on C2PA [47].

A third such concern relates to fingerprinting. This includes visually similar picture search engines, as well as the mapping of complex file formats and camera artifacts back to the device or application that generated them. Essentially, tools and techniques exist for mapping the data used for file fingerprinting from one file to another file. For example, [48] includes a demonstration of this for C2PA. These vulnerabilities underscore the dual nature of computational provenance systems, which create new forms of evidence but also new surfaces of attack.

Ultimately, security relies on some form of trust mapping, which entities are trusted, and which are not? In the case of C2PA, the content and the metadata are trusted, either explicitly or implicitly. This means that the produced digital signature validates the integrity of the content and metadata only in the sense that it is the same content and metadata from which the signature was generated. However, this does not validate the content and metadata per se, because *trusted* does not necessarily imply *trustworthy*. Furthermore, actors can conduct trust undermining or even malicious activities with little cost or risk, suggesting that such abuses may remain commonplace.

Further exacerbating the issue of trust, C2PA extends the number and types of trusted entities, including implicitly via assumptions, reliance on honesty, and opt-in (read: trivially bypassable) controls, as illustrated by Tables I and II, adapted from [44]. This is a violation of the minimal trust-based security principle [46]. Note that wherever the words 'assume,' 'trust,' 'honesty,' and/or 'not required' appear, there will be a potentially exploitable vulnerability.

TABLE III. Traditional Metadata Analysis

Content	**Assume** unaltered or acceptable alterations, and **assume** the content is not misrepresented.
Metadata	**Trust** that the metadata accurately reflects the content. This depends on the **reliability** of the person and established procedures for capturing metadata.

TABLE IV. C2PA adds

C2PA Metadata	**Trust** that it accurately reflects the content. This depends on the **reliability** of the person and/or non-human actor capturing the metadata *and* **assumes** that any new metadata is accurate.
Certificate	**Trust** that the certificate is issued to authorized source and **assumes** that the certificate is used properly.
Signer	**Trust** that signers validated the metadata and content. However, validation is **not required** by the C2PA specifications. Also, it requires **trusting** that any new signers have not altered any previous claims.
Validation	**Trust** that the tools perform proper validation, **trust** that the signature covers entire file (which is **not required**), and **trust** that the "tamper evident" system actually detects tampering.
Peer Pressure	**Trust** that thousands of reviewers *actually* review it.

VI. Conclusions

The verifiability of data provenance and the persistence of metadata that establishes and protects the authenticity of digital images (including AI-generated content) is paramount in several sectors of activity. This study examined content credentials as a cryptographic provenance framework capable of supporting the authenticity of digital images. We provide a critical summary of a blockchain-based infrastructure that involves a digital asset model, digital asset storage, a storage model, and an asset creation procedure. It has several promising features and topics for further investigation. We review the current security of C2PA, highlighting several potential issues. They include the risk of alteration of metadata, watermarking and fingerprinting vulnerability, and trust mapping challenges. For archival science, the capacity of C2PA to record multiple,

sequential manifests offers a model for documenting change over time, while maintaining authenticity. This transdisciplinary analysis contributes to the ongoing development of CAS by demonstrating the value of integrating archival diplomatics, cryptographic provenance and cybersecurity methods to better understand how the C2PA pipeline can inform and challenge existing recordkeeping theory and practices. Future research will need to explore case studies of individual and organizational implementations of content credentials in creation, use and preservation pipelines.

ACKNOWLEDGMENT

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences and Humanities Research Council of Canada (SSHRC), and San José State University (SJSU).

REFERENCES

[1] "Content Authenticity Initiative." Oct. 2025. [Online]. Available: https://contentauthenticity.org (accessed Oct. 2, 2025).

[2] *C2PA Specification*, version 2.2. Coalition for Content Provenance and Authenticity (C2PA), 2024. [Online]. Available: https://spec.c2pa.org/specifications/specifications/2.2/index.html. (accessed: Nov. 5, 2025).

[3] J. Bushey, "AI-Generated Images as an Emergent Record Format," IEEE International Conference on Big Data (BigData), Sorrento, Italy, pp. 2020-2031, 2023. doi: 10.1109/BigData59044.2023.10386946.

[4] J. Bushey, "Digital images: Trustworthiness at the dawn of AI," in *Archival Science in Interdisciplinary Theory and Practice*, C. Rogers and A. Wieland, Eds. Washington, DC: Rowman & Littlefield, 2024.

[5] N. Rivard, H. Hamouda, V. Lemieux, T. P. Lauriault, and M. Barbeau, "Towards a Prototype to Leverage Archival Diplomatics to Develop a Framework to Detect and Prevent Fake Videos," *2024 IEEE Digital Platforms and Societal Harms (DPSH)*, Washington, DC, USA, pp. 1-8, , 2024. doi: 10.1109/DPSH60098.2024.10775182.

[6] G. Rolan, G. Humphries, L. Jeffrey, E. Samaras, T. Antsoupova, and K. Stuart, More human than human? Artificial intelligence in the archive. Archives and Manuscripts, 47(2), 179–203, 2018. https://doi.org/10.1080/01576895.2018.1502088

[7] J. Bunn, Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI). Records Management Journal, 30(2), 143-153, 2020.

[8] G. Colavizza, T. Blanke, C. Jeurgens, and J. Noordegraaf, Archives and AI: An overview of current debates and future perspectives. ACM Journal on Computing and Cultural Heritage (JOCCH), 2021, 15(1), 1-15.

[9] J. Proctor and R. Marciano, "An AI-Assisted Framework for Rapid Conversion of Descriptive Photo Metadata into Linked Data," presented at the IEEE International Conference on Big Data, Institute of Electrical and Electronics Engineers, Dec. 2021. doi: 10.1109/BigData52589.2021.9671715.

[10] J. E. Davet, B. Hamidzadeh, P. C. Franks, and J. Bunn, "Tracking the Functions of AI as Paradata & Pursuing Archival Accountability," Archiving Conference, vol. 19, pp. 83–88, Jun. 2022, doi: 10.2352/issn.2168-3204.2022.19.1.17.

[11] L. Jaillant and A. Caputo, "Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data," AI & Soc, vol. 37, no. 3, pp. 823–835, Sep. 2022, doi: 10.1007/s00146-021-01367-x.

[12] L. Jaillant and A. Rees, "Applying AI to digital archives: trust, collaboration and shared professional ethics," Digital Scholarship in the Humanities, vol. 38, no. 2, pp. 571–585, Jun. 2023, doi: 10.1093/llc/fqac073.

[13] Innovation, Science and Economic Development Canada, *Consultation on Copyright in the Age of Generative Artificial Intelligence*. Ottawa, ON: Government of Canada, 2023. [Online]. Available: https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-paper-consultation-copyright-age-generative-artificial-intelligence. (accessed: Nov. 5, 2025).

[14] United States Copyright Office, "Copyright registration guidance: Works containing material generated by artificial intelligence," *Federal Register*, vol. 88, no. 51, pp. 16190–16194, Mar. 16, 2023. [Online]. Available: https://www.govinfo.gov/content/pkg/FR-2023-03-16/pdf/2023-05321.pdf. (accessed: Nov. 5, 2025).

[15] *Creator Guidelines: Making and Maintaining Digital Materials: Guidelines for Individuals.* Vancouver, BC: InterPARES 2 Project, 2008. [Online]. Available: http://www.interpares.org/public_documents/ip2(pub)creator_guidelines_booklet.pdf. (accessed: Nov. 5, 2025).

[16] *Preserver Guidelines: Preserving Digital Records – Guidelines for Organizations.* Vancouver, BC: InterPARES 2 Project, 2008. [Online]. Available: http://www.interpares.org/public_documents/ip2(pub)preserver_guidelines_booklet.pdf. (accessed: Nov. 5, 2025).

[17] J. E. Bushey, "Born digital images as reliable and authentic records," University of British Columbia, 2005. doi: 10.14288/1.0092057.

[18] J. Bushey, "The archival trustworthiness of digital photographs in social media platforms.," University of British Columbia, 2016. doi: 10.14288/1.0300440.

[19] D. Cooke, A. Edwards, A. Day, D. Nair, S. Barkoff, and K. Kelly, *Crossing the Deepfake Rubicon: The Maturing Synthetic Media Threat Landscape.* Washington, DC: Center for Strategic and International Studies (CSIS), 2024. [Online]. Available: http://www.jstor.org/stable/resrep64562 (accessed Nov. 5, 2025).

[20] D. Haden, "Now you see it... now you don't: Detecting AI use via image watermarking," *Information Today*, vol. 40, no. 9, pp. 31–32, 2023.

[21] A. Ghoshal, "Meta to label AI-generated images from Google, OpenAI and Adobe," *Computerworld (Online Only)*, p. 1, 2024. [Online]. Available: *URL*. (accessed: Nov. 5, 2025).

[22] L. Zhu, Y. Lai, W. Mou, H. Zhang, A. Lin, C. Qi, T. Yang, L. Xu, J. Zhang, and P. Luo, "ChatGPT's ability to generate realistic experimental images poses a new challenge to academic integrity," *Journal of Hematology & Oncology*, vol. 17, no. 1, pp. 1–3, 2024. doi: 10.1186/s13045-024-01543-8

[23] H. Luo, L. Li, and J. Li, "Digital watermarking technology for AI-generated images: A survey," *Mathematics*, vol. 13, no. 4, p. 651, 2025. doi: 10.3390/math13040651

[24] T. C. Helmus, *Artificial Intelligence, Deepfakes, and Disinformation: A Primer.* Santa Monica, CA: RAND Corporation, 2022. [Online]. Available: http://www.jstor.org/stable/resrep42027. (accessed: Nov. 5, 2025).

[25] C. Baquero, "All Photos are Fake Until Proven Real," *Commun. ACM* (Blog@CACM), Nov. 1 2023. [Online]. Available: https://cacm.acm.org/blogcacm/all-photos-are-fake-until-proven-real/. (accessed: Nov. 5 2025).

[26] N. Schick, "Faking it: Navigating the new era of generative AI may be the most critical challenge to democracy yet," *RSA Journal*, vol. 169, no. 2(5593), pp. 40–43, 2023. [Online]. Available: https://www.jstor.org/stable/48737215. [Accessed: Nov. 5, 2025].

[27] E. Busch and J. Ware, "Detecting the deepfake – Current mitigation strategies, proposals, & challenges," in *The Weaponisation of Deepfakes: Digital Deception by the Far-Right*, pp. 8–10. The Hague, Netherlands: International Centre for Counter-Terrorism, 2023. [Online]. Available: http://www.jstor.org/stable/resrep55429.9. (accessed: Nov. 5, 2025).

[28] T. J. Thomson, R. J. Thomas, and P. Matich, "Generative visual AI in news organizations: Challenges, opportunities, perceptions, and policies," *Digital Journalism*, pp. 1–22, 2024. doi: 10.1080/21670811.2024.2331769

[29] P. Matich, T. J. Thomson, and R. J. Thomas, "Old threats, new name? Generative AI and visual journalism," *Journalism Practice*, pp. 1–20, 2025. doi: 10.1080/17512786.2025.2451677

[30] The Linux Foundation Projects, C2PA, Advancing digital content transparency and authenticity, [Online] Available: https://c2pa.org (accessed Oct. 2, 2025).

[31] Starling Lab for Data Integrity, [Online] Available: https://www.starlinglab.org (accessed Oct. 2, 2025).

[32] Starling Lab for Data Integrity, “Documenting Stockton’s Homeless”, [Online] Available: https://www.starlinglab.org/documenting-stocktons-homeless (accessed Oct. 2, 2025).

[33] Starling Lab for Data Integrity, “78 Days – Creating a Photographic Archive of Trust”, [Online] Available: https://www.starlinglab.org/78days/ (accessed Oct. 2, 2025).

[34] R. Kumar, R. Tripathi, N. Marchang, G. Srivastava, T.R. Gadekallu, and N.N » Xiong, A secured distributed detection system based on IPFS and blockchain for industrial image and video data security, Journal of Parallel and Distributed Computing, 152, 128-143, 2021.

[35] M. P. McBee and C. Wilcox, Blockchain technology: principles and applications in medical imaging, Journal of digital imaging, 33(3), 726-734, 2020.

[36] M.Y. Jabarulla and H.N. Lee, Blockchain-based distributed patient-centric image management system. Applied Sciences, 11(1), 196, 2020.

[37] R. Mehta, N. Kapoor, S. Sourav and R. Shorey, "Decentralised Image Sharing and Copyright Protection using Blockchain and Perceptual Hashes," 2019 11th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, India, 2019, pp. 1-6, doi: 10.1109/COMSNETS.2019.8711440.

[38] P.W. Khan and Y.A. Byun, A Blockchain-Based Secure Image Encryption Scheme for the Industrial Internet of Things. Entropy, 22, 175, 2020, https://doi.org/10.3390/e22020175.

[39] B. Wang, S. Jiawei, W. Wang, and P. Zhao, Image copyright protection based on blockchain and zero-watermark, IEEE Transactions on Network Science and Engineering, 9(4), 2188-2199, 2022.

[40] T. Igarashi, T. Kazuhiko, Y. Kobayashi, H. Kuno, and E. Diehl, "Photrace: A Blockchain-Based Traceability System for Photographs on the Internet," 2021 IEEE International Conference on Blockchain (Blockchain), Melbourne, Australia, 2021, pp. 590-596, doi: 10.1109/Blockchain53845.2021.00089.

[41] Numbers, Numbers Protocol Whitepaper, Access: May 12, 2025, Online: https://ipfs-pin.numbersprotocol.io/ipfs/bafybeifu5srb5jlstk4eepsg5xaqm3dq5ty2tjetzwb6t5nqb4rrubxola

[42] C2PA, “C2PA Security Considerations”. (2025). [Online] Available: https://c2pa.org/specifications/specifications/1.0/security/Security_Considerations.html.

[43] C2PA, “C2PA Harms Modelling”. In: C2PA (2025). [Online] Available: https://c2pa.org/specifications/specifications/1.4/security/Harms_Modelling.html.

[44] Neal Krawetz, “C2PA from the Attacker’s Perspective”. In: Hcker Factor (2024). [Online] Available: https://www.hackerfactor.com/blog/index.php?/archives/1031-C2PA-from-the-Attackers-Perspective.html.

[45] Neal Krawetz, “C2PA from the Attacker’s Perspective”. In: IPTC (2024). [Online] Available: https://iptc.org/download/events/phmdc2024/krawetz-IPTC-PMD-20224.pdf.

[46] P. Van Oorschoot, Computer security and the internet: tools and jewels from malware to Bitcoin. Springer; 2020.

[47] D. Harris and L. Norden. "Meta’s AI Watermarking Plan Is Flimsy, at Best Watermarks are too easy to remove to offer any protection against disinformation". In: IEEE Spectrum (2024). [Online] Available: https://spectrum.ieee.org/meta-ai-watermarks

[48] Adam Zeloof, "Falsified Photos: Fooling Adobe’s Cryptographically-Signed Metadata". In: Hackaday (2023). [Online] Available: https://hackaday.com/2023/11/30/falsified-photos-fooling-adobes-cryptographically-signed-metadata/