Towards Trustworthy Digital Media in the AIGC Era: An Introduction to the Upcoming ISO JPEG Trust Standard

by Jiayun Mo, Xin Kang, Ziyuan Hu, Haibo Zhou, Tieyan Li, and Xiaojun Gu

Abstract

With the rapid development of Artificial Intelligence Generated Content comes the need for more advanced technologies to distinguish fake medias from authentic ones. Current standards for authenticating media assets are developed by the Coalition for Content Provenance and Authenticity and the International Press Telecommunications Council. The former focuses on providing complete provenance for media assets to allow users to backtrack the history of creation and modification, while the latter establishes a standard method for asserting and storing information related to media assets. JPEG Trust, on the other hand, is an evolving standard that aims to provide a way of proving the trustworthiness of media assets, thus directly addressing the aforementioned issue. Through the generation of a trust report, JPEG Trust allows users to directly view the evaluation results of the trustworthiness of media assets and make decisions accordingly. JPEG Trust is expected to bring application values for countries, platforms, and individuals and commercial values targeting companies and the wider society. However, there still exists potential challenges for JPEG Trust that demand a solution. With the possibilities JPEG Trust brings, it is expected to leave considerable impacts on the trustworthiness of the future after addressing challenges along the way.

Introduction

Almost all citizens living in the digitalized society have experiences engaging in a philosophical conversation with ChatGPT, admiring an awe-striking portrait from Midjourney, or enjoying the tones created by Soundraw. Indeed, the novelty and convenience brought by Artificial Intelligence Generated Content (AIGC) captured a significant amount of attentions.

However, despite their captivating nature, AIGC can be rather problematic at times. Hidden biases in the training datasets and inherent stereotypes picked up from training models can lead to inadvertent generation and dissemination of misinformation [1]. These fake medias, though unintentional, may easily twist the beliefs of the general public lacking precisions in their judgements. Furthermore, these issues further intensify as malicious parties manipulate information for financial or political gains; cyber bullying or social unrest might be aroused, posing serious threats to the well-being of societies [8]. With such technologies, they may also easily impersonate targeted individuals, celebrities or otherwise, and thus damage their reputation severely.

Throughout the years, attempts to effectively detect fake medias have been made using various technologies. Some focus on evaluating the credibility of the sources, which fail to acknowledge the possibility that an uncredible source might produce authentic content. Some choose to detect clickbait titles, but exaggerated titles do not necessarily infer their contents are flawed as well. Others use machine learning algorithms to conduct examinations and classifications of extracted features and draw reasonable conclusions [2]. While these advancing algorithms have satisfying accuracies when it comes to traditional fake medias such as news articles, they do not work as well when AI generated videos and images debuted and turned fake medias increasingly deceiving. AI models that have undergone adversarial trainings can generate fake medias that are invulnerable to such detection algorithms.

The sophistication, scale, and speed of manipulation techniques further add to the burden. Detecting the truthfulness and fakeness of each piece of media on the internet becomes impossible to accomplish in the status quo, yet the severity of their potential threats demands a solution urgently.

State of the Art

In recent years, another approach begins to emerge, targeting corroboration rather than falsification. Whereas the aforementioned algorithms aim to isolate fake contents from authentic ones, some organizations choose to establish standards used to uncover the authenticity of media assets. Two notable examples of such standards are respectively proposed by the Coalition for Content Provenance and Authenticity (C2PA) and the International Press Telecommunications Council (IPTC).

C2PA

C2PA is a project initiated by Adobe and Microsoft that builds upon the idea of providing "context and history for digital media" and "certifying the source and history (or provenance) of media content" in response to the overwhelming misinformation online [3].

The building block of C2PA is a verifiable unit named the C2PA Manifest, which consists of a claim signature, a claim, and multiple assertions; these components are bounded together by a claim generator [4]. The assertions are a series of statements representing trust signals for humans to discern the trustworthiness of an asset. Together, a set of C2PA manifests represents the history of an asset, also known as its provenance data. It allows verifiable and untamperable assertions that can help users form their own judgements, with the belief that the availability of such annotations can foster transparency and trust within the realm of digital media assets [5]. The C2PA standard has been widely embraced and left remarkable impacts on the community. Microsoft and Adobe, as part of C2PA's founding members, announced their plan to use the C2PA specification to authenticate AIGC; Leica and Nikon implemented hardware compatible with the C2PA standard, respectively in their Leica M11 camera and Nikon Z9 camera [10]. In addition, over 1500 companies are participating in the C2PA project through the Coalition Authenticity Initiative and over a hundred million photos are already using the C2PA standard.

Overall, C2PA's proposed idea of C2PA manifests has a focus on presenting the provenance of any media asset through different stakeholders' annotations. This grants individuals the ability to backtrack and look into the complete history of how an asset is created and modified, thus being able to make more informed judgements on the authenticity of the media asset.

IPTC

The IPTC Photo Metadata Standard is a widely used and recognized standard for photos, embraced by photographers, news agencies, photo agencies, or even libraries, museums, and other related industries [6].

The photo metadata include information embedded in the image files, or externally stored information in formats supported by IPTC. The data may be descriptive, explaining the image content, rights-related, documenting information about the creator and credits, or administrative, containing instructions and identifiers [6]. Together, these pieces of information are kept with the image files to help guarantee accurate identification and copyright protection. These standards have broad usages and influences; the photo metadata is supported by Google Images and over 20 other software, including Adobe Photoshop and ExifTool [6].

Overall, IPTC standardizes the formats and methods in which crucial information regarding the news media is stored. Though widely used and recognized for documenting photo metadata, it does not directly address the challenge of distinguishing authentic media assets.

An Overview of the Upcoming ISO JPEG Trust

Motivation and Progress

The massive amount of fake media accompanying the development of AI technology, as described in the Introduction, infers that people cannot easily build trust anchors and reach valid conclusions about whether to trust the media assets presented to them. How to provide technologies proving the trustworthiness of media assets and thus helping people reach valid judgements become a trending global issue.

Consequently, in January 2022, JPEG called for proposal for a solution to prove the authenticity of AIGC and discredit fake media from the pool of contents generated by artificial intelligence. JPEG Trust, a developing standard aiming to provide leadership in global establishment of trust in media assets, emerges.

Many creators want or need to declare modifications performed on media assets, to avoid hiding such manipulations that might lead to eventual spread of misinformation. JPEG Trust identifies the need for a standardized way of annotating media assets and securely linking the assets with their annotations. Compared to detecting the fakeness in medias, JPEG Trust provides a more proactive approach, where the disseminator of information needs to prove the authenticity and provide the complete provenance for the recipient, while the recipient only retains information and corresponding provenance he deems as authentic. After thorough research and investigation, JPEG Trust proposes standard mechanisms to describe and embed information for the creation of media assets and to protect these media assets' integrity.

Proposals were submitted by six different parties, including Adobe on behalf of C2PA, Huawei International Pte. Ltd., Sony Group, Vrije Universiteit Brussel, Polytechnic University of Catalonia, and Newcastle University. After evaluations of proposals in October 2022, JPEG officially initiated the development of JPEG Trust following the 98th JPEG Meeting in January 2023. The first Working Draft was presented during the 99th meeting in April, and the second Working Draft was completed in July 2023, during JPEG's 100th and most recent meeting. The Committee Draft was planned to be completed by October 2023.

Core Framework

The core principles of JPEG Trust are based on the Trust Framework displayed in Fig. 1.

A media asset consists of the metadata (including both trust record and other metadata) and content. The trust record is further separated into trust declaration and trust manifests. A trust manifest is the set of information about media asset provenance, while a trust declaration is a special type of trust manifest defined when a media asset is created, with mandatory assertions. The trust credential consists of trust indicators extracted from the two portions of the media asset, namely metadata and content. The trust profile is a set of trust indicators that is evaluated against a trust credential.

The trust report takes in a trust profile and a trust credential, and documents the result of evaluating the trust credential against the trust profile. The evaluation helps to indicate a level of trustworthiness for a given media asset. A trust indicator presented in the trust credential (extracted from the media asset) passes the evaluations if it satisfies the trust indicator requirement listed in the trust profile. For example, T1 from the trust

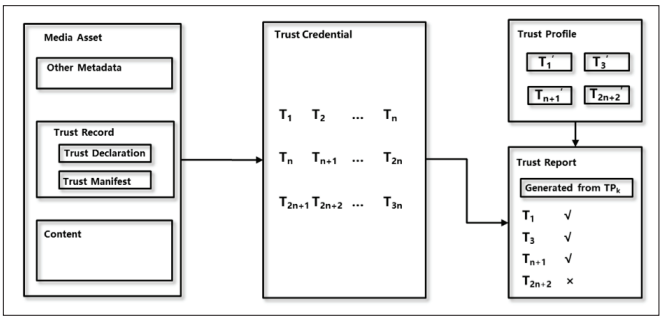

Figure 1. Trust framework.

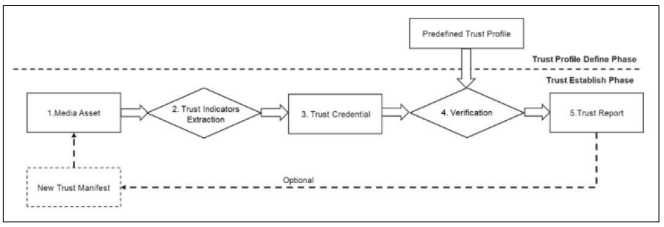

Figure 2. Trust report generation process.

credentials is checked against T1′ in the trust profile. If T1 corresponds to the city Beijing, then a T1′ value of either "location" or "China" means T1 passes the evaluation.

A new trust manifest can be generated based on the trust report optionally. The new trust manifest can be added to update the trust record of the media asset to further enhance the trustworthiness of the media asset.

Core Procedure

The procedure of trustworthiness evaluation is illustrated in Fig. 2. Firstly, the trust profile definition phase involves the creation of different trust profiles, which may be created by governments, media platforms, users, and others.

On the other hand, the trust establishment phase illustrated in the bottom part of Fig. 2 involves the uses of a generation algorithm, an extraction algorithm, and a verification algorithm, in addition to the aforementioned predefined trust profiles. First, a generation algorithm is used to generate the media asset, each consisting of metadata and content. The media asset is then passed to an extraction algorithm, which extracts the trust indicators from the metadata and content of the media asset. It also generates the trust credential for this media asset using the extracted trust indicators.

Next, the extraction algorithm sends the trust credentials and triggers the verification process. The verification algorithm requests for the predefined trust profile for the scenario from the trust profile storage. With that trust profile, the verification algorithm now checks whether the trust indicators in the trust credential satisfy the predefined trust profile. The verification algorithm also generates the trust report, which documents the result of evaluating a trust credential against a trust profile.

Afterwards, the verification algorithm may also undergo an optional process and generate a new trust manifest based on the trust report. The new trust manifest will be sent back to the generation algorithm, which will update the media asset by adding in the new trust manifest, forming an iterative process.

Differences with C2PA and IPTC

The distinguishing difference between C2PA, IPTC, and JPEG Trust lies in their different focuses. Whereas C2PA emphasizes assertions and provenance, and IPTC provides storage of information about news media, JPEG Trust is striving to evaluate trustworthiness. Through the generation of trust reports, a new

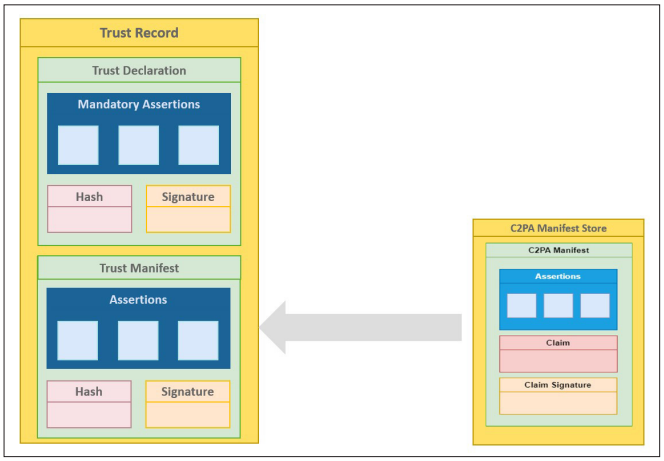

FIGURE 3. Trust record and C2PA manifest.

concept proposed by JPEG Trust, users are presented with evaluation results about the trustworthiness of media assets, helping them to decide whether to trust them. This brings a wide range of potential applications for JPEG Trust, which will be discussed in the Impacts of ISO JPEG Trust section.

JPEG Trust emphasizes the concept of trust. Thus, although it is based on similar architectures as C2PA, different terminologies are used to refer to the same component. The concepts of asset, digital content, and provenance in C2PA are respectively named media asset, media asset content, and media asset provenance in JPEG Trust. Similarly, C2PA Manifest, the first C2PA Manifest in a C2PA Manifest Store, and the C2PA Manifest Store are respectively named trust manifest, trust declaration, and trust record in JPEG Trust.

An important building block of JPEG Trust is the trust record. Its structure is derived based on C2PA's architecture of a manifest, with their differences outlined in Fig. 3.

As can be seen, a trust record in the JPEG Trust framework contains a trust declaration and a trust manifest; the former is a special instance of the latter defined during the creation of the media asset and contains mandatory assertions. The main functionality of the mandatory assertions demanded is to prevent adversaries from purposefully hiding important information, guaranteeing the trustworthiness of media assets. The concepts of Trust Declaration, which represents a new type of trust manifest, is unique to JPEG Trust, but the structure used is compatible with C2PA. The trust manifest remains similar to the C2PA manifest, displaying the provenance of media assets. Together, one trust declaration and multiple trust manifests make up the trust profile.

Impacts of ISO JPEG Trust

The impacts of implementing JPEG Trust can be separated into application values and commercial values, which will be discussed in the following sections respectively.

Application Values

After implementing the trust framework, it will no longer be necessary to individually assess the trustworthiness of each media asset. Instead, different trust profiles will be retrieved based on different scenarios and trust reports will be generated after evaluations, which effectively enhances the convenience and efficiency of proving trustworthiness. The advantages of directly retrieving trust profiles and checking the results of the trustworthiness evaluations can be seen in the following three scenarios.

Firstly, different countries may have different trust indicators when assessing trustworthiness. Using the trust framework, generated trust reports can directly showcase whether one media asset satisfies the criteria set by each country. The standardization and admission problems across different countries will thus be resolved. Media assets do not need to be assessed repeatedly for each country's criteria, and sharing trustworthy media assets internationally will become more efficient.

Secondly, different platforms have different criteria for assessing uploaded contents. Trust reports will explicitly display whether the media asset meets each particular platform's criteria, without the need to reevaluate media assets continuously for different platforms. It saves both time and effort to allow trustworthy media assets to be approved by multiple platforms efficiently.

Thirdly, individuals will have different specified needs for the judgement of trustworthiness. Evaluation results in trust reports as well as the combinations of trust indicators used will help them reach a conclusion about the trustworthiness of the media assets, thus helping with their decision-making process.

As can be seen, applying JPEG Trust will generate positive impacts for a wide range of different kinds of users, addressing the differences between the needs of countries, platforms, and individuals in an efficient manner.

Commercial Values

Fake media gains incrementing attention with the development of large models and AIGC. The mainstream technologies' focus on detecting fakeness in medias struggles to keep up with the pace of the massive generation of increasingly deceiving fake medias. Comparably, JPEG Trust is a more efficient and advanced solution that precisely addresses the needs to authenticate AIGC demanded by society urgently. The launch of such a standard will be expected to face overwhelming responses.

Moreover, the aforementioned wide embracement for C2PA shortly after its initial release indirectly shows the potential market for JPEG Trust. It is reasonable to expect software companies to incorporate JPEG Trust standards into authentication processes of media assets and expect manufactures to manufacture their hardware products to be compatible with JPEG Trust standards. With JPEG Trust's more thorough considerations and comprehensive functionalities compared to C2PA, it will possess significant commercial prospects ahead.

Furthermore, with recent progresses in the research about 5.5G and 6G, immersive communication methods are under development. Voice and gesture controls, face recognitions, avatars, extended realities, video rendering technologies will all contribute to increase the fun of communication and provide users with multimedia, visualizable, and fully interactable new communication experiences. The launch and promotion of these new communication methods require the assistance of authentication technologies [9]. People need to understand the trustworthiness of media assets before enjoying the entertainment brought by such technologies [7]. JPEG Trust will be able to complete the task adequately, delving into a large market with immense possibilities in the near future.

In addition, the European Union (EU) plans to pass an EU Artificial Intelligence Act at the end of 2024, which will be the first concrete regulation of AI. This timeline coordinates nicely with JPEG Trust's development plan. Thus, JPEG Trust has the potential of being cited by the EU AI Act and becoming a core standard within it. The authority and influence it may assert on Europe and the world as a whole will bring bright potentials for JPEG Trust.

The escalating advancements in technologies require accompanying efforts about media trustworthiness; JPEG Trust will have ample opportunities and noteworthy business prospects after its launch.

The Road Ahead

As JPEG Trust is a comparably new standard, still under development and constantly evolving, there exist a few potential challenges crucial to the success and effectiveness of JPEG Trust that demand decent solutions.

Trust Indicators

The first and perhaps most significant challenge centers around defining trust indicators. Trust indicators play one of the most crucial roles in determining the trustworthiness of a media asset, since such a conclusion is derived after evaluating extracted trust indicators from a media asset against trust indicators in a trust profile. Thus, deciding which factors can count as trust indicators and determining their effectiveness is vital. However, there exists plentiful of information contained in the Exchangeable Image File Format (EXIF) of an image. Choosing the relevant information as trust indicators from the excessive available information poses a challenge. Moreover, trust is a complex and rather dynamic concept; it is not possible to come up with a static standard fitting to all circumstances. For example, trust indicators for photographs may be distinctively different to those for artworks. Defining trust indicators for each individual scenario further magnifies the challenge. Another problem is about whether to make a trust indicator mandatory. Mandatory trust indicators may incur privacy problems, such as when the creator field becomes mandatory information while the person prefers to remain anonymous. Overall, defining a set of comprehensive and effective trust indicators is a challenge faced by JPEG Trust.

Privacy Issues

In addition to mandatory trust indicators, privacy concerns that may emerge from elsewhere within the trust framework need to be thoroughly examined and addressed. Cryptographic keys and signed certificates help ensure the security of data and private information. As a result, any mistreatment and leakage of keys will directly release such data to the adversaries, diminishing all efforts for privacy. Proper key management and certificate management are required to ensure the privacy of each stakeholder involved.

Watermarking

Another challenge is related to watermarking media assets. These watermarks may be visible or invisible. The fact that visible watermarks can appear too conspicuous and interfere with the aesthetics of images justifies the preferences for invisible watermarks. This is especially true in the domain of creative works and AIGC, where watermarks can greatly interfere with people's appreciation of the media assets. In order to apply invisible watermarks, one need to be well aware of the algorithm; however, knowledge of the algorithm implies that one possesses the ability to reverse the watermark. How to address the challenge and best incorporate watermarks on media assets remain a problem to be resolved.

Compatibility

JPEG Trust's compatibility with existing standards, mainly the aforementioned C2PA and IPTC, need to be accounted for. A solution that takes the existing standards and their respective requirements into considerations is demanded. Finding a way to require only minor and minimal updates to stay compatible without unnecessary divisions between the different standards will be ideal.

These potential challenges require satisfying solutions in order to ensure that JPEG Trust can serve its purposes and achieve its visions and missions. However, given that JPEG Trust is still progressing, eventual solutions to address each challenge can be expected in future editions.

Conclusion

In this paper, solutions tackling the prevalence of fake media are explored, with a specific focus on examining the trust framework of JPEG Trust and how the concept of trust report helps to display the trustworthiness of media assets to users. The expected impacts of JPEG Trust include the efficiency and effectiveness for countries, platforms, and users to authenticate media assets, as well as the commercial values after being embraced by manufactures, companies, 6G immersive communication technologies, or incorporated into the EU AI Act. However, the project is still under development; potential challenges for JPEG Trust include the definition of trust indicators, measures to ensure privacy, watermarking techniques, as well as compatibility with existing standards. After addressing these challenges and overcoming other obstacles along the way, JPEG Trust is expected to leave remarkable impacts in the domain of trust within modern society.

References

[1] L. Lyu, C. Chen, and J. Fu, "A Pathway Towards Responsible AI Generated Content," *Proc. 2nd Int'l. Joint Conf. Artificial Intelligence*, 2023; 10.24963/ijcai.2023/803.

[2] B. Al Asaad and M. Erascu, "A Tool for Fake News Detection," *2018 20th Int'l. Symp. Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 2018; 10.1109/synasc.2018.00064.

[3] Coalition for Content Provenance and Authenticity, "Overview," C2PA, 2023; https://c2pa.org/.

[4] Coalition for Content Provenance and Authenticity. "C2PA Technical Specifications," C2PA, 2023; https://c2pa.org/specifications/specifications/1.3/specs/C2PA_Specification.html.

[5] Coalition for Content Provenance and Authenticity. "Guiding Principles for C2PA Designs and Specifications," C2PA, 2023; https://c2pa.org/principles/.

[6] International Press Telecommunications Council, "IPTC Photo Metadata Standard," IPTC, 2023; https://www.iptc.org/standards/photo-metadata/iptc-standard/.

[7] H. L. J. Ting *et al.*, "On the Trust and Trust Modeling for the Future Fully-Connected Digital World: A Comprehensive Study," *IEEE Access*, vol. 9, 2021, pp. 106,743–83.

[8] H. Wang *et al.*, "An Overview of Trust Standards for Communication Networks and Future Digital World," *IEEE Access*, vol. 11, 2023, pp. 42,991–98.

[9] Y. Wang *et al.*, "SIX-Trust for 6G: Toward a Secure and Trustworthy Future Network," *IEEE Access*, vol. 11, 2023, pp. 107,657–68

[10] A. Parsons, "Major Steps Forward for the CAI: Partnerships with Leica and Nikon, New Content Credentials features in Photoshop and Beyond at MAX 2022," Adobe Blog, 2022; https://blog.adobe.com/en/publish/2022/10/18/major-steps-forward-cai-partnerships-leica-nikon-new-content-credentials-features-photoshop-beyond-max-2022.

Biographies

Jiayun Mo (jmo005@e.ntu.edu.sg) is currently pursuing the Bachelor of Science degree in Computer Science at Nanyang Technological University, Singapore. During her internship at Shield Laboratory, the Singapore Research Center of Huawei International Pte. Ltd., she delved into research about trust modeling and its related applications. She intends to aim for further studies in order to better pursue her current research interests of human-centered applications of technologies.

Xin Kang (kang.xin@huawei.com) is senior researcher at Huawei Singapore Research Center. He received his Ph.D. Degree from National University of Singapore. He has more than 15 years' research experience in wireless communication and network security. He is the key contributor to Huawei's white paper series on 5G security. He has published 70+ IEEE top journal and conference papers, and received the Best Paper Award from IEEE ICC 2017, and Best 50 Papers Award from IEEE GlobeCom 2014. He has also filed 60+ patents on security protocol designs, and contributed 30+ technical proposals to 3GPP SA3. He is also the initiator and chief editor for ITU-T standard X.1365, X.1353, and the on-going work item Y.atem-tn.

Ziyuan Hu (huziyuan@huawei.com) received the BSc degree and Ph.D. degree from the Department of Computer Science and Engineering, Shanghai JiaoTong University. Currently, he is working as Senior Engineer in Huawei Technologies Co., Ltd.. His research interests include information security and media authenticity.

Haibo Zhou (zhouhaibo10@huawei.com) received the Ph.D. degree from the Department of Mathematics, Shandong University. Currently, he is working as postdoc in Huawei Technologies Co., Ltd.. His research interests include information security and media authenticity.

Tieyan Li (li.tieyan@huawei.com) is currently leading Digital Trust research, on building the trust infrastructure for future digital world, and previously on mobile security, IoT security, and AI security at Shield Lab, Singapore Research Center, Huawei Technologies. He is also the director of Trustworthy AI C-TMG and the vice-chairman of ETSI ISG SAI.

Xiaojun Gu (guxiaojun1@huawei.com) received science of computer master degree of Beijing Institute of Technology in 2000 and EMBA of University International Business and Economics in 2010. He is the standards director at Huawei, has more than 20 years of media relevant international standards experience, and is active in ITU-T SG16, ISO TC42 and ISO/IEC/JTC1/SC24,29.