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 technologies[1].

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
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
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

[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.

No.	Category	Recommendation
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

Recommendations in Detail

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.

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

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 on web pages:

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

- The IPTC Photo Metadata Standard Copyright Notice property embedded in all copyrighted image files
- The IPTC Video Metadata Hub Copyright Notice property embedded in all copyrighted video files

3. 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. (The EU Copyright Directive does instruct AI engines to respect "machine readable" opt-out declarations, but without naming any specific protocols).

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 they can never see copyrighted content.

Several sites and databases[2] 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.

[2] 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.

4. 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.

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 3 above (but also note the potential disadvantages).

5. 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.

6. 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.

7. 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.[3]

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

[3] 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.

declarations. This mechanism allows for third-party assets to be treated differently without publishers having to do any additional work.

This should be in addition to the embedded Copyright Notice property described above in Recommendation 2.

8. 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 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'"
      }
}
```

9. 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 interpreted as directing bots not to index content at all.

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

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

Note that according to some AI providers such as Microsoft Bing Chat, 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.

To block content from generative AI crawlers but allow search engine crawlers, we recommend using the “noarchive” directive along with “noai” and “noimageai”:

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

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

Another way to implement TDMRep 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">
```

Again, 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.

11. Send Robots Exclusion Protocol directives in HTTP headers where appropriate

While not defined in RFC9309 or the robots.txt site, many crawlers respect the X-Robots-Tag header on an individual HTTP request.

The same guidance given in Recommendation 9 applies to this mechanism, so an HTTP response may look like the following:

```
X-Robots-Tag: noarchive,noai,noimageai
```

12. Use TDMRep HTTP headers where appropriate to implement TDM declarations on a per-URL basis

Yet another way to use the TDMRep declaration is to include it as an HTTP header on an individual HTTP request. It is possible to implement this recommendation using configuration files on Apache, nginx or other web servers.

```
HTTP/1.1 200 OK
Date: Wed, 17 Apr 2025 12:07:48 GMT
Content-type: text/html
tdm-reservation: 1
```

This can be combined with the X-Robots-Tag directive described in Recommendation 11. We only recommend these mechanisms if fine-grained signals are needed for individual assets.

About the International Press Telecommunications Council (IPTC)

The IPTC is the global technical standards body of the news media. Our mission is to simplify the distribution of information.

A non-profit organisation based in London with members from 22 countries around the world, the IPTC brings together technical representatives from the world's leading news agencies, publishers, broadcasters, industry vendors and consultants. Members join together at IPTC events and at regular Working Group meetings to share best practices and create technical standards for sharing media content.

Current focus areas of IPTC's work include content provenance and authenticity (including IPTC's Origin Verified News Publishers List, which is a "trust list" of publisher certificates that can be used to sign media content using C2PA technology), AI transparency, data mining opt-in and opt-out rules, and metadata for content accessibility. For more information on IPTC, contact Managing Director Brendan Quinn at mdirector@iptc.org or use the IPTC Contact form.