IPTC Generative AI Opt-Out Best Practice Recommendations

Version 2.0, 30 March 2026

Introduction

The IPTC has fielded many requests from members and others in the digital media publishing industry. Publishers want to be able to express the fact that they do not wish their content to be used as training data for Generative AI models without an agreed licence. However, the currently available technologies make it very difficult to express these intentions in a way that can easily be read by web crawling tools used by AI companies.

In this document[1], 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. Following these guidelines, publishers give themselves the best chance of instructing web crawlers of their rights.

All of these techniques use technologies that are currently available and free to use.

Regulatory Context

Governments around the world are looking at laws governing AI engines' use of human-created content for the purposes of training and running Large Language Models (LLMs) and other AI tools. Currently there is a patchwork of rules, both in-use and proposed, with varying approaches to default use, from opt-in by default to opt-out by default.

The IPTC advocates for a usable, practical AI usage protocol to be explicitly acknowledged by law, and we have submitted responses to the European Union, to the EU AI Office, 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[2]. To help publishers while we wait for such standards to be published and adopted, we have created this guidance document to show how current technologies can be used to reserve the rights of content creators.

---

[1] This document is inspired by similar guidance defined by the International Association of Scientific, Technical & Medical Publishers (STM). We thank STM for their work.

[2] The AIPrefs Working Group within the IETF has been working on a potential update to robots.txt that will include expressions of AI usage preferences. Unfortunately this work has been delayed due to a lack of consensus and is not expected to deliver a final RFC until August 2026. In the meantime, we are recommending other methods such as those recommended in this document.

Summary of Recommendations

Web Content Recommendations	
1	Display a plain-language, visible rights reservation declaration for all copyrighted content
HTML metadata Recommendations	
2	Display a rights reservation declaration in metadata tags on copyrighted web pages
3	Use in-page metadata to declare whether robots can archive or cache page content
4	Use TDMRep HTML meta tags where appropriate to implement TDM declarations on a per-page basis
Robots.txt and related recommendations	
5	Instruct AI crawler bots using their user agent IDs in your robots.txt file
6	Add Cloudflare Content Signals directives to your site's robots.txt file
7	Implement a site-wide tdmrep.json file instructing bots which areas of the site can be used for Generative AI training
8	Add a generic Really Simple Licensing (RSL) license to your site's robots.txt file
9	Use the trust.txt "datatrainingallowed" parameter to declare site-wide data mining restrictions or permissions
Network-level recommendations	
10	Use Internet firewalls to block AI crawler bots from accessing your content
Media file metadata recommendations	
11	Use the IPTC Photo Metadata Data Mining property on images and video files
12	Use the CAWG Training and Data Mining Assertion in C2PA-signed images and video files
13	Use TDMRep embedded metadata in epub and PDF files

Web Content Recommendations

1. Display a plain-language, visible rights reservation declaration for all copyrighted content

Ensure that there can be no misinterpretation of your intent to reserve your rights through a plain-language, visible rights reservation sentence such as

> *Copyright © YEAR ENTITY. All rights, including for text and data mining, AI training, and similar technologies, are reserved.*

or

> *Users of this website are prohibited from using any data mining, robots or similar data gathering or extraction methods.*

This text should be consistently displayed: for example, within website terms of service, at the bottom of each web page on your website or alongside copyright works.

HTML metadata Recommendations

2. Display a rights reservation declaration in metadata tags on copyrighted web pages

Add one of the above sentences in a dedicated metadata field (if available) that is part of the copyrighted content item.

Examples include:

- schema.org metadata links on web pages:

```
<link rel="schema.dcterms" href="http://purl.org/dc/terms/">
<meta name="dcterms.rights" content="copyright statement">
```

- schema.org embedded JSON-LD:

```
<script type="application/ld+json">
{ "@context":"https://schema.org",
  "@type":"NewsArticle",
  "author":"Brendan Quinn",
  "publisher":{"@type":"Organization",
               "name":"International Press Telecommunications
Council",
               "Logo":{ "@type":"ImageObject",
                        "url":"/img/logo_dpa.svg"}},
  "copyrightNotice": "copyright statement"
  ... etc ...
}
</script>
```

3. Use in-page metadata to declare whether robots can archive or cache page content

The “noindex”, “noarchive” and “nocache” HTML meta tag directives, not explicitly defined in the Robots Exclusion Protocol but widely implemented, can be used to influence what crawling robots do with HTML-based content.

- “noindex” is interpreted as directing bots to never use the content at all. Note that this may also include search engine robots.

- The “nocache” directive is generally interpreted as allowing bots to index high-level headings (including allowing search-engine crawler bots) but not to index the details of content.

- “noarchive” is no longer used by Google search, but is still used by other search engines.. Microsoft’s robots tag documentation for Bing and Copilot, and a related blog post (dating from when Copilot was called Bing Chat) explain that content with a “noarchive” tag will be ignored by AI but retained for search purposes.

- The additional “noai” and “noimageai” directives, defined by DeviantArt in 2022 and used by other sites including fab.com, can also be used to indicate data mining reservation for all content and for image content respectively.

These directives are additive, and they can be declared either by using multiple HTML meta tags or by separating the values with commas.

Note that according to some AI providers such as Microsoft Copilot, content that is described as both “nocache” and “noarchive” is interpreted as only “nocache”, meaning the “noarchive” directive is ignored. This may not be in line with the publisher’s intent. Therefore we recommend against using both “nocache” and “noarchive” at the same time.

Google supports other non-standard rules to guide search results and links. These include “nosnippet” and “max-snippet” which govern whether search and AI results should include summaries or extracts of content, and how many characters should be displayed.

Google’s documentation for “nosnippet” says “This applies to all forms of search results (at Google: web search, Google Images, Discover, AI Overviews, AI Mode) and will also prevent the content from being used as a direct input for AI Overviews and AI Mode.”[3]

To block content from generative AI crawlers but allow search engine crawlers, we recommend using the “noarchive” directive, plus “noai” and “noimageai”. “nosnippet” may also be added if publishers wish to require only verbatim extracts and direct links:

```
<meta name="robots"
content="noarchive,nosnippet,noai,noimageai">
```

[3] This idea can also be used to exclude specific sections of a page’s content from appearing in Google search result snippets. To do this, use the `data-nosnippet` HTML attribute.

4. Use TDMRep HTML meta tags where appropriate to implement TDM declarations on a per-page basis

One way to implement the W3C TDMRep protocol is to include directives in meta tags in HTML pages. The following HTML tag should be included in the <head> section of relevant web pages:

```
<meta name="tdm-reservation" content="1">
```

We only recommend this mechanism if fine-grained control is needed. It is better to use a site-wide tdmrep.json file as explained in Recommendation 7 if possible.

Robots.txt and related recommendations

This set of recommendations assumes that you have sufficient access to your web server to create or edit text files at the top level of your site's domain, such as example.com/textfile.txt.

5. Instruct AI crawler bots using their user agent IDs in your robots.txt file

Robots.txt is the most common mechanism for instructing web crawlers to tailor the way that they crawl a site (although they are not always respected, see the note in Recommendation 3 above). Currently there is no universal mechanism to add a "generative AI opt-out" to robots.txt. The only way to allow general Internet robots (such as search engine crawlers) but block AI engines is to disallow them one by one, using their "User Agent" names.

For example, to tell the Perplexity bot not to crawl any resources on your web site, add the following text to the yourdomain.com/robots.txt file:

```
User-agent: PerplexityBot
Disallow: /
```

Maintaining a list of bot user agent strings to be used in robots.txt files is not a simple task. We are considering maintaining a list of such services for the benefit of IPTC member organisations. In the meantime, looking at other publishers' robots.txt files can give a good indication of which services can be disallowed using robots.txt.

A non-exhaustive list is provided in Appendix A.

We remind site owners that robots.txt is only a recommendation to site crawlers, and it does not guarantee that it will be followed by AI providers in any jurisdiction. The only way to be sure that bots will not index your site's content is to block them at the HTTP level, as described in Recommendation 10 below (but also note the potential disadvantages).

6. Add Cloudflare Content Signals directives to your site's robots.txt file

Content Signals is a specification introduced in late 2025 by Cloudflare, based on early drafts of the IETF AI Preferences Working Group. While it has not been ratified by a standards body, it is a simple and effective proposal and we see no downside in publishers using it. Content Signals is currently the only way to declare opt-out information in a robots.txt file in a generic way (i.e. without having to name each individual AI crawler user agent).

To use Content Signals to allow search but block all AI access from crawlers, simply add the following text to your robots.txt file:

```
User-Agent: *
Content-Signal: ai-train=no, search=yes, ai-input=no
Allow: /
```

7. Implement a site-wide tdmrep.json file instructing bots which areas of the site can be used for Generative AI training

Use the TDMRep Protocol (the output of a W3C Community Group) to instruct bots and processors that text and data mining rights for all content on a web server are reserved, by creating a file on your web server at the URL `yourdomain.com/.well-known/tdmrep.json`. This file contains an array of locations, optionally including wild-cards, and a "tdm-reservation" key with a value of 0 (unreserved) or 1 (reserved).

The simplest tdmrep.json file, reserving data-mining rights across an entire site, is as follows:

```
[
  {
    "location": "/",
    "tdm-reservation": 1
  }
]
```

We note that it is possible to set a more detailed text and data mining policy using the `tdm-policy` directive defined in the TDMRep specification. However to our knowledge this is not yet implemented by any crawler bots, so we do not currently recommend using the `tdm-policy` property. The `tdm-reservation` property should be sufficient.

8. Add a generic Really Simple Licensing (RSL) license to your site's robots.txt file

Really Simple Licensing (RSL) is another specification that was released in late 2025. It is not aligned with a standards body at this time.

RSL is capable of expressing more complex rights and licensing information, but to use it for simple AI opt-out, you can add the following to your site's robots.txt file: either within a User Agent group, or outside of a group, separated from other groups by an empty line.

```
License: https://example.com/license.xml
```

The "license.xml" file should then contain the following to opt out from AI training but opt in to search indexing:

```
<rsl xmlns="https://rslstandard.org/rsl">
  <content url="/">
    <license>
      <permits type="usage">search</permits>
      <prohibits type="usage">ai-train ai-input
ai-index</prohibits>
    </license>
    <terms>https://example.com/legal/data-terms.html</terms>
  </content>
</rsl>
```

If your site's needs are more nuanced, you may wish to consult the Really Simple Licensing specification for more details and examples, e.g. to allow AI indexing but only for non-commercial or academic purposes.

9. Use the trust.txt "datatrainingallowed" parameter to declare site-wide data mining restrictions or permissions

The trust.txt specification allows a publisher to declare a single, site-wide data mining reservation with a simple command: `datatrainingallowed=no` (or alternatively `datatrainingallowed=yes` if a site wishes to allow training on its content).

If you already maintain a trust.txt file, we recommend that you add the property to ensure consistency across all rights-reservation mechanisms.

Network-level recommendations

10. Use Internet firewalls to block AI crawler bots from accessing your content

It's important to remember that at present, all rights declaration protocols such as robots.txt, TDMRep and HTML meta tags do not guarantee that crawlers will comply with publisher requirements. At present, no government has required web crawlers to follow any specific machine-readable standard.[4]

---

[4] One government that has made steps towards this is the European Union with its AI Act. The Code of Practice of the EU AI Act states: "for the purpose of text and data mining as defined in Article 2(2) of Directive (EU) 2019/790 and the training of their general-purpose

As a result, many publishers report to us that some AI bots simply ignore any robots.txt declarations and crawl copyrighted content anyway.

Therefore, a technical solution could be to block crawlers at the HTTP level, so that they never see copyrighted content.

Several sites and databases[5] keep records of the User-Agent strings and IP addresses of the major AI crawler bots. Using these records, publishers can construct firewall configuration files that block known AI provider bots before they can access copyrighted content. Publishers could also block third-party crawlers that provide crawled content to AI providers, such as Common Crawl and LAION.

Infrastructure solutions such as Amazon Web Services Web Application Firewall Bot Control or Google Cloud Armor Bot Management can be used to set rules concerning which bots are allowed to access publisher content.

Licensed content can be shared with selected AI providers via exceptions to these firewall rules.

Disadvantages to blocking specific crawler-bots are:

- The approach places additional cost and burden on publishers to monitor AI crawler bots and to adjust their settings on a dynamic basis; and
- SEO performance might be negatively impacted, e.g. if search engine crawler bots are also inadvertently blocked or search engine ranking systems take into account whether crawler-bots are blocked.

CloudFlare can be configured to block all AI crawler access by default.

Note that we recommend all firewalls should respond to requests for /robots.txt files. In particular, firewalls should not respond with a 4xx response status to requests for /robots.txt files, because RFC 9309 says that "server status code indicates that the robots.txt file is unavailable to the crawler, then the crawler MAY access any resources on the server."

AI models, Signatories commit [...] to employ web-crawlers that read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt), as specified in the Internet Engineering Task Force (IETF) Request for Comments No. 9309, and any subsequent version of this Protocol for which the IETF demonstrates that it is technically feasible and implementable by AI providers and content providers, including rightsholders." But no such "subsequent version" of RFC9309 has yet been published and it is not clear whether the work of the IETF AIPrefs Working Group will meet the definition of a "subsequent version of RFC9309". So our point remains valid.

[5] One such service is DarkVisitors.com, which maintains a database of web crawlers including User Agent and IP known IP address blocks. We welcome suggestions of similar services that we could recommend.

Media file metadata recommendations

11. Use the IPTC Photo Metadata Data Mining property on images and video files

The IPTC Photo Metadata Working Group and IPTC Video Metadata Working Group have adopted a metadata property defined by the PLUS Coalition. This property, called simply "Data Mining", allows rights holders to declare their wishes for if and how their content can be used for data mining, embedded directly into the metadata packet within an image or video file.[6]

The set of terms that can be declared as values of the Data Mining property includes values such as blanket "Allowed" and "Prohibited" statements, but also variations such as "Prohibited for Generative AI/ML training" and "Prohibited except for search engine indexing". The full list of terms is available on the PLUS website.

There are several benefits of using embedded metadata to declare data mining reservations for media assets:

a. This mechanism allows fine-grained signalling at the individual asset level, which would be difficult and time-consuming to do with robots.txt or tdmrep.json techniques
b. The data mining reservation metadata is carried along with the asset when it is moved (as long as metadata is not stripped out from media files. We caution against metadata stripping but this unfortunately cannot be prevented by third-parties and commonly does occur.)
c. Third-party assets (such as images from news wires or picture agencies) may have different ownership rights and therefore different data mining reservation declarations. This mechanism allows for third-party assets to be treated differently without publishers having to do any additional work.

In addition, the Copyright Notice property in IPTC Photo Metadata and the IPTC Video Metadata Hub Copyright Notice property should be used to convey a human-readable version of copyright information.

12. Use the CAWG Training and Data Mining Assertion in C2PA-signed images and video files

The Creator Assertions Working Group (CAWG), a group incorporated under the Decentralised Identity Foundation which provides solutions compatible with the Coalition for Content Provenance and Authenticity (C2PA), publishes the Training and Data Mining

[6] The IPTC Photo Metadata Standard makes use of the XMP packet in media files, which has been used to embed metadata in image files for over 20 years and is supported by all major image-editing tools. This is different from the C2PA standard which also allows metadata to be embedded in signed files but is less well-adopted at present. Both can exist in the same file.

Assertion. This provides a mechanism for publishers to embed rights reservation information into C2PA-signed content such as images and video files.

The assertion could look like the following:

```
{
  "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'"
      }
}
```

13. Use TDMRep embedded metadata in epub and PDF files

Recent versions of the TDMRep spec outline a way that metadata can be included in epub and PDF files. While TDMRep doesn't yet support fine grained metadata, it is a way to tell crawlers that the content is not to be indexed for any text and data-mining purposes, which includes AI training.

For example, TDMRep metadata would be embedded in a PDF file using the XMP metadata section as follows:

```
<rdf:RDF
xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#>
<rdf:Description rdf:about=""
  ...
  xmlns:tdm="http://www.w3.org/ns/tdmrep/">
      <tdm:reservation>1</tdm:reservation>

<tdm:policy>https://publisher.com/policies/policy.json</tdm:p
olicy>
    </rdf:Description>
  </rdf:RDF>
```

Appendix A: Partial list of AI engine user agents

This list of user agents has been compiled from public sources, including the existing robots.txt files of several major publishers and content owners. We welcome suggestions for additions to the list, which we will add in future revisions of this document.

NOTE: Blocking some user agents risks your content not appearing in many parts of the web. For example, blocking "Googlebot" would stop your site from being indexed by Google for inclusion in search results.

User-agent: AliyunSecBot
User-agent: anthropic-ai
User-agent: Applebot-Extended
User-agent: archive.org_bot # Internet Archive / Wayback Machine
User-agent: AudigentAdBot
User-agent: AwarioRssBot
User-agent: AwarioSmartBot
User-agent: Amazonbot # AI Search crawler (Amazon)
User-agent: bingbot # Microsoft Bing search and Copilot
User-agent: BLEXBot
User-agent: Bytespider
User-agent: CCBot # Common Crawl indexer
User-agent: ChatGPT-User # "inference-time" requests from ChatGPT
User-agent: ClaudeBot
User-agent: Claude-SearchBot
User-agent: Claude-User
User-agent: Claude-Web
User-agent: cohere-ai
User-agent: cohere-training-data-crawler
User-agent: DataForSeoBot
User-agent: Diffbot
User-agent: DuckAssistBot
User-agent: EchoboxBot
User-agent: FacebookBot
User-agent: FriendlyCrawler
User-agent: Google-CloudVertexBot # Used for Vertex AI Bots (more info)
User-agent: Google-Extended # Gemini AI training (more info)
User-agent: GPTBot # OpenAI training data crawling
User-agent: ImagesiftBot
User-agent: Jetslide
User-agent: magpie-crawler
User-agent: Meta-ExternalAgent
User-agent: meta-externalagent
User-agent: Meta-ExternalFetcher
User-agent: meta-externalfetcher
User-agent: MyCentralAIScraperBot

User-agent: NewsNow
User-agent: news-please
User-agent: OAI-SearchBot # AI Search Crawler (OpenAI / ChatGPT)
User-agent: omgili
User-agent: omgilibot
User-agent: peer39_crawler
User-agent: PerplexityBot # AI Search Crawler (Perplexity)
User-agent: Perplexity-User
User-agent: Poseidon Research Crawler
User-agent: quillbot.com
User-agent: Quora-Bot
User-agent: Scrapy
User-agent: SeekrBot
User-agent: SeznamHomepageCrawler
User-agent: TaraGroup Intelligent Bot
User-agent: Timpibot
User-agent: TurnitinBot
User-agent: ViennaTinyBot
User-agent: YouBot # AI Search Crawler (You.com)