IPTC and PLUS: The “Data Mining” Embedded Image/Video Metadata Property

A solution for communicating critical data mining rights information with metadata embedded into image and video files.

Abstract

This document describes the PLUS "Data Mining" property developed by the PLUS Coalition in partnership with the International Press Telecommunications Council (IPTC). This property provides a means for stakeholders to communicate essential data mining rights information via embedded Extensible Metadata Platform (XMP) metadata in digital image and video formats. This mechanism allows for clear communication of data mining permissions, prohibitions, and constraints, which can be readily accessed and interpreted by crawlers and AI platforms.

Introduction

Image creators, copyright owners, distributors, and publishers (“stakeholders”) had concerns about images being used in AI training data.. Stakeholders were frustrated that their images were being mined for AI training and generative purposes without their knowledge or authorization — simply because those images were accessible on the internet. 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, therefore a solution that is embedded in the media file itself is most appropriate.

In response to 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.

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 value from the standardised 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 vocabulary 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 documents and audio files.

While the IETF request for proposals refers to a “right to opt-out” of data mining, most countries have yet to implement laws or regulations defining or providing such a right. It is possible that some countries may implement “opt-in” requirements, to conform with existing laws governing the exploitation of intellectual property. We encourage the IETF to identify a solution that will support both “opt-in” and “opt-out” paradigms.

We respectfully suggest that Robots.txt alone is not a viable solution. Robots.txt may allow for communication of rights information applicable to all image assets on a website, or within a web directory, or on specific web pages. However, it is not an efficient method for communicating rights information for individual image files published to a web platform or website; as rights information typically varies from image to image, and as 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. As a result, agents can only be blocked retrospectively — after they have already indexed a site once. This requires that publishers must constantly check their server logs, to search for new user agents crawling their data, and to identify and block bad actors.

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 can embed a coded signal into each image file, allowing downstream systems to read the embedded XMP metadata and to use that information to sort/categorise 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

https://iptc.org/standards/photo-metadata/quick-guide-to-iptc-photo-metadata-and-google-images/.

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.

Supplementing the Data Mining property, the IPTC’s “Digital Source Type” property allows generative AI systems to avoid accidentally ingesting/re-ingesting AI-generated images. https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata#digital-source-type

Application developers who work with embedded metadata and images are already deploying the data mining field in their applications in anticipation that a data mining field will gain more widespread adoption in the future.

Organisations

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 are at:
https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata#history

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.

References

[IPTC-DataMining] from the International Press Telecommunications Council (IPTC) Photo Metadata Specification
https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata#data-mining

[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 International Press Telecommunications Council (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

Goals

With this IETF proposal, we aim to encourage and promote widespread awareness and adoption of the Data Mining property as a solution for communicating data mining permissions, prohibitions, and constraints, suitable for use in opt-in and opt-out scenarios. We seek to allow stakeholders and the internet community at large to easily access information about the Data Mining Property and how to freely use this open standard.

Author Contacts

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