

Overovanie pôvodu výstupov generatívnej umelej inteligencie

Technológie, možnosti a limity watermarkingu

Martin ZÁBORSKÝ, Slavka PITOŇÁKOVÁ

Abstrakt

Rozvoj generatívnej umelej inteligencie prináša nové výzvy pri určovaní pôvodu digitálneho obsahu, keďže synteticky vytvorený obsah je v mnohých prípadoch ťažké odlíšiť od ľudskej tvorby. Článok sa zaoberá watermarkingom AI výstupov a súvisiacimi technológiami digitálnej autenticity, ktoré podporujú overovanie pôvodu a histórie obsahu. Približuje vybrané prístupy k označovaniu a overovaniu obsahu vrátane textového a obrazového watermarkingu, systému SynthID a štandardu C2PA. Tieto prístupy môžu uľahčiť určenie pôvodu obsahu a jeho moderovanie, neposkytujú však jednoznačný dôkaz o jeho pravdivosti. Ich účinnosť môžu znižovať napríklad parafrázovanie, preklad, orezanie, kompresia, odstránenie metadát, nejednotné používanie technológií a chyby pri detekcii. Výskum používateľských reakcií naznačuje, že označovanie AI obsahu môže zvýšiť transparentnosť a pomôcť používateľom lepšie posúdiť jeho pôvod. Ich účinok však závisí od spôsobu označenia, kontextu použitia a schopnosti používateľov rozlíšiť pôvod obsahu od jeho faktickej správnosti. Ako perspektívne sa javia prístupy kombinujúce watermarking, technológie overovania pôvodu obsahu a profesionálne postupy overovania informácií.

Kľúčové slová

overovanie pôvodu obsahu, generatívna umelá inteligencia, digitálna autenticita, dôveryhodnosť médií, informačná integrita



Abstract

The rapid development of generative artificial intelligence has introduced new challenges in determining the origin of digital content, as synthetically generated content is often difficult to distinguish from human-created content. This article examines the watermarking of AI-generated outputs and related digital authenticity technologies that support the verification of content origin and provenance. It reviews selected approaches to content labeling and verification, including text and image watermarking, the SynthID system, and the C2PA standard. These approaches can facilitate the identification of content origin and support content moderation; however, they do not provide conclusive evidence of the truthfulness of the content itself. Their effectiveness may be reduced by paraphrasing, translation, cropping, compression, metadata removal, inconsistent technology adoption, and detection errors. Research on user responses suggests that AI content labeling can increase transparency and help users better assess content provenance. Its impact, however, depends on the labeling method, the context of use, and users' ability to distinguish content origin from its factual accuracy. Approaches combining watermarking, provenance systems, and professional information verification practices appear particularly promising.

Keywords

content provenance verification, generative artificial intelligence, digital authenticity, media credibility, information integrity



Úvod

Rozvoj a použitie generatívnej umelej inteligencie zásadne znížil náklady na výrobu presvedčivého textu, obrazu, zvuku a videa. Dôsledkom nie je len nárast objemu synteticky vytvoreného obsahu, ale aj oslabovanie tradičných indikátorov, na základe ktorých publikum posudzovalo pôvod a spôsob vzniku digitálneho obsahu (či bol vytvorený fotografovaním, nahrávaním, redakčným spracovaním alebo počítačovou syntézou). Keďže tieto indikátory sú čoraz menej spoľahlivé, autenticitu digitálneho obsahu už nemožno posudzovať iba na základe ľudského vnímania a skúseností. Problém spoľahlivého určovania pôvodu a autenticity digitálneho obsahu preto nemožno redukovať na otázku, či človek dokáže rozpoznať synteticky vytvorený alebo manipulovaný obsah zrakom či sluchom. Týka sa širšej informačnej infraštruktúry, v ktorej sa pôvod, história úprav a zodpovednosť za publikovanie často strácajú pri prenose medzi aplikáciami a platformami (Chandra et al., 2024; UNESCO, 2023).

Výskum vplyvu synteticky vytvoreného digitálneho obsahu ukazuje, že hlavným účinkom nemusí byť priame presvedčenie publika o nepravdivom tvrdení. Rovnako závažná je informačná neistota, ktorá oslabuje dôveru v digitálne informačné zdroje vrátane správ šírených na sociálnych médiách (Vaccari & Chadwick, 2020). Narušenie dôvery v autenticitu digitálneho obsahu však nevytvára len priestor pre úspešné falzifikáty. Existencia ľahko vytvoriteľných falzifikátov zároveň poskytuje aktérom možnosť označiť autentický kompromitujúci materiál za podvrh, čo Chesney a Citron (2019) pomenovali ako *dividendu klamára*. Overovanie pôvodu teda chráni nielen pred falošným obsahom, ale aj pred účelovým spochybňovaním pravých záznamov.

Potreba spoľahlivejšie dokumentovať pôvod obsahu preto viedla k vývoju technických mechanizmov označovania a overovania digitálnych výstupov. Watermarking, teda vkladanie identifikačného signálu do vygenerovaného obsahu, patrí medzi najdiskutovanejšie odpovede na tento problém. Často sa používa ako zastrešujúci pojem pre rozdielne techniky: viditeľné logá, neviditeľné signály, metadáta a kryptografické záznamy. Takéto zjednodušenie vedie k prehnaným očakávaniam. Watermark môže pomôcť preukázať, že konkrétny systém vložil do



výstupu určitý signál, ale sám o sebe nepotvrďuje, že zobrazená udalosť je pravdivá, že autor mal súhlas dotknutých osôb alebo že obsah nebol vytrhnutý z kontextu (C2PA, 2026).

Zámerom článku je predstaviť možnosti súčasných techník watermarkingu v širšom kontexte technológií digitálnej autenticity a posúdiť jeho význam pre mediálnu dôveryhodnosť. Ťažisko spočíva v otázke, aký druh tvrdenia jednotlivé systémy umožňujú urobiť, kto ho môže overiť a čo sa stane, keď sa obsah upraví, preloží, skopíruje alebo zámerne zbaví identifikačných znakov.

Watermarking ako súčasť digitálnej autenticity

Digitálny watermark je informácia vložená do nosného obsahu tak, aby sa dala neskôr zistiť špecializovaným detektorom. Môže byť viditeľný alebo skrytý, verejne overiteľný alebo závislý od tajného kľúča a môže obsahovať jednoduchú informáciu alebo podrobnejšie údaje o modeli, službe či používateľovi. Pri syntetickom obsahu sa značka môže pridať po vygenerovaní alebo priamo počas generatívneho procesu. Kvalitná schéma musí súčasne zachovať vnímanú kvalitu, dosiahnuť nízku mieru falošne pozitívnych a falošne negatívnych výsledkov, prežiť bežné úpravy a odolávať zámernému odstráneniu (Chandra et al., 2024).

Tieto požiadavky si navzájom konkurujú. Silnejší signál sa spravidla ľahšie deteguje, môže však viac zasiahnuť do štýlu textu alebo vizuálnej kvality. Slabý signál je menej rušivý, no stráca sa v krátkom texte, pri kompresii alebo úprave. Vyššia kapacita umožní zakódovať identifikátor zdroja, ale zväčšuje priestor pre chyby. Preto sa úspech watermarkingu nedá hodnotiť jedným percentom presnosti; potrebný je opis konkrétneho modelu, typu obsahu, dĺžky vzorky a povolených transformácií (Liu et al., 2024; Pang et al., 2024).

Pri posudzovaní digitálneho obsahu je užitočné rozlišovať tri vrstvy dôveryhodnosti:

1. integrita súboru, teda otázka, či bol digitálny objekt po podpise zmenený,
2. pôvod obsahu (proveniencia), teda dokumentovaná história jeho vzniku a následných úprav vrátane informácií o použitých nástrojoch, zdrojoch a autoroch jednotlivých zásahov,



3. dôveryhodnosť samotného tvrdenia, teda či zodpovedá skutočnosti a či je uvedené v správnom kontexte.

Kryptografia a watermarking môžu posilniť prvé dve vrstvy, ale tretiu nemožno automatizovať bez práce novinárov, znalcov a dôveryhodných inštitúcií (C2PA, 2026).

Prístupy k označovaniu AI výstupov

Označovanie AI výstupov možno realizovať viacerými spôsobmi, ktoré sa líšia tým, či informáciu o pôvode priamo vkladajú do obsahu, pripájajú ju vo forme metadát alebo sa ju snažia spätne zistiť pomocou detekcie. Rozdiely medzi najčastejšie používanými mechanizmami označovania a overovania obsahu naznačuje tabuľka 1.

Tabuľka 1: Mechanizmy označovania a overovania digitálneho obsahu

Mechanizmus	Typ podporovaného tvrdenia	Obmedzenie
Textový watermark	Miera pravdepodobnosti pôvodu v konkrétnej schéme	Nepotvrďuje autorstvo ani pravdivosť
Obrazový alebo audiovizuálny watermark	Pôvod v systéme používajúcom danú značku	Nepotvrďuje pravdivosť zobrazeného obsahu
SynthID	Pôvod v systéme kompatibilnom so SynthID	Absencia značky nevylučuje AI pôvod
C2PA / Content Credentials	Deklarovaný pôvod a história úprav	Nepotvrďuje pravdivosť deklarácií
Fingerprinting	Zhoda s referenčným obsahom	Neposkytuje provenienciu
Pasívna detekcia AI obsahu	Miera pravdepodobnosti syntetického pôvodu	Poskytuje len pravdepodobnostný výsledok

Textový watermarking

Text je pre skryté označovanie náročný, pretože na rozdiel od obrazu ponúka len obmedzené možnosti vloženia dodatočnej informácie bez ovplyvnenia významu. Keďže význam textu závisí od konkrétneho výberu slov, aj drobné úpravy môžu ovplyvniť jeho interpretáciu alebo faktickú presnosť. Najznámejšie generatívne schémy preto zasahujú do výberu tokenov počas tvorby textu. Kirchenbauer a kolegovia (2023) navrhli mechanizmus, ktorý pre každý krok pseudonáhodne rozdelí slovník modelu a mierne zvýhodní takzvané zelené tokeny. Detektor potom zisťuje, či ich je v texte štatisticky viac, než by sa očakávalo náhodou.



Výhodou tohto princípu je, že čitateľ nevidí osobitnú značku a detektor nemusí mať prístup k parametrom pôvodného jazykového modelu. Pri dostatočne dlhom texte môže poskytovať interpretovateľné skóre. Zároveň však nejde o dôkaz autorstva v bežnom zmysle. Detekcia vyhodnocuje pravdepodobnosť prítomnosti konkrétneho štatistického vzoru. Krátke odpovede, deterministické formulácie, kód, tabuľkové údaje alebo text s obmedzenou slovnou zásobou ponúkajú menej priestoru na bezpečné vloženie signálu (Liu et al., 2024).

Na tento princíp nadväzujú novšie implementácie. SynthID Text od Google DeepMind využíva pokročilejšie postupy výberu tokenov, pri ktorých je voľba generovaných slov ovplyvnená tajným watermarkingovým kľúčom. Metóda nevyžaduje pretrénovanie modelu a je použiteľná aj v produkčných systémoch. Autori pri takmer 20 miliónoch odpovedí Gemini nezistili prakticky významné zhoršenie používateľskej skúsenosti, čo naznačuje dobrú škálovateľnosť riešenia (Dathathri et al., 2024).

Produkčná realizovateľnosť však nerieši problém univerzálneho pokrytia. Watermark možno nájsť iba vo výstupe systému, ktorý ho vložil, a detektor musí poznať príslušnú schému alebo kľúč. Text možno navyše parafrázovať, preložiť, skrátiť alebo kombinovať s ľudskými zásahmi. Rastogi a Pruthi (2024) ukázali, že schémy deklarujúce odolnosť proti parafrázovaniu možno pri obmedzenom počte vzoriek analyzovať a následne účinnejšie obchádzať. Teoretická práca Zhang a kolektívu (2024) zase dokazuje, že za všeobecných predpokladov nemožno garantovať silný watermark, ktorý by informovaný útočník neodstránil pri zachovaní prijateľnej kvality.

Textový watermarking je užitočný najmä vtedy, keď ho poskytovateľ AI systému používa a umožňuje jeho overovanie. Môže tiež pomôcť pri kontrole veľkého množstva textov alebo slúžiť ako upozornenie na obsah, ktorý je potrebné ďalej preveriť. Nie je však vhodné používať ho ako jediný podklad na disciplinárne, pracovnoprávne alebo právne rozhodnutie. Pravdepodobnostný výsledok sa navyše môže vzťahovať len na konkrétnu metódu, nie na všetky texty vytvorené umelou inteligenciou.



Obrazový a audiovizuálny watermarking

Obraz poskytuje viac priestoru na vloženie skrytého signálu. Postprodukčné metódy menia pixely alebo frekvenčné zložky tak, aby rozdiel nebol pre človeka viditeľný, no dal sa strojovo prečítať. Ich prednosťou je použiteľnosť na výstup ľubovoľného generátora. Slabinou je oddelenie od samotného modelu: ak je označovanie iba posledným krokom exportu, používateľ s kontrolou nad pracovným postupom ho môže vypnúť alebo obísť (Fernandez et al., 2023).

Novšie metódy obrazového watermarkingu, napr. Stable Signature (Fernandez et al., 2023) alebo Tree Rings (Wen et al., 2023), integrujú identifikačný signál priamo do procesu generovania obrazu, čím zvyšujú jeho odolnosť voči bežným úpravám. Ani tieto prístupy však nepredstavujú nezničiteľný dôkaz pôvodu, keďže kompresia, snímka obrazovky, prekreslenie, zmena formátu, koláž alebo opätovné vygenerovanie obsahu môžu detekciu oslabiť. Zhao a kolegovia (2024) ukázali útok, pri ktorom sa obraz najprv zašumí a potom rekonštruje generatívnym modelom. Pri skúmaných pixelových watermarkoch útok výrazne znížil detekciu pri zachovaní vizuálne podobného obsahu. Výsledok neznamená, že všetky watermarky sú bezcenné; znamená, že odolnosť musí byť testovaná proti aktuálnym generatívnym nástrojom, nielen proti kompresii alebo zmene jasu.

Pri videu a zvuku sa pridáva časový rozmer. Značka musí prežiť vystrihnutie úseku, zmenu snímkovú frekvenciu, prevzorkovanie, miešanie zvukových stôp a platformovú kompresiu. Súčasne nesmie zhoršiť zrozumiteľnosť reči ani kvalitu obrazu. Praktické systémy preto rozkladajú signál naprieč segmentmi alebo rámcami a detekciu vyhodnocujú kumulatívne. Takýto prístup zvyšuje šancu, že zostane čitateľná aspoň časť signálu, no opäť vyžaduje definovať minimálnu dĺžku a podmienky overenia (Google DeepMind, 2026).

Multimodálny systém SynthID

Praktickým príkladom snahy o jednotné označovanie naprieč viacerými modalitami je systém SynthID. SynthID je súbor príbuzných metód pre obraz, video, zvuk a text. V obraze a videu sa neviditeľný signál vkladá pri vytváraní obsahu; pri zvuku má zostať nepočuteľný a prežiť



bežné úpravy; pri texte sa menia pravdepodobnosti výberu tokenov. Spoločným menovateľom je spojenie vloženia a detekcie s ekosystémom poskytovateľa (Google DeepMind, 2026).

Pre používateľa je dôležité správne interpretovať negatívny výsledok. Nenájdenie SynthID neznamena, že obsah nevytvorila AI. Môže pochádzať z iného generátora, zo staršej verzie služby, z nepodporovaného formátu alebo mohol byť signál degradovaný. Pozitívny výsledok zas dokladá detekciu kompatibilnej značky, nie faktickú pravdivosť. Tento význam výsledkov by mal byť súčasťou každého používateľského rozhrania.

Čoraz viac sa preto uplatňuje kombinovanie viacerých spôsobov označovania a overovania obsahu. OpenAI oznámilo doplnenie C2PA proveniencie o watermark SynthID pri obrázkoch generovaných v ChatGPT, Codex a API. Dôvodom je, že metadáta poskytujú podrobnejšie informácie o pôvode a úpravách obsahu, no môžu sa stratiť; skrytý signál môže prežiť niektoré transformácie, ale poskytuje menej informácií (OpenAI, 2026). Kombinácia týchto prístupov môže zvýšiť ich účinnosť, zároveň však poukazuje na potrebu nezávislého testovania a verejne zrozumiteľných pravidiel detekcie. Súčasné snahy o overovanie pôvodu obsahu sa preto neobmedzujú len na vkladanie značiek do samotného obsahu, ale zahŕňajú aj systémy dokumentovania jeho proveniencie.

Mechanizmy C2PA a Content Credentials

Jedným z najvýznamnejších systémov proveniencie, s ktorými sa watermarking môže kombinovať, je štandard C2PA. C2PA nie je klasický watermark. Ide o otvorený technický štandard, v ktorom sa k digitálnemu aktívu pripája kryptograficky podpísaný manifest. Ten môže obsahovať tvrdenia o zariadení alebo softvéri, čase vytvorenia, použitých úpravách, vložených zdrojoch a zapojení generatívnej AI. Hash spája záznam s konkrétnou podobou súboru a digitálny podpis umožňuje overiť, kto tvrdenie vydal a či sa po podpise zmenilo. Platný podpis v systéme C2PA umožňuje overiť, že deklarácia o pôvode bola vydaná konkrétnym subjektom a od podpisu nebola zmenená. Nepotvrďuje však správnosť samotnej deklarácie (C2PA, 2026).



Content Credentials poskytujú odpoveď na otázku, odkiaľ súbor pochádza a akú deklarovanú históriu nesie. Neodpovedajú automaticky na otázku, či zachytená udalosť je pravdivá. Dôveryhodnosť záznamu závisí od identity podpisujúceho subjektu, bezpečnosti jeho nástrojov a úplnosti reťazca. C2PA výslovne rozlišuje platnú provenienciu od hodnotenia pravdivosti. Autentický záznam môže dokumentovať inscenovanú scénu a syntetický obraz môže byť korektnou ilustráciou (C2PA, 2026).

Hlavnou praktickou slabinou je strata manifestu pri exporte, kopírovaní alebo nahrávaní na platformu, ktorá metadáta odstráni. Samotná absencia Content Credentials preto nesmie byť chápaná ako dôkaz manipulácie. Iniciatíva Content Authenticity Initiative presadzuje takzvané trvácne poverenia, ktoré kombinujú vložené metadáta s neviditeľným watermarkom a obsahovým fingerprintom. Ak sa manifest oddelí od súboru, skrytý signál alebo podobnostný odtlačok môže sprostredkovať vyhľadanie jeho kópie v externom úložisku (Content Authenticity Initiative, 2024).

Význam C2PA presahuje rámec označovania obsahu generovaného umelou inteligenciou. Rovnaká infraštruktúra môže overiť aj pôvod fotografií z kompatibilného fotoaparátu, následné úpravy v editore a identitu vydavateľa. Vhodnejšie je preto hovoriť o pozitívnej proveniencii dôveryhodných médií než o univerzálnom stigmatizovaní syntetických výstupov. Takéto použitie pomáha chrániť autentický materiál pred dividendou klamára (Chesney & Citron, 2019).

Viditeľné označenie a detekcia AI obsahu

Skrytý strojový signál sa stáva účinným až vtedy, keď ho platforma alebo overovací nástroj preloží do zrozumiteľnej informácie. Viditeľné označenie môže byť súčasťou obrazu, prehrávača alebo sprievodného rozhrania. Má okamžitý komunikačný účinok, ale možno ho orezať, napodobniť alebo oddeliť od obsahu. Preto by mal štítok uvádzať aj zdroj a typ dôkazu, napríklad platný podpis vydavateľa, watermark konkrétneho systému alebo iba používateľské priznanie.



Nie všetky prístupy k identifikácii AI obsahu sú založené na vkladaní značky do obsahu. Fingerprinting je technika, ktorá vytvára jedinečný digitálny „odtlačok“ obsahu na účely jeho rozpoznania alebo porovnania s už známymi objektmi. Na rozdiel od watermarkingu nevkladá informáciu priamo do obsahu. Z obsahu vypočíta odtlačok, ktorý sa následne porovnáva s databázou známych objektov. Percepčný hash môže nájsť aj upravené verzie podobného obrazu, potrebuje však referenčný register a sám nevysvetľuje históriu úprav. Pasívne detektory naopak hľadajú štatistické znaky syntézy bez toho, aby generátor vopred vložil značku. Majú širší záber, ale ich výkon sa mení s novými modelmi a transformáciami. Americký Národný inštitút pre štandardy a technológie (NIST) preto odporúča vnímať tieto metódy ako doplnkové signály s merateľnou mierou neistoty, nie ako jednoznačný dôkaz o pôvode obsahu (Chandra et al., 2024).

Možnosti a limity overovania pôvodu obsahu

Najväčšou výhodou watermarkingu je škálovateľnosť. Platforma môže automaticky kontrolovať veľké množstvo položiek, pridať kontextové označenie alebo odovzdať prípad na ľudské posúdenie. Watermark môže pomôcť určiť, z akej služby alebo zdroja obsah pochádza. Pre médiá môže takýto signál uľahčiť prvotnú kontrolu materiálu, nemôže však nahradiť redakčné overovanie.

Druhou výhodou je možnosť overiť pôvod obsahu s väčšou mierou istoty. Klasifikátor často iba odhaduje podobnosť so syntetickým obsahom, kým správne overený watermark alebo podpis môže odkazovať na konkrétneho poskytovateľa. Táto vlastnosť je cenná pri incidentoch, keď treba rozlíšiť pôvodný súbor od množstva kópií. Je však účinná iba pri dostatočnom pokrytí generátorov a kompatibilite detektorov.

Medzi hlavné technické limity patria:

1. bežné transformácie oslabujú signál a útočník môže použiť parafrázovanie alebo inú úpravu obsahu,
2. každý detektor robí chyby a jeho prah musí zodpovedať dôsledkom rozhodnutia,



3. otvorené modely možno upraviť tak, aby watermark nevkladali, kým uzavreté služby môžu ponechať detekciu iba vo vlastnom ekosystéme,
4. neexistuje jedna značka pre všetky modalitty a všetkých poskytovateľov.

Teoretické a empirické práce preto hovoria o kompromisoch, nie o absolútnej odolnosti (Pang et al., 2024; Zhang et al., 2024; Zhao et al., 2024). Technické obmedzenia však nepredstavujú jedínú výzvu. Účinnosť systémov overovania pôvodu závisí aj od spôsobu ich správy, prijatia a dôvery v inštitúcie, ktoré ich prevádzkujú.

Inštitucionálne limity sú rovnako dôležité. Systémy technickej proveniencie zároveň presúvajú časť otázky dôveryhodnosti od samotného obsahu k subjektom, ktoré vydávajú a overujú tvrdenia o jeho pôvode. Podrobné identifikátory môžu podporiť zodpovednosť, ale aj sledovanie tvorcov alebo používateľov. Systém preto predpokladá oddelenie verejne dostupných informácií od forenzných údajov a jasne definované pravidlá prístupu (C2PA, 2026).

Regulačný a inštitucionálny rámec

Technické možnosti označovania obsahu sa v posledných rokoch premietli aj do regulačných požiadaviek. Európsky AI Act zakotvil označovanie syntetického obsahu do právneho rámca. Článok 50 ods. 2 vyžaduje, aby poskytovatelia relevantných systémov zabezpečili strojovo čitateľné a detekovateľné označenie výstupov ako umelo generovaných alebo manipulovaných. Technické riešenia majú byť podľa možností účinné, odolné a spoľahlivé. Odsek 4 ukladá používateľom systémov povinnosť zverejniť umelý pôvod synteticky vytvoreného obrazu, zvuku alebo videa a za určených okolností aj textu publikovaného na informovanie verejnosti o veciach verejného záujmu (Európsky parlament a Rada Európskej únie, 2024).

Tieto povinnosti sa uplatňujú od 2. augusta 2026. Európska komisia v roku 2026 dokončila dobrovoľný Kódex postupov k transparentnosti AI generovaného obsahu, ktorý konkretizuje pravidlá pre poskytovateľov a nasadzujúce subjekty. Kódex zdôrazňuje strojovú čitateľnosť,



detekovateľnosť, interoperabilitu a transparentné označovanie synteticky vytvoreného obsahu, no nenahrádza právne povinnosti samotného nariadenia (Európska komisia, 2026).

Praktická pripravenosť trhu zostáva nerovnomerná. Rijsbosch, van Dijck a Kollnig (2026) preskúmali 50 systémov na generovanie obrazu a zistili primerané strojovo čitateľné označovanie len pri 38 percentách a viditeľné označenie synteticky vytvoreného obsahu pri 18 percentách. Meranie sa uskutočnilo ešte pred začatím uplatňovania povinností a autori upozorňujú na rýchle zmeny služieb, výsledok však dobre dokumentuje implementačnú medzeru.

Regulácia musí zachovať proporcionalitu. Povinné označenie môže podporiť transparentnosť, nesmie sa však zmeniť na všeobecný signál nízkej kvality alebo nepravdivosti. Umelecké, satirické, vzdelávacie a asistované redakčné použitie má iný komunikačný význam než podvodná imitácia. UNESCO preto zasadzuje technické riešenia do širšieho rámca ľudských práv, transparentnosti platforiem, mediálnej gramotnosti a zodpovednosti viacerých aktérov (UNESCO, 2023).

Záver

Watermarking môže znížiť neistotu o pôvode digitálneho obsahu, ak je vložený dôveryhodným systémom, prežije relevantné transformácie a jeho výsledok je správne vysvetlený. Textové schémy ukazujú, že štatistickú značku možno vložiť bez zjavného zhoršenia kvality, no parafrázovanie a krátky text obmedzujú detekciu. Obrazové metódy dokážu značku integrovať do generatívneho procesu, ale môžu byť oslabené pri opätovnom vytvorení podobného obrazu pomocou AI. SynthID a C2PA ukazujú, že spoľahlivejšie označovanie AI obsahu bude pravdepodobne vyžadovať kombináciu viacerých prístupov, nie jedno univerzálne riešenie.

Žiadny z týchto nástrojov nepotvrdzuje pravdivosť komunikovaného obsahu. Ich reálny prínos spočíva v poskytovaní overiteľných indícií pre platformy, redakcie a publikum. Najvhodnejší prístup preto kombinuje technické značenie s viditeľnou transparentnosťou, profesionálnym overovaním, mediálnou gramotnosťou a pravidlami zodpovednosti. Watermarking preto nemožno považovať za samostatný dôkaz pravosti obsahu, ale za jednu zo súčastí širšieho



procesu jeho overovania. Budúcnosť digitálnej dôveryhodnosti pravdepodobne nebude založená na jedinom technickom riešení, ale na kombinácii proveniencie, kryptografického overovania, transparentnosti platforiem a ľudského posudzovania.

Zoznam použitých zdrojov

CHANDRA, B. et al., (2024). *Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency* (NIST AI 100-4). National Institute of Standards and Technology. Online. Dostupné na: <https://doi.org/10.6028/NIST.AI.100-4>. [zobrazené 2026-05-08].

CHESNEY, R. a D. K. CITRON, (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, vol. 107 (2019), no. 6, s. 1753–1820. Online. Dostupné na: <https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security>. [zobrazené 2026-05-06].

COALITION FOR CONTENT PROVENANCE AND AUTHENTICITY, (2026). *C2PA Specifications 2.4*. Online. Dostupné na: <https://spec.c2pa.org/specifications/specifications/2.4/index.html>. [zobrazené 2026-05-09].

CONTENT AUTHENTICITY INITIATIVE, (2024). *Durable Content Credentials*. Online. Dostupné na: <https://contentauthenticity.org/blog/durable-content-credentials>. [zobrazené 2026-05-06].

DATHATHRI, S. et al., (2024). Scalable watermarking for identifying large language model outputs. *Nature*. Online. vol. 634 (2024), no. 8035, s. 818–823. Dostupné na: <https://doi.org/10.1038/s41586-024-08025-4>. [zobrazené 2026-05-09].

EUROPEAN COMMISSION, (2026). *Code of practice on transparency of AI-generated content*. Online. Dostupné na: <https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content>. [zobrazené 2026-05-10].



EUROPEAN PARLIAMENT AND COUNCIL OF THE EUROPEAN UNION, (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*. Online. Dostupné na: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>. [zobrazené 2026-05-10].

FENG, K. J. K. et al., (2023). Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions. *Proceedings of the ACM on Human-Computer Interaction*. Online. vol. 7 (2023), no. CSCW2, s. 1–42. Dostupné na: <https://doi.org/10.1145/3610061>. [zobrazené 2026-05-08].

FERNANDEZ, P.; G. COUAIRON; H. JÉGOU; M. DOUZE a T. FURON, (2023). The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Online. s. 22466–22477. Dostupné na: https://openaccess.thecvf.com/content/ICCV2023/papers/Fernandez_The_Stable_Signature_Rooting_Watermarks_in_Latent_Diffusion_Models_ICCV_2023_paper.pdf. [zobrazené 2026-05-06].

GOOGLE DEEPMIND, (2026). *SynthID: A tool to watermark and identify content generated through AI*. Online. Dostupné na: <https://deepmind.google/models/synthid>. [zobrazené 2026-05-07].

KIRCHENBAUER, J. et al., (2023). A Watermark for Large Language Models. In: *Proceedings of the 40th International Conference on Machine Learning*. Online. vol. 202, s. 17061–17084. Dostupné na: <https://proceedings.mlr.press/v202/kirchenbauer23a/kirchenbauer23a.pdf>. [zobrazené 2026-05-06].

LI, F. a Y. YANG, (2024). Impact of Artificial Intelligence–Generated Content Labels on Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation: Web-Based, Randomized, Controlled Experiment. *JMIR Formative Research*. Online. vol. 8 (2024), e60024. Dostupné na: <https://doi.org/10.2196/60024>. [zobrazené 2026-05-08].



LIU, A. et al., (2024). A Survey of Text Watermarking in the Era of Large Language Models. *ACM Computing Surveys*. Online. vol. 57 (2024), no. 2, s. 1–36. Dostupné na: <https://doi.org/10.1145/3691626>. [zobrazené 2026-05-06].

OPENAI, (2026). *Advancing content provenance for a safer, more transparent AI ecosystem*. Online. Dostupné na: <https://openai.com/index/advancing-content-provenance>. [zobrazené 2026-05-08].

PANG, Q.; S. HU; W. ZHENG a V. SMITH, (2024). No Free Lunch in LLM Watermarking: Trade-offs in Watermarking Design Choices. In: *Advances in Neural Information Processing Systems 37*. Online. Dostupné na: <https://doi.org/10.52202/079017-4402>. [zobrazené 2026-09-09].

RASTOGI, S. a D. PRUTHI, (2024). Revisiting the Robustness of Watermarking to Paraphrasing Attacks. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024, s. 18100–18110. Dostupné na: <https://doi.org/10.18653/v1/2024.emnlp-main.1005>. [zobrazené 2026-05-06].

RIJSBOSCH, B.; G. VAN DIJCK a K. KOLLNIG, (2026). Missing the Mark: Adoption of Watermarking for Generative AI Systems in Practice and Implications Under the New EU AI Act. *Policy & Internet*. Online. vol. 18 (2026), no. 2. Dostupné na: <https://doi.org/10.1002/poi3.70041>. [zobrazené 2026-05-15].

TRATTNER, C.; S. L. FORSTNER; A. D. STARKE a E. KNUDSEN, (2026). C2PA Provenance Labels Increase Trust in Digital News Platforms Across Western Countries. *Proceedings of the International AAAI Conference on Web and Social Media*. Online. vol. 20 (2026), no. 1, s. 2267–2279. Dostupné na: <https://doi.org/10.1609/icwsm.v20i1.42749>. [zobrazené 2026-05-10].

UNESCO, (2023). *Guidelines for the governance of digital platforms: Safeguarding freedom of expression and access to information through a multistakeholder approach*. Online. Dostupné na: <https://unesdoc.unesco.org/ark:/48223/pf0000387339>. [zobrazené 2026-05-10].



VACCARI, C. a A. CHADWICK, (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*. Online. vol. 6 (2020), no. 1. Dostupné na: <https://doi.org/10.1177/2056305120903408>. [zobrazené 2026-05-08].

WEN, Y.; J. KIRCHENBAUER; J. GEIPING a T. GOLDSTEIN, (2023). Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images. In: *Advances in Neural Information Processing Systems 36*. Online. Dostupné na: <https://doi.org/10.52202/075280-2529>. [zobrazené 2026-05-10].

WITTENBERG, C.; Z. EPSTEIN; G. PÉLOQUIN-SKULSKI; A. J. BERINSKY a D. G. RAND, (2025). Labeling AI-generated media online. *PNAS Nexus*. Online. vol. 4 (2025), no. 6. Dostupné na: <https://doi.org/10.1093/pnasnexus/pgaf170>. [zobrazené 2026-05-08].

ZHANG, H. et al., (2024). Watermarks in the Sand: Impossibility of Strong Watermarking for Language Models. In: *Proceedings of the 41st International Conference on Machine Learning*. Online. vol. 235, s. 58851–58880. PMLR. Dostupné na: <https://raw.githubusercontent.com/mlresearch/v235/main/assets/zhang24o/zhang24o.pdf>. [zobrazené 2026-05-06].

ZHAO, X. et al., (2024). Invisible Image Watermarks Are Provably Removable Using Generative AI. In: *Advances in Neural Information Processing Systems 37*. Online. Dostupné na: <https://doi.org/10.52202/079017-0276>. [zobrazené 2026-05-06].

Vyhlasenie o použití umelej inteligencie

Na prípravu a úpravu tohto článku boli použité nástroje umelej inteligencie. Na tvorbu konceptov, prípravu rešeršného podkladu, jazykovú a štylistickú úpravu bol použitý model GPT-5.6 Sol od spoločnosti OpenAI.



Autori

Ing. Martin Záborský
martin.zaborsky@umkd.uniza.sk

Ústav mediamatiky a kultúrneho dedičstva
Žilinská univerzita v Žiline
Univerzitná 8215/1
010 26 Žilina
Slovenská republika

Martin Záborský sa venuje informačným technológiám, správe informačných systémov a podpore vzdelávacích a výskumných aktivít. Zaujíma sa o spracovanie textových dát, strojové učenie a využívanie digitálnych technológií v akademickom prostredí. Zabezpečuje technickú podporu, správu a rozvoj digitálnych služieb a informačných systémov pre potreby vzdelávania, výskumu a vedeckej spolupráce.

doc. PhDr. Slavka Pitoňáková, PhD.
slavka.pitonakova@umkd.uniza.sk

Ústav mediamatiky a kultúrneho dedičstva
Žilinská univerzita v Žiline
Univerzitná 8215/1
010 26 Žilina
Slovenská republika

Slavka Pitoňáková sa zameriava na prepojenie médií a kultúry, pričom vo svojom výskume venuje pozornosť najmä spôsobom, akými médiá sprostredkujú, interpretujú a utvárajú význam kultúrneho dedičstva v rôznych mediálnych produktoch vrátane produkcie mainstreamových médií. Jej výskumný záujem sa sústreďuje na hodnotu mediálnych produktov a ich komplexné posudzovanie s dôrazom na kultúrny a spoločenský kontext. V tejto súvislosti sa parciálne zaoberá problémami, akými technologické zmeny ovplyvňujú vnímanie pôvodu a významu mediálneho obsahu.

The manuscript was received on 2nd July 2026, underwent peer review, was accepted for publication on 28th September 2026, and was published on 30th September 2026.

