



Best Practices on AI opt-out

Brendan Quinn
Managing Director, International Press
Telecommunications Council (IPTC)



What's the problem?

- Many “standards” for AI opt-out exist – which to choose?
- What solutions work for images, video, text, web pages?
- How to handle third-party supplied content, e.g. images from wire services?
- Are the required steps different in different jurisdictions? (Automatic opt-in vs opt-out)
- How do we work with agents, RAG, inference, AI search?
- How do opt-out instructions work with third-party crawlers eg Common Crawl and LAION?
- Which solutions are actually respected by AI providers today?
- Are there any simple steps that we can recommend to publishers?

Robots.txt

- Most commonly used approach
- Standardised by the IETF as [RFC9309](#) in 2022
- Allow / disallow at "user agent" level i.e. using the identifier text for each crawler
- Publishers must list each bot's "user agent" separately

RFC 9309

Robots Exclusion Protocol

Table of Contents

Abstract

This document specifies and extends the "Robots Exclusion Protocol" method originally defined by Martijn Koster in 1994 for service owners to control how content served by their services may be accessed, if at all, by automatic clients known as crawlers. Specifically, it adds definition language for the protocol, instructions for handling errors, and instructions for caching.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force, representing the consensus of the IETF community. It has received public review and approval by the Internet Engineering Steering Group (IESG). For more information on the status of this document, any errata, or how to provide feedback on it may be obtained at <https://www.rfc-editor.org/info/rfc9309>.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe the rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.

- 1. Introduction
 - 1.1. Requirements Language
- 2. Specification
 - 2.1. Protocol Definition
 - 2.2. Formal Syntax
 - 2.2.1. The User-Agent Line
 - 2.2.2. The "Allow" and "Disallow" Lines
 - 2.2.3. Special Characters

```
User-Agent: *
Disallow: *.gif$
Disallow: /example/
Allow: /publications/

User-Agent: foobot
Disallow: /
Allow: /example/page.html
Allow: /example/allowed.gif


User-Agent: barbot
User-Agent: bazbot
Disallow: /example/page.html

User-Agent: quxbot
```



TDM Reservation Protocol

- Created in response to EU Copyright Directive in early 2024
- Works for all types of content
- Based on HTTP
- Three methods (can be mixed):
 - /.well-known/tdmrep.json file
 - HTTP headers
 - <meta> tag in HTML pages
- Rules can differ based on URL
- Can optionally offer detailed policy statements using ODRL (RightsML)



TDM Reservation Protocol (TDMRep)

Final Community Group Report 10 May 2024

This version:
<https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/>

Editor:
Laurent Le Meur ([EDRLab](#))

Previous version
[TDMRep, Second Version](#)

Useful links
[TDMRep Documents](#)
[TDMRep CG Home Page](#)

Feedback:
[GitHub w3c/tdm-reservation-protocol](#) (pull requests, new issue, open issues)
public-tdmrep@w3.org with subject line [tdmrep] – message topic – ([archives](#))

Copyright © 2021-2024 the Contributors to the TDM Reservation Protocol (TDMRep) Specification, published by the [Text and Data Mining Reservation Protocol Community Group](#) under the [W3C Community Final Specification Agreement \(FSA\)](#). A human-readable summary is available.

TABLE OF CONTENTS

- Abstract
- Status of This Document
- 1. Introduction
- 2. Terminology
- 3. Conformance
- 4. Requirements
- 5. Declaring the reservation of TDM Rights
 - 5.1 tdm-reservation
 - 5.2 tdm-policy
- 6. Protocol
 - 6.1 TDM File on the Origin Server
 - 6.1.1 Use of regular expressions
 - 6.1.2 Examples
 - 6.2 TDM Header Field in HTTP Responses
 - 6.3 TDM Metadata in HTML Content
 - 6.4 TDM Metadata in EPUB 2 files

```
[
  {
    "location": "/directory-a/",
    "tdm-reservation": 1
  },
  {
    "location": "/directory-b/html/",
    "tdm-reservation": 1,
    "tdm-policy": "https://provider.com/policies/policy.json"
  },
  {
    "location": "/directory-b/images/*.jpg",
    "tdm-reservation": 0
  }
]
```



IPTC/PLUS Data Mining property

- Applies to images and video via embedded XMP metadata
- Carried along with the file
- Vocabulary defined at <https://ns.useplus.org/LDF/ldf-XMPSpecification#DataMining>

11.7. Data Mining

Row header	Specification
Name	Data Mining
Definition	Data mining prohibition or permission, optionally with constraints.
User Note(s)	<p>Regional laws applying to an asset may prohibit, constrain, or allow data mining for certain purposes (such as search indexing or research), and may overrule the value selected for this property. Similarly, the absence of a prohibition does not indicate that the asset owner grants permission for data mining or any other use of an asset.</p> <p>The prohibition “Prohibited except for search engine indexing” only permits data mining by search engines available to the public to identify the URL for an asset and its associated data (for the purpose of assisting the public in navigating to the URL for the asset), and prohibits all other uses, such as AI/ML training.</p> <p>The PLUS <i>Other Constraints</i> property is human-readable. The IPTC</p>

Cardinality	0..1
Controlled Vocabulary	<ul style="list-style-type: none">• http://ns.useplus.org/ldf/vocab/DMI-UNSPECIFIED (Unspecified - no prohibition defined)• http://ns.useplus.org/ldf/vocab/DMI-ALLOWED (Allowed)• http://ns.useplus.org/ldf/vocab/DMI-PROHIBITED-AI ML TRAINING (Prohibited for AI/ML training)• http://ns.useplus.org/ldf/vocab/DMI-PROHIBITED-GEN AI ML TRAINING (Prohibited for Generative AI/ML training)• http://ns.useplus.org/ldf/vocab/DMI-PROHIBITED-EXCEPT SEARCH ENGINE INDEXING (Prohibited except for search engine indexing)• http://ns.useplus.org/ldf/vocab/DMI-PROHIBITED (Prohibited)• http://ns.useplus.org/ldf/vocab/DMI-PROHIBITED-SEE CONSTRAINT (Prohibited, see Other Constraints property)• http://ns.useplus.org/ldf/vocab/DMI-PROHIBITED-SEE EMBEDDED RIGHTS EXPR (Prohibited, see Embedded Encoded Rights Expression property)• http://ns.useplus.org/ldf/vocab/DMI-PROHIBITED-SEE LINKED RIGHTS EXPR (Prohibited, see Linked Encoded Rights Expression property)



CAWG Training and Data Mining Assertion

- Started as a C2PA assertion, removed from C2PA spec for version 2.0
- Now maintained by CAWG
- Version 1.1 released in May 2025

The screenshot shows the DIF Creator Assertions Working Group website. The main heading is "Training and Data Mining Assertion". Below the heading, there is a paragraph explaining that the C2PA technical specification allows actors in a workflow to make cryptographically signed assertions about the produced C2PA asset. Another paragraph states that the training and data mining assertion enables a human actor to provide a C2PA Manifest Consumer information about whether an asset with C2PA metadata may be used as part of a data mining or AI/ML training workflow. On the left side, there is a sidebar with links to "Endorsement Assertion", "Identity Assertion", "Metadata Assertion", "Training and Data Mining Assertion", and "Meeting Notes". On the right side, there is a "Contents" section with links to "1. Introduction", "1.1. Scope", "2. Normative references", "3. Assertion definition", "3.1. Overview", "3.2. Schema and example", and "Appendix A: Version history".

```
{
  "entries": {
    "cawg.ai_training": {
      "use": "allowed"
    },
    "cawg.ai_generative_training": {
      "use": "notAllowed"
    },
    "cawg.data_mining": {
      "use": "constrained",
      "constraint_info": "may only be mined on days whose names end in 'y'"
    }
  }
}
```



Firewall blocking of AI bots

Cloud services offer bot protection services as part of their Web Application Firewall systems:

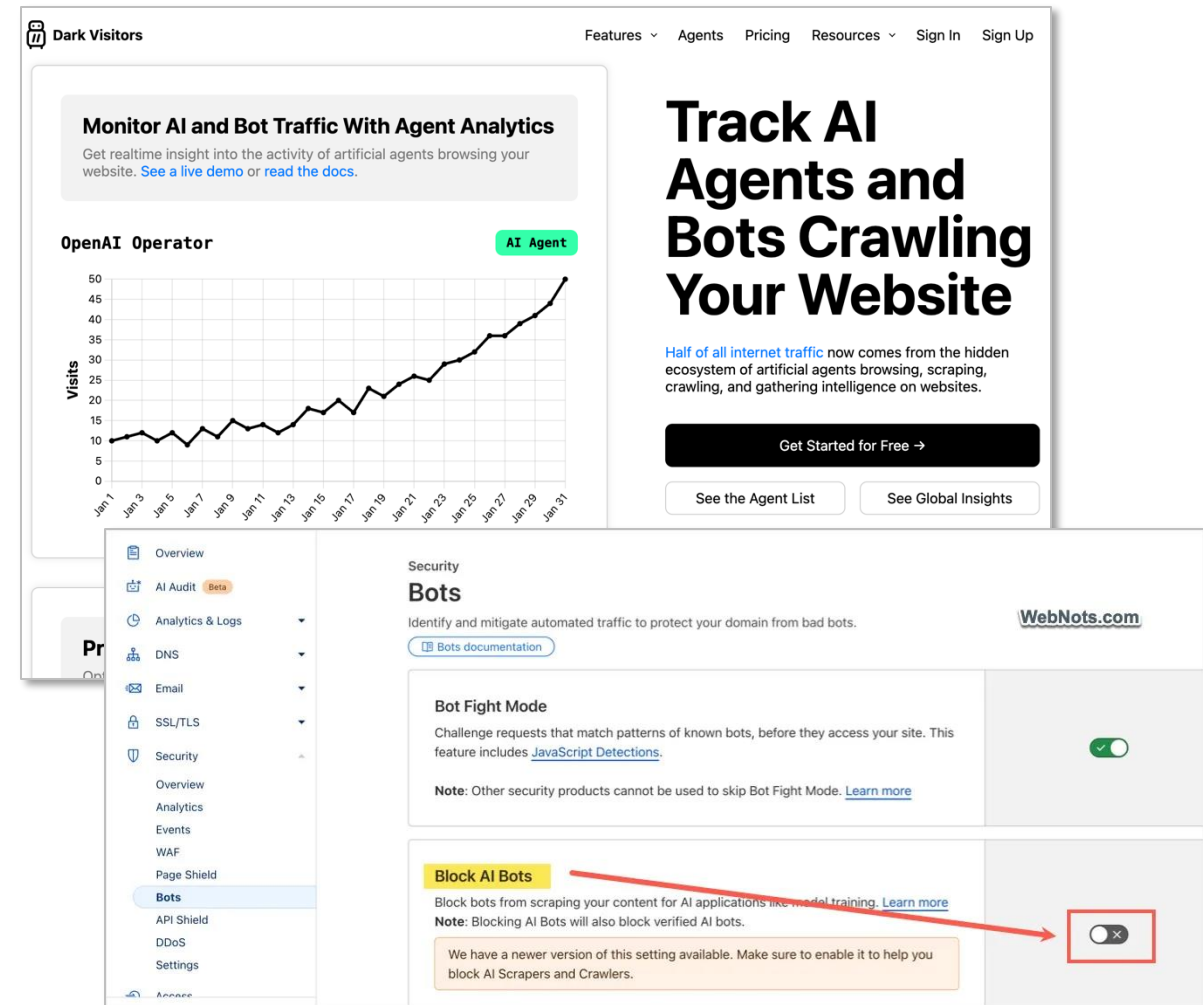
- [AWS WAF Bot Control](#)
- [Google Cloud Armor bot management](#)
- [Azure WAF Bot Protection](#)

The image displays three overlapping screenshots of documentation for bot protection services from major cloud providers:

- AWS WAF Bot Control:** The top-left screenshot shows the AWS documentation page for Bot Control, featuring a sidebar with navigation links and a main content area with a search bar and a 'Focus mode' toggle.
- Google Cloud Armor bot management overview:** The middle-right screenshot shows the Google Cloud documentation page for Cloud Armor, highlighting its bot management capabilities and providing a 'Send feedback' button.
- Azure Web Application Firewall on Azure Application Gateway bot protection overview:** The bottom-right screenshot shows the Azure documentation page for WAF bot protection, including a 'Note' section that specifies the Bot Protection Ruleset is only supported in the Azure public cloud, Azure China, and Azure US Government, and is not supported in air-gapped clouds.

Commercial tools to block AI bots

- DarkVisitors.com: “Automated robots.txt service”
 - they update your robots.txt for you when new bots appear
 - They also publish a [list of bot user-agent IDs](#) categorised into many varieties
- Cloudflare’s “Bot Fight Mode”
 - Option to block AI bots in the admin interface
 - Blocks at HTTP level, bots never see your site – they have no choice (unless [they use devious means](#))
 - Enabled by default since July 2025

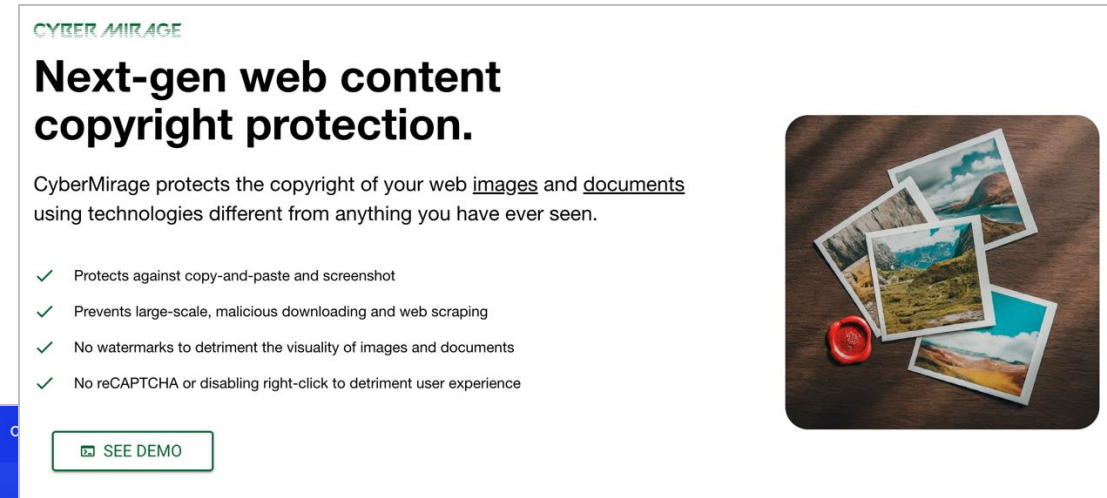


CloudFlare screenshot from <https://www.webnots.com/>

Other commercial tools that can help

Some tools obfuscate images to stop them from being downloaded by bots (or anyone else), e.g. by turning images into video files or using JavaScript to manipulate images so they are visible to humans but not to bots.

- [CyberMirage](#)
- [Smartframe](#)



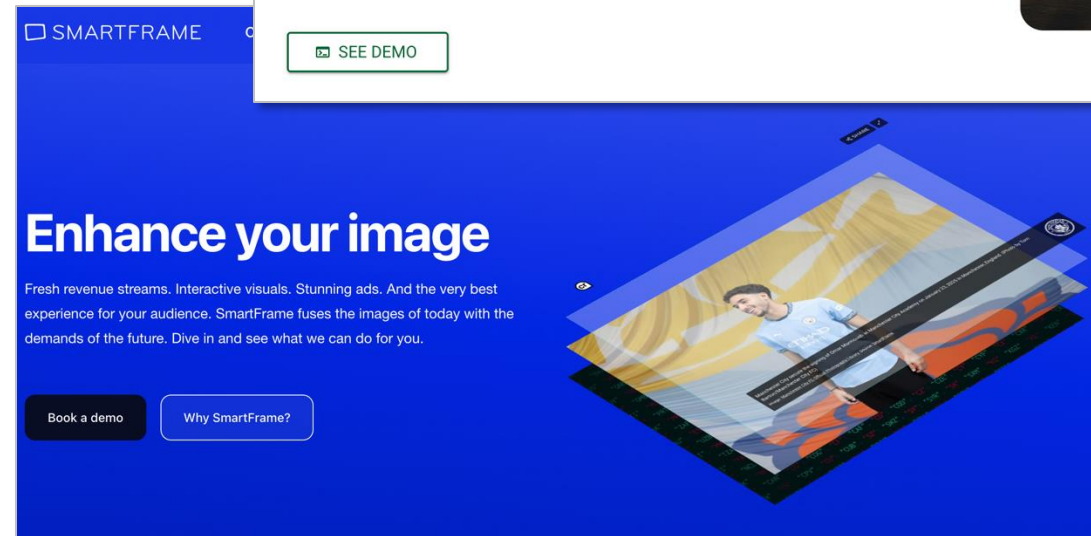

CYBER MIRAGE

Next-gen web content copyright protection.

CyberMirage protects the copyright of your web images and documents using technologies different from anything you have ever seen.

- ✓ Protects against copy-and-paste and screenshot
- ✓ Prevents large-scale, malicious downloading and web scraping
- ✓ No watermarks to detriment the visibility of images and documents
- ✓ No reCAPTCHA or disabling right-click to detriment user experience

[SEE DEMO](#)

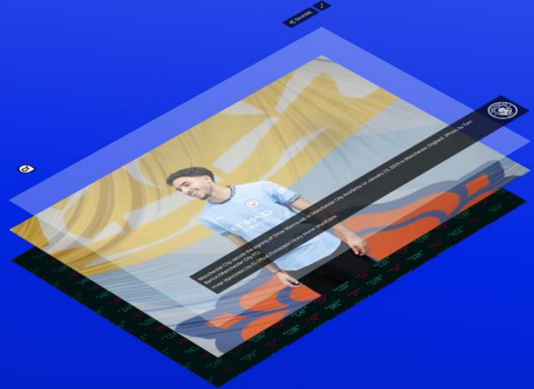


SMARTFRAME

Enhance your image

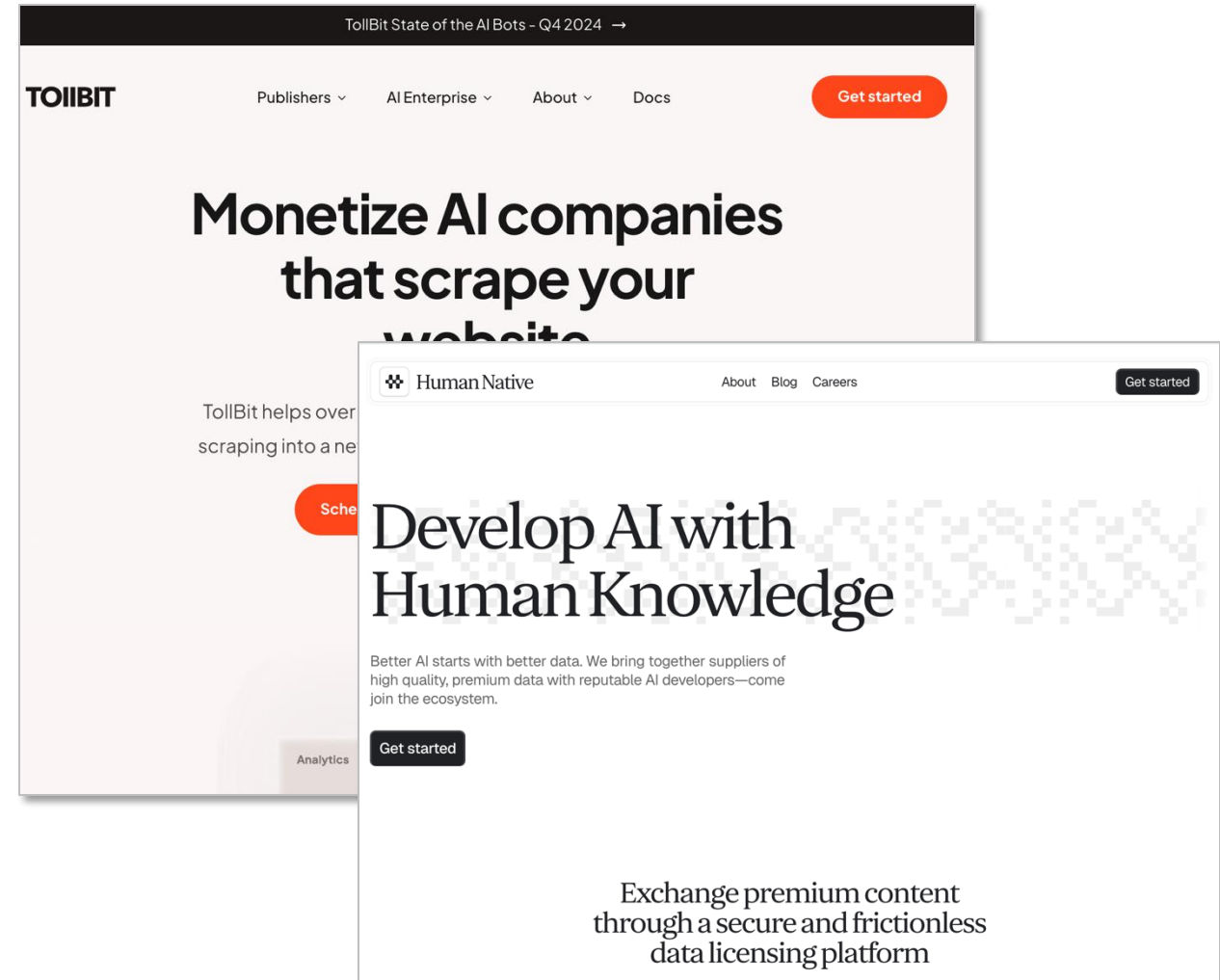
Fresh revenue streams. Interactive visuals. Stunning ads. And the very best experience for your audience. SmartFrame fuses the images of today with the demands of the future. Dive in and see what we can do for you.

[Book a demo](#) [Why SmartFrame?](#)



AI licensing systems / services

- Redirect all crawlers to a site/API that allows AI engines to fairly licence your data
- Works for crawling / training and also searching/RAG/agentic AI
 - Some AI providers only license content for search and agent purposes
- Vendors:
 - [TollBit](#)
 - [Human Native AI](#)
 - [ProRata.ai](#)





Summary of options available today (1/2)

TDM Reservation Protocol

- Works for any type of content
- Can express opt-in, opt-out, complex permissions
- Not well implemented

IPTC / PLUS Data Mining property

- Works for images and video via embedded metadata
- Can be carried along with the asset eg for wire images
- Not yet widely implemented yet

C2PA CAWG data mining assertion

- Works for images and video via custom C2PA assertion
- Requires content to be C2PA-signed
- Not yet widely implemented



Summary of options available today (2/2)

robots.txt

- Disallow option must be specified for each bot individually
- “Allow” or “disallow” based on URLs (including wild cards)
- Widely implemented

Manual forms from AI providers

- Some AI providers have their own forms to allow content providers to opt out per domain or per URL
- Not scalable

Firewalls

- Huge effort to maintain database of IP addresses and bot user agents
- Most effective – works even if the AI crawlers ignore robots.txt etc



IPTC Best Practices guidelines

- Inspired by [guidance created by the International Association of Scientific, Technical & Medical Publishers](#), we created a Best Practices guide for the news media industry
- It could also apply to image libraries and other publishers
- [Available from the IPTC website](#)

IPTC Generative AI Opt-Out Best Practice Recommendations

Version 1.0, 28 May 2025



In this document, we lay out a series of best practices that content publishers can follow to express the fact that they reserve data-mining rights on their copyrighted content. All of these techniques use currently available technologies¹.

We are advocating for more of these techniques to be explicitly acknowledged by law, and have submitted responses [to the European Union](#), [to the UK government](#) and [to the Internet Engineering Task Force \(IETF\)](#) on this subject.

In addition, we are actively working on future technical standards that may be used to express publisher rights and requirements to AI providers and data crawlers in other effective and scalable ways. But until those standards are published and adopted, we have created this guidance document to show how current technologies can be used to reserve the rights of content creators.

Summary of Recommendations

No.	Category	Recommendation
1	Non-technical	Display a plain language, visible rights reservation declaration for all copyrighted content
2	HTML, Image metadata	Display a rights reservation declaration in metadata tags on copyrighted content
3	Web infrastructure	Use Internet firewalls to block AI crawler bots from accessing your content



IPTC Best Practices Recommendations: 1-6

No.	Category	Recommendation
1	Non-technical	Display a plain language, visible rights reservation declaration for all copyrighted content
2	HTML, Image metadata	Display a rights reservation declaration in metadata tags on copyrighted content
3	Web infrastructure	Use Internet firewalls to block AI crawler bots from accessing your content
4	Robots Exclusion Protocol	Instruct AI crawler bots using their user agent IDs in your robots.txt file
5	TDMRep	Implement a site-wide tdmrep.json file instructing bots which areas of the site can be used for Generative AI training
6	Trust.txt	Use the trust.txt “datatrainingallowed” parameter to declare site-wide data mining restrictions or permissions



IPTC Best Practices Recommendations: 7-12

No.	Category	Recommendation
7	Image metadata	Use the IPTC Photo Metadata Data Mining property on images and video files
8	Image metadata / C2PA	Use the CAWG Training and Data Mining Assertion in C2PA-signed images and video files
9	HTML / Robots Exclusion Protocol	Use in-page metadata to declare whether robots can archive or cache page content
10	HTML / TDMRep	Use TDMRep HTML meta tags where appropriate to implement TDM declarations on a per-page basis
11	HTTP / Robots Exclusion Protocol	Send Robots Exclusion Protocol directives in HTTP headers where appropriate
12	HTTP / TDMRep	Use TDMRep HTTP headers where appropriate to implement TDM declarations on a per-URL basis



Ongoing work: IETF AI Preferences group

- The Internet Engineering Task Force (IETF) define internet standards (“RFCs”) for the plumbing of the internet – email, HTTP, HTTPS...
- This WG is looking at evolving current internet standards (including robots.txt) to account for AI crawlers
- Meetings in Washington DC, Brussels, London. Zurich soon.
- Attendees from Google, Mozilla, OpenAI, Microsoft, Creative Commons, Anthropic, Financial Times, IPTC + many more

The screenshot displays two overlapping web pages from the IETF Datatracker. The top page, titled "AI Preferences (aipref)", shows a summary of the working group with fields for Name, Acronym, Area, State, Charter, and Document dependencies. The bottom page, titled "Charter for Working Group", provides a detailed overview of the group's mission, its goals, and the topics it will and will not address.

AI Preferences (aipref)

WG	Name	AI Preferences
	Acronym	aipref
	Area	Web and Internet Transport (wit)
	State	Active
	Charter	charter-ietf-aipref-01 Approved
	Document dependencies	Show

Charter for Working Group

The AI Preferences Working Group will standardize building blocks that allow for the expression of preferences about how content is collected and processed for Artificial Intelligence (AI) model development, deployment, and use.

There are many ways that preferences regarding content might be expressed. The Working Group will focus on attaching preferences to content either by including preferences in content metadata or by signaling preferences using the protocol that delivers content.

The Working Group will deliver:

- A standard track document covering vocabulary for expressing AI-related preferences, independent of how those preferences are associated with content.
- Standard track document(s) describing means of attaching or associating those preferences with content in IETF-defined protocols and formats, including but not limited to using Well-Known URIs ([RFC 8615](#)) such as the Robots Exclusion Protocol ([RFC 9309](#)), and HTTP response header fields.
- A standard method for reconciling multiple expressions of preferences.

The working group is expected to liaise as appropriate with:

- International Press Telecommunications Council (IPTC) and the PLUS Coalition, regarding attachment in formats controlled by these bodies
- WHATWG and/or W3C, regarding attachment in HTML and other formats controlled by these bodies
- Other bodies responsible for content formats, as appropriate

Liaisons are intended to aid the incorporation into protocols, mechanisms, frameworks and content formats developed and controlled by external bodies.

The following topics are out of scope for this charter:

- Technical enforcement of preferences
- Application layer protocols for authenticating or authorizing clients and/or crawlers



AIPrefs WG working draft documents

- “A Vocabulary For Expressing AI Usage Preferences”
 - Defining hierarchical levels of AI preferences, without reference to *how* they would be associated with content
- “Associating AI Usage Preferences with Content in HTTP”
 - Defines how the above preferences would be “attached” to content

draft-ietf-aipref-attach-03

AI Preferences
Internet-Draft
Updates: 9309 (if approved)
Intended status: Standards Track
Expires: 9 March 2026

G. Illyes
Google
M. Thomson
Mozilla
5 September 2025

Associating AI Usage Preferences with Content in HTTP
draft-ietf-aipref-attach-03

Abstract

Content creators and other stakeholders define preferences about how their content is processed by automated systems. This document defines how these preferences are associated with content as part of the acquisition of content.

This document updates RFC 9309 to add AI usage preferences.

About This Document

This note is to be removed before publishing as an RFC.

The latest revision of this draft can be found at <https://ietf-wg-aipref.github.io/drafts/draft-ietf-aipref-attach.html>. Status information for this document may be found at <https://datatracker.ietf.org/doc/draft-ietf-aipref-attach/>.

Discussion of this document takes place on the AI Preferences Working Group mailing list (<mailto:ai-control@ietf.org>), which is archived at <https://mailarchive.ietf.org/arch/browse/ai-control/>. Subscribe at <https://www.ietf.org/mailman/listinfo/ai-control/>.

draft-ietf-aipref-vocab-03

AI Preferences
Internet-Draft
Intended status: Standards Track
Expires: 9 March 2026

P. Keller
Open Future
M. Thomson, Ed.
Mozilla
5 September 2025

A Vocabulary For Expressing AI Usage Preferences
draft-ietf-aipref-vocab-03

Abstract

This document defines a vocabulary for expressing preferences regarding how digital assets are used by automated processing systems. This vocabulary allows for the declaration of restrictions or permissions for use of digital assets by such systems.

About This Document

This note is to be removed before publishing as an RFC.

The latest revision of this draft can be found at <https://ietf-wg-aipref.github.io/drafts/draft-ietf-aipref-vocab.html>. Status information for this document may be found at <https://datatracker.ietf.org/doc/draft-ietf-aipref-vocab/>.

Discussion of this document takes place on the AI Preferences Working Group mailing list (<mailto:ai-control@ietf.org>), which is archived at <https://mailarchive.ietf.org/arch/browse/ai-control/>. Subscribe at <https://www.ietf.org/mailman/listinfo/ai-control/>.

AIPrefs current draft: vocabulary

- **Automated Processing**

The act of using automated processing on one or more assets to analyze text and data in order to generate information which includes but is not limited to patterns, trends and correlations.

- **AI Training**

The act of training machine learning models or artificial intelligence (AI).

- **Generative AI Training**

The act of training general purpose AI models that have the capacity to generate text, images or other forms of synthetic content, or the act of training more specialized AI models that have the purpose of generating text, images or other forms of synthetic content.

- **Search**

Using one or more assets in a search application that directs users to the location from which the assets were retrieved.

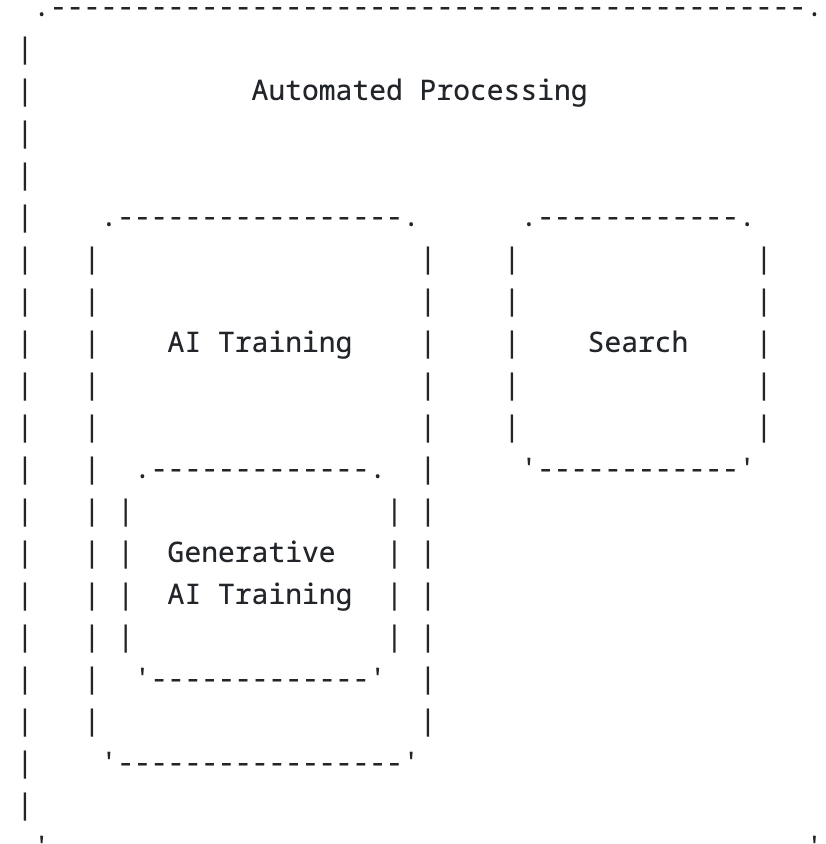


Figure 1: Relationship Between Categories of Use



AIPrefs Vocabulary: how to express, when it applies

- “y” for allow, “n” for disallow
- Example:
 - `train-ai=y`, `train-genai=n`
- After processing a statement of preferences the recipient associates each category of use one of three preference values: "allowed", "disallowed", or "unknown". In the absence of a statement of preference, all usage categories are assigned a preference value of "unknown".
- “Contractual agreements or other specific arrangements might override statements of preference.”
- When combining preferences, the most restrictive preference applies.
- Examples / use cases for when preferences might be ignored by the recipient: Accessibility, cultural heritage, scholarly research, detecting harmful content
- “the consequences of ignoring preferences could vary depending upon how a given legal jurisdiction recognizes preferences”



AIPrefs current draft: attachments

“

The automated consumption of content by crawlers and other machines has increased significantly in recent years. This is partly due to the training of machine-learning models. Content creators and other stakeholders, such as distributors, might wish to express a preference regarding the types of usage they consider acceptable. Entities that might use that content need those preferences to be stated in a way that is easily consumed by an automated system.

This document describes two mechanisms for associating preferences with content:

- A Content-Usage header field for HTTP [HTTP]; see Section 2.
- A Content-Usage directive for the Robots Exclusion Protocol (colloquially known as "robots.txt") [ROBOTS]; see Section 3.

For automated systems that use HTTP to gather content, these allow for the automated gathering of preferences in the same way that content is obtained.

”



AIPrefs attachments - examples

robots.txt file:

```
User-Agent: *  
Allow: /  
Content-Usage: train-ai=n  
  
User-Agent: *  
Allow: /  
Disallow: /never/  
Content-Usage: train-ai=n  
Content-Usage: /ai-ok/ train-ai=y
```

HTTP sever response:

```
HTTP/1.1 200 OK  
Date: Wed, 23 Apr 2025 04:48:02  
GMT  
Content-Type: text/plain  
Content-Usage: train-ai=n  
  
This is some content.
```



AIPrefs attachments – embedded metadata

1.3.1. Embedded Preferences

Embedding preferences is expected to be an effective means of associating preferences with content, because it ensures that metadata is always associated with content. This document, however, does not define any specific means of embedding preferences in content.

[...]

1.3.2. Registry-Based Preferences

A preferences registry is a database that stores usage preference statements associated with both content identifiers and a means of identifying the declaring party. Registry-based approaches might be applicable in certain contexts, particularly where embedding is impractical or unavailable. Additionally, a registry might enable persistent association of preferences across distribution channels.



AIPrefs WG – outstanding issues

- All issues are [visible in GitHub](#)
- Main issues raised:
 - Definitions are too broad. Does “automated processing” include printing? Saving to PDF?
 - Idea of focusing on “processing” vs “use” of the content, eg RAG and “grounding”
 - Or to focus on display of the crawled content, eg “noindex” “nosnippet” “exact match” “image preview” etc
 - Proposal of “substitutive use” as another vocab element (under Automated Processing)
 - Should “search” really be a subset of “Automated Processing”?
 - Does it make sense not to define default values?

Overall approach of the Vocabulary 0 / 3 triage vocabulary wglc
#159 · mnot opened 8 hours ago

Bots Collect Data for Multiple Purposes triage vocabulary wglc
#158 · mnot opened yesterday

AI Training and Generative AI Training are Too Broad triage vocabulary
wglc
#157 · mnot opened yesterday

Search's Parent triage wglc
#156 · mnot opened yesterday

Automated Processing is too broad triage vocabulary wglc
#155 · mnot opened yesterday

Defaults triage vocabulary wglc
#154 · mnot opened yesterday

Difference between "unknown" and "allowed" triage vocabulary wglc
#153 · mnot opened yesterday

Use of 'Machine Learning' triage vocabulary wglc
#152 · mnot opened yesterday

Definition of AI triage vocabulary wglc
#151 · mnot opened yesterday

Proposal to add a Substitutive Use Category to the Vocab Doc triage
vocabulary wglc
#150 · BradSilver99 opened 2 days ago



AIPrefs WG – next steps

- Drafts are in “Working Group Last Call” state
 - this is “to elicit final comments” within the working group
 - This period lasts until 23rd September
 - Outstanding issues will be discussed at the next group meeting in Zurich (and online) 30 Sep – 2 Oct
- After the WG achieves a level of consensus, the document is submitted to the Internet Engineering Steering Group for discussion and approval (and probably a round of changes)
- Note that there is no consensus right now – the draft will probably change before the final version is released



Thanks!

Brendan Quinn
mdirector@iptc.org