До проблеми активних методів детекції та верифікації ШІ-генерованого контенту: метадані провенансу *C2PA*

Ігор Римар
кандидат історичних наук, завідувач відділу дослідження звуко- та відеозапису, лабораторія досліджень у сфері інформаційних технологій, Київський науково-дослідний експертно-криміналістичний центр Міністерства внутрішніх справ України, м. Київ, Україна, ORCID: https://orcid.org/0000-0001-6153-0900, e-mail: ihor.rymar@outlook.com

Ганна Тулайдан
старший судовий експерт, відділ фототехнічних та портретних досліджень, лабораторія досліджень у сфері інформаційних технологій, Київський науково-дослідний експертно-криміналістичний центр Міністерства внутрішніх справ України, м. Київ, Україна, ORCID: https://orcid.org/0009-0003-1192-5981, e-mail: annatulaydan@gmail.com

Докладно розглянуто стандарт C2PA та інструментарій c2patool, виявлено основні обмеження стандарту. Обґрунтовано необхідність гібридного підходу для аналізу ШІ-генерованого контенту та сформульовано рекомендації щодо практики документально орієнтованого обміну цифровими доказами в діяльності слідчих та оперативних співробітників.

Ключові слова: *C2PA; c2patool; штучний інтелект; SynthID.*

On the Issue of Active Detection and Verification Methods for AI-Generated Content: *C2PA* Provenance Metadata

Ihor Rymar, Anna Tulaydan

The paper provides a detailed examination of the C2PA standard and the c2patool utility, identifying the key limitations of the standard. It justifies the necessity of a hybrid approach to analyzing AI-generated content and formulates recommendations for document-oriented digital evidence exchange in the activities of investigators and operational officers.

Keywords: *C2PA, c2patool; artificial intelligence; SynthID.*

Стрімкий розвиток генеративних моделей штучного інтелекту (далі — *ШІ*), зокрема дифузійних мереж, великих мовних моделей і відеогенераторів спричинив якісний зсув у характері цифрових загроз: синтетичний контент набув масштабу, доступності та фотореалістичності, що внеможливлює його візуальну ідентифікацію без застосування спеціальних знань і ставить під сумнів доказову силу цифрових матеріалів у судових провадженнях. Глобальний звіт про ризики Всесвітнього економічного форуму (2024) назвав дезінформацію та маніпуляцію з медіаконтентом найбільшими глобальними ризиками найближчих років [1]. Для судово-експертної спільноти це означає нагальну необхідність формування верифікованих методологічних підходів до виявлення, ідентифікації та оцінки допустимості як доказу ШІ-генерованого контенту в умовах відсутності усталених галузевих стандартів і методик.

Продовжуючи дослідження проблеми детекції медіаконтенту, створеного генеративним ШІ [2], звернімося до активних методів виявлення: до *C2PA*.

C2PA (*Coalition for Content Provenance and Authenticity*) є відкритим технічним стандартом, розробленим у 2021 році консорціумом *Adobe, Microsoft, BBC, Intel, Truepic, Arm* та інших організацій під егідою *Joint Development Foundation*. Поточна версія специфікації (*C2PA 2.2*, травень 2025 р. [3]) визначає формат *маніфесту провенансу*, що вбудовується безпосередньо у файл цифрового контенту (зображення, відео, аудіо, документ). Маніфест провенансу є своєрідним «ланцюжком зберігання доказів» (*chain of custody*) у цифровому форматі та містить основні компоненти, зазначені у табл. 1.

Таблиця 1

Компоненти маніфесту провенансу *C2PA* [4; 3]

Компонент	Визначення
Claim	Центральний підписаний об'єкт, що підтверджує сукупність *assertions* контенту
Assertions	Структуровані записи про дії над контентом (створення, редагування, публікація), використані інструменти, час, геолокацію, ідентифікатор моделі ШІ
Claim Signature	Цифровий підпис *PKI (X.509)*, що прив'язує *claim* до конкретного актора
Ingredient Hashes	Хеші (*SHA-256/SHA-512*) вхідних матеріалів для верифікації цілісності
Hard Bindings	Хеші самого файлу, що унеможливлюють підміну контенту без інвалідації підпису

Маніфест зберігається у форматі *JUMBF (ISO 19566-5)* і вбудовується у відповідний контейнер: *JPEG* (маркер *APP11*), *PNG* (власний *chunk*), *MP4*/*MOV* (*uuid box*), *WebP*, *HEIC*, *PDF* тощо [3]. Стандарт забезпечує *верифіковану* (а не декларативну) прозорість: факт користування ШІ підтверджено криптографічним підписом виробника інструмента, а не самодекларацією автора. Відкрита специфікація, активна екосистема та поступова інтеграція у соціальні мережі (*LinkedIn*, *TikTok*), упровадження підтримки на рівні апаратного забезпечення провідними виробниками (*Sony, Leica, Nikon*) формують реальну індустріальну базу стандарту [5]. Національний інститут стандартів і технологій США (*NIST*) схарактеризував *C2PA* як «оптимальний стандарт» для масштабованої прозорості цифрового контенту [5]. Водночас важливо розуміти, що *C2PA не призначений для автоматичної верифікації достовірності інформації*. Він підтверджує лише походження та ланцюжок змін контенту [6].

Офіційною референтною реалізацією стандарту з відкритим вихідним кодом (*Rust*, ліцензія *Apache 2.0/MIT*), який підтримують *Adobe* і спільнота *C2PA*, є пакет *c2patool*, що забезпечує: читання та відображення маніфесту; створення та підписання нового маніфесту із застосуванням сертифікатів *X.509*; перевірку валідності підпису та цілісності хешів; підтримку широкої номенклатури форматів (*JPEG, PNG, MP4, WebP, HEIC, TIFF* та ін.); *CLI*-інтерфейс та бібліотечний *API*.

Незважаючи на переваги, фундаментальним недоліком є добровільний (*opt-in*) характер стандарту: відсутність маніфесту *не є* доказом автентичності [6]. *C2PA* не вирішує проблеми «першого підпису» (*first-mile problem*): якщо зловмисник підписує синтетичний контент як автентичний, криптографія підтверджує лише факт підпису, а не відповідність дійсності, що є суттєвим обмеженням у контексті судово-експертної діяльності. Окрім того, офіційна документація стандарту фіксує можливість обходу захисних механізмів через підробку провенансних метаданих та імітацію цифрових відбитків [7]. Інфраструктура відкритих ключів (*PKI*) також потребує верифікованих центрів сертифікації, що залишається невирішеним на рівні глобальної довіри. Основною операційною проблемою *C2PA* в реальних умовах обігу контенту є систематичне знищення маніфестів під час передавання через популярні платформи, що відображено у табл. 2 [5; 8—11].

Таблиця 2

Збереження метаданих провенансу під час передавання через популярні платформи

Платформа / месенджер	Перекодування (стандартний режим)	EXIF / XMP	*C2PA*	Режим «документ / файл»	Деталізація збереження *C2PA* (результати експерименту)
WhatsApp	*JPEG* (40–80%); *H.264* (відео); *Opus* (аудіо)	Ні	Ні	Так (зберігає оригінал для зображень, відео й аудіо)	Перетягування (стандартний): .wav → Так; .jpeg/.jpg, .mp4 (Якість: HD/SD) → Ні; Файлом: .wav/.jpg/.png/ .mp4 → Так

Завершення таблиці 2

Платформа / месенджер	Перекодування (стандартний режим)	EXIF / XMP	*C2PA*	Режим «документ / файл»	Деталізація збереження *C2PA* (результати експерименту)
Telegram	*JPEG* (стандартний режим); *H.264/H.265* (відео); *OGG Opus* (аудіо)	Ні	Умовно [1]	Так («без стиснення» / «файлом») — зберігає оригінал і метадані	Без стиснення (стандартний): .wav/.jpg/.png/.mp4 → Так; Файлом: .wav/.jpg/.png/.mp4 → → Так; Швидко (стандартний): .mp4 → Так; .jpeg/.jpg → Ні
Viber	*JPEG* («миттєво»); *H.264 (≤720p)*	Ні	Умовно [2]	Так (зберігає оригінал зображень, відео й аудіо)	Миттєво (стандартний): macOS 26.4.1: .jpg/.png → Так; Android 16: .jpg → Ні; Оригінал: .jpg/.png → Так; Перетягуванням: .wav/.mp4 → Так; Файлом: .wav/.jpg/.png/.mp4 → Так
Facebook	*JPEG/WebP; H.264*	Ні	Ні	Недоступний	Повідомленням/Публікація: .jpg/.jpeg/.mp4/.webp → Ні [3]
Instagram	*JPEG/WebP; H.264*	Ні	Ні	Недоступний	Публікація (yt-dlp [4]/gallery-dl [5])/ Повідомленням: .jpg/.mp4 → Ні
Twitter / X	*JPEG/WebP; H.264*	Ні	Ні	Недоступний	Публікація (yt-dlp): .jpg/.mp4 → Ні
TikTok	*H.264/H.265* (обов'язкове)	Ні	Ні	Недоступний	Публікація: .mp4 → Ні
LinkedIn	Помірне *JPEG*	Ні [6]	Ні [6]	–	Повідомленням (зображенням, файлом): .jpeg/.jpg/.png → Ні; Публікація: macOS 26.4.1: .jpeg → Ні; Windows 11 (Chrome DevTools [7]): .mp4 → Ні
YouTube	*VP9/AV1/H.264*	Умовно	Ні	–	Публікація: .mp4 → Ні

Примітка. «Так» — метадані збережені в усіх перевірених сценаріях; «Умовно» — часткове збереження або залежить від режиму передавання / операційної системи; «Ні» — метадані не збережені в жодному зі стандартних сценаріїв. Результати експерименту (квітень 2026 р.) є попередніми, потребують підтвердження / уточнення / корегування із застосуванням більшого масиву експериментальних даних (фото / відео / аудіо) і платформ / месенджерів.

1 Telegram / *C2PA*: у стандартному режимі («швидко») *C2PA* не зберігається для зображень. Режими «без стиснення» і «файлом» забезпечують збереження *C2PA* для всіх форматів (.wav, .jpg, .png, .mp4) на macOS 26.4.1 та Android 16. У режимі «швидко» відеофайли (.mp4) зберігають *C2PA*, зображення (.jpeg/.jpg) — ні.

2 Viber / *C2PA*: у стандартному режимі «миттєво» *C2PA* зберігається на macOS 26.4.1 (.jpg, .png), але не зберігається на Android 16 (.jpg). Режими «оригінал», «перетягуванням» і «файлом» забезпечують збереження *C2PA* на обох платформах. Режим «файлом» підтримує всі формати (.wav, .jpg, .png, .mp4).

3 У деяких файлах частково збережено тег DigitalSourceType (MediaInfo, v. 26.01), що фіксує «Created using Generative AI». *C2PA*-маніфест відсутній.

4 yt-dlp — багатофункціональний завантажувач аудіо / відео командного рядка з підтримкою тисяч сайтів. Проект є форком youtube-dl, заснованим на нині неактивному youtube-dlc.

5 gallery-dl — кросплатформний інструмент з багатьма параметрами командного рядка й налаштування, а також потужними можливостями іменування файлів.

6 LinkedIn / *C2PA* та EXIF/XMP: експеримент не виявив збереження *C2PA*-маніфесту в жодному із перевірених сценаріїв завантаження (macOS 26.4.1., Android 16 і Windows 11), підтримка Content Credentials стосується лише відображення інформації платформою для завантажених матеріалів, але не гарантує збереження *C2PA* у файлах під час завантаження.

7 Chrome DevTools — набір інструментів веброзробника, убудованих безпосередньо в браузер Google Chrome (та ін. на базі Chromium). Вони дають змогу редагувати сторінки в реальному часі, діагностувати проблеми, налагоджувати HTML, CSS і JavaScript, аналізувати продуктивність сайту й мережевий трафік.

Отже, поведінка платформ є свідомим архітектурним рішенням, спрямованим на забезпечення стандартизації, конфіденційності й оптимізації трафіку [12; 10], а не наслідком компресії даних. За стандартної передачі медіаконтенту через соціальні мережі та месенджери маніфести *C2PA* майже гарантовано знищуються, що внеможливлює верифікацію походження ШІ-генерованого контенту на стороні отримувача. Критичним чинником для відео- (*MP4/MOV*) та аудіофайлів (*MP3, WAV*) є процедура транскодування: навіть за збереження вихідного контейнера, будь-яке перекодування потоку в стрімінгові формати (такі як *Opus, AAC* або *VP9*) призводить до незворотного видалення службових блоків метаданих (*uuid box* у відеофайлах, *ID3v2*, *RIFF-INFO* в аудіофайлах), що узгоджується з попередніми спостереженнями щодо проблематики збереження метаданих у цифрових розслідуваннях [12]. Тож для потреб судової експертизи єдиним надійним методом передачі медіаданих працівниками слідчих та оперативних підрозділів, що забезпечує цілісність метаданих провенансу, залишається режим «документ / файл».

Наявні сьогодні шляхи розв'язання проблеми метаданих провенансу наведено у схемі далі.

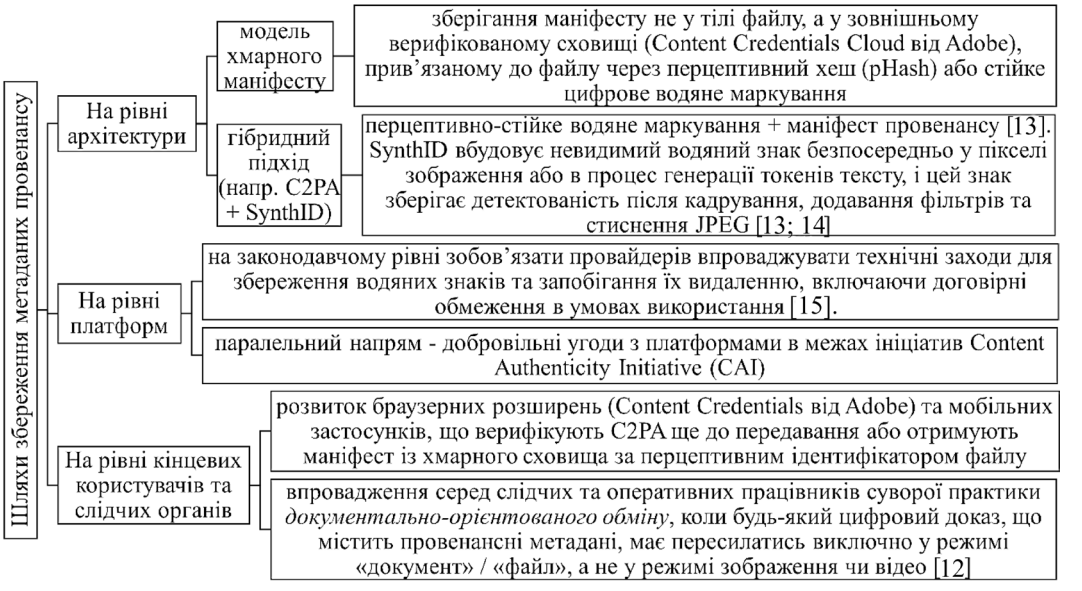

Схема «Шляхи збереження метаданих провенансу»

Отже, незважаючи на те, що метадані провенансу *C2PA* є найбільш технічно зрілими і криптографічно обґрунтованими з-поміж наявних активних механізмів маркування ШІ-генерованого контенту [4; 3] і вирішують проблему *верифікованої атрибуції*, вони мають низку суттєвих обмежень, що стосуються проблем *обов'язкового маркування* [15], крихкості маніфестів — це суттєво обмежує практичну придатність *C2PA* в контексті криміналістичного дослідження [7; 6; 10] — і потребують гібридного підходу (аналіз *C2PA*-маніфесту (за наявності) + верифікація через хмарне сховище *Content Credentials* + аналіз перцептивно-стійкого водяного маркування (наприклад, *SynthID*) + пасивні методи дослідження) для аналізу ШІ-генерованого контенту. Подальший розвиток стандарту *C2PA* покликаний виробити дійсно корисний і потужний інструмент, який у поєднанні з аналітичними методами дасть змогу з великою часткою ймовірності досліджувати й ідентифіковувати контент, генерований ШІ.

Перелік джерел посилання

1. World Economic Forum. Global Risks Report 2024. Geneva : WEF, 2024. URL: https://www.weforum.org/reports/global-risks-report-2024 (дата звернення: 12.04.2026).

2. Римар І., Тулайдан Г. До проблеми методів детекції медіаконтенту, створеного за допомогою генеративного штучного інтелекту, під час проведення судової експертизи. *Актуальні питання експертного забезпечення правосуддя у справах адміністративної юрисдикції: сучасний стан та перспективи розвитку* : зб. мат-лів Міжнар. наук.-практ. конф. (Харків — Київ, 25.02.2026). Харків, 2026. С. 360—363. URL: https://res2.weblium.site/res/64e5d4a1196a4b000eb26345/69ae823c2235cac5e127d5c7 (дата звернення: 12.04.2026).
3. Coalition for Content Provenance and Authenticity (*C2PA*). *C2PA* Technical Specification v2.2. *Joint Development Foundation.* 2025. URL: https://surl.li/isnmvk (дата звернення: 12.04.2026).
4. Coalition for Content Provenance and Authenticity (*C2PA*). *C2PA* Technical Specification v2.1. *Joint Development Foundation.* 2024. URL: https://surl.li/yfhfrt (дата звернення: 12.04.2026).
5. Coalition for Content Provenance and Authenticity (*C2PA*). *Response to NIST AI 100-4: Reducing Risks Posed by Synthetic Content*. 2024. URL: https://surl.li/tlckhd (дата звернення: 12.04.2026).
6. Kaye K., Dixon P. Privacy, Identity and Trust in *C2PA*: A Technical Review and Analysis of the *C2PA* Digital Media Provenance Framework. *World Privacy Forum.* 2025. URL: https://worldprivacyforum.org/posts/privacy-identity-and-trust-in-c2pa/ (дата звернення: 12.04.2026).
7. Coalition for Content Provenance and Authenticity (*C2PA*). Security Considerations. *Joint Development Foundation*. 2024. URL: https://surl.li/dfjlww (дата звернення: 12.04.2026).
8. Arshad M., Ahmad A., Onn C. W., Sam E. A. Investigating methods for forensic analysis of social media data to support criminal investigations. *Frontiers in Computer Science.* 2025. Vol. 7. Art. 1566513. DOI: 10.3389/fcomp.2025.1566513 (дата звернення: 12.04.2026).
9. Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era. *U. S. National Security Agency. Cybersecurity Information Sheet. 2025.* URL: https://surl.li/bjpths (дата звернення: 12.04.2026).
10. Soni N. Forensic Value of Exif Data: An Analytical Evaluation of Metadata Integrity across Image Transfer Methods. *Perspectives in Legal and Forensic Sciences. SCIEPublish.* 2025. Vol. 2. Is. 2. DOI: 10.70322/plfs.2025.10006 (дата звернення: 12.04.2026).
11. The Promise and Risk of Digital Content Provenance. *Center for Democracy and Technology (CDT).* 2025. URL: https://surl.lu/wwjqyn (дата звернення: 12.04.2026).
12. Marin M. Sending Encrypted Photos While Preserving Metadata. *Citizen Evidence Lab.* 2020. URL: https://surl.li/woxkqm (дата звернення: 12.04.2026).
13. Watermarking AI-Generated Text and Video with SynthID. *Google DeepMind.* 2024. URL: https://deepmind.google/blog/watermarking-ai-generated-text-and-video-with-synthid/ (дата звернення: 12.04.2026).
14. Dathathri S., See A., et al. Scalable watermarking for identifying large language model outputs. *Nature.* 2024. Vol. 634. P. 818–823. DOI: https://doi.org/10.1038/s41586-024-08025-4 (дата звернення: 12.04.2026).
15. Code of Practice on Marking and Labelling of AI-generated Content (second draft). *European Commission.* March 2026. URL: https://surl.li/fovlkv (дата звернення: 12.04.2026).