STM Association
Publisher Recommendations

STM Association Publisher Guidelines for Protecting Content from Unauthorized AI/TDM activity

Art. 4 EU 2019 Directive for Copyright in the Digital Single Market (DSM Directive) governs text and data mining (TDM) for commercial purposes. It uses a broad definition of TDM that can be interpreted in a way that is relevant in the context of AI technology.

This requires rights holders to express clearly what usages are (and are not possible with their content.)

Robust human- and machine-readable policy indicators are needed to ensure that commercial content collectors and processes recognize that content is protected, even when they have lawful access to it (via subscription or open access). This safeguards the commercial licensing market for STM publishers.

STM suggests publishers take steps to inform content collectors about the rights, policies, and rules that apply to text and data mining of their content. Publishers are encouraged to implement some or all recommendations, depending on their needs. Guidelines fall in four categories of action below. Please note that this guidance is offered by the STM Association on complimentary basis.

Human Readable Policies

Machine Readable Policies

Prevent Access for Prohibited Content Collectors

Provide Instructions for Permitted Content Collectors

STM

Human Readable Policies

STM recommends publishers display human-readable language in the copyright line of content and in their website's Terms and Conditions that clarify the rights reserved by the rightsholder and the policies and instructions for carrying out text and data mining.

Content Copyright Line

Display a humanly readable and visible rights reservation sentence alongside every copyrighted item.

Suggested wording: *Copyright © [YEAR] [ENTITY]. All rights, including for text and data mining, AI training, and similar technologies, are reserved.*

Website Terms & Conditions

Include instructions in the Terms and Conditions of your website

Machine Readable Policies

STM recommends publishers attach machine-readable indicators to content that clarify the rights, policies, and instructions for carrying out text and data mining. These indicators flag to TDM actors the duties and restrictions that may apply to content, and on how to approach the publisher.

Crossref

Use Crossref DOIs to specify TDM license information, using the license_ref element defined in the Crossref metadata schema.

ePUB & PDF formats

Use the TDMRep Protocol to instruct crawlers and processors of PDF or EPUB files that TDM rights are reserved.
The TDMRep Protocol specifies two flags which are added to the metadata of a PDF or EPUB.

Suggested Reading: More detail how to add these flags is available in the full guidelines document.

Provide Instructions for Content Collectors

Unlike measures preventing content collectors from accessing STM content, guidelines for content collectors do not technically impede any action. Bad actors may choose not to respect these instructions. However, implementing these measures clarifies how content may be used. This benefits both publishers and content collectors. STM recommends the following techniques to provide instructions for collectors.

TDMRep Protocol

- **High priority:** Host the TDM file tdmrep.json in the /.well-known directory of the web server
- **Medium priority:** Add HTTP response headers of tdm-reservation (value 1) and tdm-policy (URL pointing to a json policy file)
- **Low priority:** Add HTML meta-tags of tdm-reservation (value 1) and tdm-policy (URL pointing to a json policy file)

HTML Meta Tags

Medium priority: Implement HTML meta-tags to instruct bots and processors (like Microsoft Bing) not to use content for use in Generative AI functionality.

This is especially useful for actors with separate processes for page collection and page processing.

Proposed Full-Use & Obtain Consent Workflow

The TDM Actor must follow the three stage TDM workflow: 1. Determine if a policy exists, 2. Identify the duty, 3. Negotiate with the rights holder if necessary. The content accessed and the technique used to access determine how the publisher's policy must be identified.

Request content	Read rights policy	Request use rights	Fulfill duties
• **TDM Actor** sends HTTP request for content OR downloads PDF • **Publisher Website** sends HTTP 200 with header and content	• **TDM actor** identifies policy location • **TDM actor** requests tdm-policy • **Publisher** tdm-policy duty/action indicates which duty may apply	• **TDM actor reads "assigner" fields for contact info** • **TDM actor** contacts publisher to determine collection of content under target values in JSON	• **TDM actor** is expected to comply with any duty that applies. The TDM actor may contact the Publisher if necessary.

Prevent Access for Prohibited Content Collectors

One basic step publishers can take to protect their content from bots, crawlers, or other actors, is to prevent those entities from accessing content in the first place. Two methods recommended by STM are full prevention through firewalls, or partial prevention using robots.txt file.

Firewall

Prevent bots that are not welcome to collect any of your content, for any reason, by blocking their traffic. You can do this using techniques such as web application firewalls.

Robots.txt file

Instruct bots that are not welcome to collect some, or all, of your content by listing their known names in the robots.txt file.

Although robots.txt has limitations, it is a de-facto standard for content collectors and is easy to implement. STM recommends this implementation as a high priority for publishers.

STM

Contributors

SPRINGER NATURE

WILEY