IPTC and PLUS: Response to NITRD NCO Request for Information on the Development of an AI Action Plan

A solution for communicating critical data mining rights information using embedded metadata in image and video files

This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI Action Plan and associated documents without attribution.

Executive Summary

This document describes suggestions from the International Press Telecommunications Council (IPTC) and the PLUS Coalition (PLUS) regarding several metadata properties that can be embedded directly into digital image and video files to let others know where and how the file was created, as well as directions on how the file may be used by others that are data mining such content.

Based in the UK since its foundation in 1965, the IPTC brings together the world's leading news agencies, publishers and industry vendors. The IPTC's mission is to simplify the distribution of information. IPTC develops and promotes efficient technical standards to improve the management and exchange of information between content providers, intermediaries and consumers. The IPTC Photo Metadata Standard is the most widely-adopted means of embedding descriptive, administrative and rights metadata into image and video files, used by thousands of photographers, agencies, publishers and distributors of visual media content.

Based in the US and founded in 2004, PLUS is a global coalition of communities engaged in creating, distributing, using and preserving images. The PLUS mission is to simplify the communication and management of image rights. The PLUS License Data Format, a standard schema for communicating image rights information, was first published in 2006 and is integrated in all manner of software used for creating, distributing, using or preserving images.

On behalf of our respective memberships, the IPTC and PLUS respectfully suggest that existing copyright law is sufficient to enable licensing of content to AI platforms. A “fair use” provision does not cover commercial AI training. Existing United States copyright law should be enforced.

The IPTC and PLUS solution provides a technical means for expressing the creator's intent as to whether their creations may be used in generative AI training data sets. This takes the form of metadata embedded in image and video files. This solution, in combination with other technical mechanisms such as the Text and Data Mining Reservation Protocol ([TDMRep]), could take the place of a formal licence agreement between parties. This would make an opt-in approach scalable and therefore technically feasible.

A solution based solely on robots.txt, or another site-wide mechanism for expressing a website owner's willingness to offer the site's content to AI providers as training data, belies an oversimplified and inaccurate understanding of the modern media ecosystem. Many images and media files that are published by small and large media organisations are not actually owned by those organisations. Wire services, image agencies and individual photographers maintain license agreements with publishers, who then use their valuable content under licence, under a restricted set of permissions. Simply put, much of the content on publisher websites is not theirs to give. For this reason, we request that any approach endorsed or mandated by the US government includes item-specific opt-in using embedded media metadata formats such as the IPTC Photo Metadata Standard and the PLUS License Data Format.

It is true that our technical solutions would also be relevant if the US government chooses to implement an "opt-out" approach; a broad data mining exception with a provision for content owners to reserve their rights through electronic means. However, an opt-out-based approach does not currently protect owners' rights, due to the common practice of "metadata stripping." This is an activity, built into many publisher workflows and social media platforms, that removes metadata embedded in media files, even critical rights and accessibility based information, in the misguided belief that it will improve site performance[1]. Even though Google recommends against it [Google-Metadatastripping], metadata stripping is routinely performed by many publishers and publishing systems, often inadvertently. Due to this practice, the absence of an opt-out directive in an image's metadata section should not be understood as equivalent to an opt-in. We can only recommend that the US government adopts an opt-in approach.

Introduction

Image creators, copyright owners, distributors, and publishers ("stakeholders") approached IPTC and PLUS with concerns about their images being used in AI training data. Stakeholders were frustrated that their images were being mined for AI training and generative purposes

[1] see the IPTC study "Rights Information and Social Media Networks" for detailed information on which platforms strip embedded metadata from uploaded image and video files. The IPTC and PLUS Coalition have both campaigned on this issue for more than ten years: see the Embedded Metadata Manifesto which we created in 2011.

without their knowledge or authorisation, simply because those images were accessible on the internet.

The opt-out solutions proposed by AI providers, such as robots.txt directives or manual web forms, were not suitable to these stakeholders: image creators often do not have access to web servers to be able to create or edit robots.txt files. Image creators also want their rights requirements to travel along with the image if it is downloaded and re-used on another web service. Therefore a solution that is embedded in the media file itself is most appropriate.

In response to these stakeholder requests, the PLUS Coalition ("PLUS"), in partnership with the International Press Telecommunications Council ("IPTC"), developed the PLUS "Data Mining" property to provide stakeholders with a simple, readily accessible way to communicate essential data mining information via an embedded Extensible Metadata Platform (XMP) property supported by a large number of digital image and video formats.

The Digital Source Type property (first released in 2008, and refreshed in 2014 and 2024) can be used to "tag" an increasing number of image types, including those created using generative AI (or incorporating generative AI systems to alter the image).

Details

The IPTC/PLUS "Data Mining" property uses a short list of controlled terms to communicate whether data mining is prohibited or allowed — either in general, for AI or Machine Learning purposes, or for generative AI/ML purposes. Selecting one term from the standardized controlled list is sufficient to express data mining permissions, constraints and prohibitions applicable to crawlers, AI platforms and others. This readily accessible rights information allows any system to read and interpret data mining information embedded in these image files — and AI platforms and others can rely on that metadata to make informed decisions about mining and using images published to the internet.

The definitions for the controlled list were derived through an open process which included public review by numerous organisations and individuals across 140 countries, representing diverse stakeholders engaged in creating, distributing, using and preserving images. Inspired in part by recent efforts in global legislation reforms such as the EU Data Act, this process included extensive deliberation by working groups in the PLUS and IPTC communities, to define the terms and controlled vocabulary for use in the context of data mining, with as much clarity as possible.

IPTC and PLUS periodically update their standards, and will adopt terminology as it evolves to accommodate real-world workflows. The Data Mining property is also suitable for adoption for use in communicating data mining rights for other types of media such as text documents and audio files.

To balance the needs of the creative industries to protect their content with the desire to promote innovation in AI, we encourage the US government to identify a solution based on an “opt-in” paradigm.

It is important that rights reservation information is truly machine-readable in a way that is practical for both AI developers and content owners to implement. Such a technology already exists: the Text and Data Mining Reservation Protocol ([TDMRep]), which can be used in combination with embedded metadata in media files.

We propose that AI developers should be required to take embedded metadata into account when determining which content is authorized for their use in training data sets. The use of these properties by rights holders is voluntary, and it should not be presumed that every file might contain this information. However it would be helpful if developers look for this information, and encourage others to include (embed) this information in their content.

Our standardized approach already allows creators to reserve their rights using machine-readable formats. The PLUS/IPTC Data Mining property is designed to be both machine-readable and simple enough to be read and understood by humans. Just like all other IPTC and PLUS standards, this property is open and free for anyone to use. There are a number of software tools, both commercial and open source, that have support for this property, which should make it easy for anyone to apply this value to one or hundreds of images and/or video files. This work can and should be added to an existing workflow as part of a good licensing practice/strategy.

AI developers can also filter based on the “Digital Source Type” property. As some studies suggest that AI models collapse when trained on recursively generated data [Nature]. it could be useful to have an automated way to avoid processing existing generative AI images. This also could be part of a standard workflow. Last year, Google announced that images with embedded tags identifying them as generative AI outputs will display the label 'Image self-labeled as AI generated' in Google Images search results under “About this image.”

The most widely-suggested mechanism for AI opt-out, the robots.txt standard [RobotsTxt], cannot provide the granular control over the use of works that many rights holders seek. It allows works to be blocked from web crawling at the site or directory level, but does not recognize reservations associated with individual works displayed on web pages. It also does not enable rights holders to distinguish between uses of works. For example, they may be satisfied that web crawlers use their works for search indexing or language training, but not for generative AI. Robots.txt does not currently allow for this degree of control. Furthermore, Robots.txt is not an efficient method for communicating rights information for individual image files published to a web platform or website: rights information typically varies from image to image, and the publication of images to websites is increasingly dynamic.

In addition, the use of robots.txt requires that each user agent must be blocked separately, repeating all exclusions for each AI engine crawler robot. This creates challenges when new, unknown AI crawlers emerge, as they can index a site before being identified and added to the robots.txt file. To maintain control, publishers must regularly check their server logs to identify new user agents crawling their data and update their blocking protocols accordingly.

In contrast, embedding rights declaration metadata directly into image and video files provides media-specific rights information, protecting images and video resources whether the site/page structure is preserved by crawlers or the image files are scraped and separated from the original page/site. The owner, distributor, or publisher of an image or video can embed a coded signal into each file, allowing downstream systems to read the embedded XMP metadata and to use that information to sort/categorize images and to comply with applicable permissions, prohibitions and constraints.

IPTC, PLUS and XMP metadata standards have been widely adopted and are broadly supported by software developers, as well as in use by major news media, search engines, and publishers for exchanging images in a workflow as part of an "operational best practice." For example, Google Images currently uses a number of the existing IPTC and PLUS properties to signal ownership, licensor contact info and copyright. For details see [IPTC-GoogleImages].

Adoption and Use

While the PLUS Data Mining Property was first published in September 2023 (and immediately adopted by the IPTC in October 2023), the underlying method of tagging/embedding images and for reading embedded XMP metadata is commonplace worldwide and has existed for several decades.

By simply reading the information stored in the "Data Mining" property within image files, systems can automatically sort and process image files by several criteria, to ensure respect for (and compliance with) the rights of copyright owners. For example ExifTool, the most popular open-source image metadata tool, has supported the property since September 2023. Through the use of plugins, tools such as Adobe Photoshop have also adopted the new property.

Supplementing the Data Mining property is the IPTC's "Digital Source Type" field [IPTC-DigitalSourceType]. This allows the content creator to declare the origin of a piece of content, embedded directly in the file. This property is being used by Google, Meta, Apple and others to signal content that was generated by AI, and to surface that information to users. The IPTC Digital Source Type property has also been adopted into the C2PA specification, used to create One interesting side-point is that the Digital Source Type property can be used to allow generative AI systems to avoid accidentally ingesting AI-generated images (some studies suggest that using AI-generated content in a training set can “poison” an AI model, leading to garbage output).

Summary and Conclusion

1. On behalf of our memberships, IPTC and PLUS respectfully suggest that existing copyright law is sufficient to enable licensing of content to AI platforms. A “fair use” provision does not cover commercial AI training. Existing United States copyright law should be enforced.
2. IPTC and PLUS Photo Metadata provide a technical means for expressing the creator’s intent as to whether their creations may be used in generative AI training data sets. This takes the form of metadata embedded in image and video files. This solution, in combination with other solutions such as the Text and Data Mining Reservation Protocol [TDMRep], could take the place of a formal licence agreement between parties, making an opt-in approach technically feasible and scalable.
3. It is true that our technical solutions would also be relevant if the US government chose to implement an opt-out based approach. However, this does not currently protect owners’ rights well due to the routine activity of “metadata stripping” - removing important rights and accessibility metadata that is embedded in media files, in the misguided belief that it will improve site performance. Metadata stripping is performed by many publishers and publishing systems - often inadvertently.
4. As a result, we can only recommend that the US adopts an opt-in approach. We request that the US government ensures that metadata embedded in media files be declared as a core part of any technical mechanism to declare content owner’s desire for content to be included or excluded from training data sets.

Content creators are a core part of the US economy and have a strong voice. We agree with their position, but we don’t simply come with another voice of complaint: we bring a viable, ready-made technical solution that can be used today to implement true opt-in data mining permissions and reservations.

About our organizations

Founded in 1965 and based in London, the IPTC brings together the world's leading news agencies, publishers and industry vendors. The IPTC's mission is to simplify the distribution of information. IPTC develops and promotes efficient technical standards to improve the management and exchange of information between content providers, intermediaries and consumers. The IPTC’s media-type agnostic Information Interchange Model (IIM) format of the IPTC photo metadata standard was introduced in 1990, and the version using Adobe's Extensible Metadata Platform (XMP) debuted in 2004. Additional historical details about the IPTC specification can be found at [IPTC-PMDHistory].

Founded in 2004, PLUS is a global coalition of communities engaged in creating, distributing, using and preserving images. The PLUS mission is to simplify the communication and management of image rights. The PLUS License Data Format, a standard schema for communicating image rights information, was first published in 2006 and is integrated in all manner of software used for creating, distributing, using or preserving images.

Author Contacts

We welcome inquiries for more details and further discussions on this topic, and would be happy to be involved in any workshops or reviewing draft policy papers.

Brendan Quinn, Managing Director, IPTC <mdirector@iptc.org>
Michael Steidl <mwsteidl@newsit.biz>
David Riecks <david@riecks.com>
Jeff Sedlik <js@plus.org>
Margaret Warren <mwarren@ihmc.org>

References

[Google-Metadatastripping] Google wants you to label AI-generated images used in Merchant Center
https://searchengineland.com/google-wants-you-to-label-ai-generated-images-used-in-merchant-center-437645

[IPTC-DataMining] section in the IPTC Photo Metadata Specification
https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata#data-mining

[IPTC-PMDHistory] History of the IPTC Photo Metadata Standard
https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata#history

[IPTC-UserGuide-Datamining] from the IPTC Photo Metadata User Guide
http://www.iptc.org/std/photometadata/documentation/userguide/#_data_mining

[PLUS-DataMining] Picture Licensing Universal System (PLUS) License Data Format for Data Mining
https://ns.useplus.org/LDF/ldf-XMPSpecification#DataMining

[IPTC-DigitalSourceType] from the IPTC Photo Metadata Specification
https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata#digital-source-type

[IPTC-DigitalSourceType-Userguide] from the IPTC Photo Metadata User Guide
https://www.iptc.org/std/photometadata/documentation/userguide/#_guidance_for_using_digital_source_type

[IPTC-GoogleImages] Quick guide to IPTC Photo Metadata and Google Images
https://iptc.org/standards/photo-metadata/quick-guide-to-iptc-photo-metadata-and-google-images/.

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

[RobotsTxt] RFC9309: Robots Exclusion Protocol
https://www.rfc-editor.org/rfc/rfc9309.html

[TDMRep] W3C Community Group: Text and Data Mining Reservation Protocol
https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/