STM Advancing trusted research

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands

TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

Recommendations

Guidelines on using STM content for Text and Data Mining and for Training of Artificial Intelligence models/systems

Technical Summary – Version 2024-03-27

Starting positions:

- In general, a Text and Data Mining (TDM) reservation for any content is assumed to be contained in a general reservation of rights, like "© All rights reserved."
- With respect to TDM, it is advisable to provide additional human-readable and visible disclaimers against the non-licensed TDM-like use of STM content in the content itself. A rights reservation sentence could be used to this end such as "*Copyright © YEAR ENTITY. All rights, including for text and data mining (TDM), Artificial Intelligence (AI) training, and similar technologies, are reserved.*"
- If such a sentence is repeated in a dedicated metadata field that is part of the STM content, this might be considered as being machine-readable.
- The EU 2019 Directive for Copyright in the Digital Single Market (DSM Directive)[1] provides a definition of TDM that is quite broad and could be interpreted to include AI Training, and therefore Articles 3 and 4 (that define under which conditions TDM can be carried out without authorisation of the rightsholder, i.e. the publisher) of the same Directive are relevant in the AI context and would apply to (at least parts of) AI training.
- Following these articles in this EU Directive, it is strongly advisable to add machine-readable indicators to content, (i) to provide and clarify rightsholders' rights, policies and instructions when using content for TDM (and AI Training) purposes and (ii) to flag to crawlers and other TDM actors that they should not engage in non-licensed use for TDM (and AI Training) purposes. If this is not done, either as recommended below or in a different manner, TDM actors can claim that the EU Directive allows them to perform TDM, including for commercial purposes, on all content that can be lawfully accessed, either through subscription or made available open access.
- Possible mechanisms to inform content collectors and processors about rightsholders' rights, policies and instructions with regards to TDM (and AI Training) are through:
    - Preventing autonomous content collectors from accessing content.
    - Providing machine-readable instructions to content collectors not to collect content.
    - Providing machine-readable Instructions (either at web or content level or both) to content processors on how to license content for TDM (and AI training) purposes, and which content is not to be used for TDM (and AI training) purposes.
    - Including instructions in the Terms and Conditions of the website.
    - Including human-readable and visible disclaimers in the content itself.
- The above-listed solutions are not mutually exclusive, rather complementary and compatible between each other. Publishers can consider implementing all or some of them, depending on the different necessities, technology equipment, and know-how.
- Unlike measures preventing content collectors ("bots", crawlers, etc.) from accessing STM content, the instructions not to or how to collect or process content for TDM purposes are guidelines - they don't technically impede any action and TDM actors can respect them or not. We recommend that each publisher familiarizes themselves with the Terms and Conditions of AI actors for their reference.

[1] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/E, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1591810794966&uri=CELEX:32019L0790

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

Recommendations (High / Medium / Low priority) to prevent unlicensed TDM/AI Training:

Non-technical recommendations

1. **High** Display a humanly readable and visible rights reservation sentence like "*Copyright © YEAR ENTITY. All rights, including for text and data mining, AI training, and similar technologies, are reserved*." alongside every copyrighted content item (web page or PDF ...).

2. **High** Add the same sentence in a dedicated metadata field (if available) that is part of the copyrighted content item.

Web application firewalls

3. **High** Prevent bots that are not welcome to collect any of your content (for any reason) by blocking their traffic using techniques such as web application firewalls.

Robots.txt

4. **High** Instruct bots that are not welcome to collect some, or all, of your content (for any reason) by instructing them not to collect, by listing their known names in the so-called robots.txt file. See also:
    a. https://www.rfc-editor.org/rfc/rfc9309.html
    b. https://platform.openai.com/docs/gptbot
    c. https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers
    d. https://blog.google/technology/ai/an-update-on-web-publisher-controls
    e. https://commoncrawl.org/ccbot

Note: The robots.txt file is not specifically focused on the collection (and the use) for the purpose of text and data mining, but on the collection of data by bots in general. However, as the robots.txt standard is a de-facto standard that is both widely respected and easily implemented, its use is recommended with a high implementation priority.

The TDMRep protocol

5. **Medium** Implement HTML meta-tags to instruct bots and processors that act on HTML meta-tags, like Microsoft Bing, not to use content for use for Generative AI functionality. See also:
    a. https://blogs.bing.com/webmaster/september-2023/Announcing-new-options-for-webmasters-to-control-usage-of-their-content-in-Bing-Chat

Note: This HTML meta-tag can be useful in those cases where a TDM actor has separate page *collection* and page *processing* processes, and the HTTP header information (see below) is unavailable at the time of processing. The TDM File and HTTP header recommendations below are easier and faster to implement and already cover the EU DSM Directive aspect, and thus the priority of this HTML meta-tag implementation is set to medium.

The following recommendations are based on the TDM Reservation Protocol (TDMRep).

6. **High** Use the TDMRep Protocol to instruct bots and processors that TDM rights for all content on a web server are reserved, by hosting the TDM File tdmrep.json in the /.well-known directory of the web server. This json file tdmrep.json will set a value of 1 for the tdm-reservation property and a URL pointing to a json policy file (see below) for the tdm-policy property. Example:

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

```
[
    {
        "location": "/",
        "tdm-reservation": 1,
        "tdm-policy": "https://publisher.com/policies/policy.json"
    }
]
```

7. **Medium** ("high" if the TDM File tdmrep.json mentioned above has not been implemented) Use the TDMRep Protocol to instruct bots and processors that web content TDM rights are reserved, by adding HTTP response headers of tdm-reservation, with a value of 1,and and tdm-policy, with a value of a URL pointing to a json policy file (see below). Example:

    HTTP/1.1 200 OK
    Date: Wed, 14 Jul 2021 12:07:48 GMT
    Content-type: text/html
    tdm-reservation: 1
    tdm-policy: https://publisher.com/policies/policy.json

8. **Medium** (only if there is a need for keeping the TDM property values embedded in the HTML content - in all other cases "Low") Use the TDMRep Protocol to instruct bots and processors that web content TDM rights are reserved, by adding HTML meta-tags of tdm-reservation, with a value of 1, and tdm-policy, with a value of a URL pointing to a json policy file (see below). Example:

```
<head>
    <meta charset="utf-8">
    <meta name="tdm-reservation" content="1">
    <meta name="tdm-policy"
        content="https://publisher.com/policies/policy.json">
    <title>Document title</title>
</head>
```

Note: The above note on the implementation priority of the HTML meta-tag mentioned earlier, also applies to the implementation priority of the TDMRep HTML meta-tag.

Note: In the case that a publisher wants to make it explicit that TDM rights are not reserved, this can always be done by explicitly using the value of 0 for the tdm-reservation flag. This holds for the value provided in the TDM File tdmrep.json, in the HTTP header, and in the HTML meta-tag, but also for the value provided in the PDF and EPUB (see below).

Note: Per the specification of the tdm-reservation in the TDMRep Protocol, if there is a difference between its values (for instance, between the header and a meta-tag), the most detailed (deepest) value of the flag is the one to be observed. So, if a publisher uses a 0 and 1 at different levels, it has to be aware of this hierarchy (for instance, the tdm-reservation value in the meta-tag supersedes the one in the header).

9. **High** Use the TDMRep Protocol to instruct bots and processors of PDF / EPUB files that TDM rights are reserved, by adding tags to the XMP / XML of the files for reservation, with a value of 1, and policy, with a value of a URL pointing to a json policy file (see below).

    a. For PDF, this means: use the TDMRep namespace and add two elements tdm-reservation and tdm-policy to the TDM namespace like this:

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about=""
        xmlns:xmp="http://ns.adobe.com/xap/1.0/"
        xmlns:xmpMM="http://ns.adobe.com/xap/1.0/mm/"
        xmlns:xmpRights="http://ns.adobe.com/xap/1.0/rights/"
        xmlns:tdm="http://www.w3.org/ns/tdmrep/">
    <tdm:reservation>1</tdm:reservation>
    <tdm:policy>
      https://publisher.com/policies/policy.json
    </tdm:policy>
  </rdf:Description>
</rdf:RDF>
```

b. For EPUB, the prefix approach allows for the addition of tdm <meta> tags to support both reservation and policy. This applies at the container level and is represented in the Open Package Format (OPF) file. Subordinate tdm allocations for chapter-level content would use the HTML <meta> tags shown above. Within the OPF the following is used:

```
<package ... prefix="tdm: http://www.w3.org/ns/tdmrep/" ... >
<meta property="tdm:reservation">1</meta>
<meta property="tdm:policy">
    https://publisher.com/policies/policy.json
</meta>
</package>
```

Note: The previous note that the most detailed (deepest) value of the tdm-reservation flag is the one to be observed is especially valuable in the context of a PDF or EPUB file that has been retrieved from a publisher website. For instance, a publisher can choose to set this flag to 1 at the HTTP header level, indicating that TDM rights are reserved for all website content, as even the HTML of OA articles might be surrounded by publisher- and website-specific content for which the rights are reserved. However, for OA articles downloaded from the website as PDF or EPUB files, the tdm-reservation flag can be set to 0, indicating that TDM rights are not reserved for that specific copy. In this case the "deeper" setting of tdm-reservation in the PDF or EPUB file supersedes the higher-level setting in the HTTP header.

10. **High** Use Crossref DOIs to specify TDM license information. Use the license_ref element defined in the Crossref metadata schema (see https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/license-information/) with the applies_to="tdm" attribute setting to specify one or more publisher specific TDM license URLs that contains machine readable and/or human readable TDM policy information.

```
<ai:program xmlns:ai= "http://www.crossref.org/AccessIndicators.xsd" name=
"AccessIndicators">
        <ai:license_ref applies_to =
"tdm">https://www.publisher.com/licensing/v1.html</ai:license_ref>
</ai:program>
```

11. **Medium** Use Crossref DOIs to specify TDM resource links. Use the collection element with the property= "text-mining" attribute setting to specify the resource link(s) at which the content is available for text mining. Note that it is possible to specify multiple resource links and multiple mime types. However, it is not possible to associate individual resource links with specific TDM licenses.

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

See https://gitlab.com/crossref/schema/-/blob/be649453e6c6f73f24980611e0d0279442d743f2/examples/515151.xml/ .

```
<collection property="text-mining">
  <item>
     <resource mime_type =
"application/pdf">https://www.publisher.com/article.pdf</resource>
  </item>
  <item>
     <resource mime_type =
"application/xml">https://www.publisher.com/article.xml</resource>
  </item>
</collection>
```

The afore-mentioned json policy file linked to in the tdm-policy fields is a machine-readable means of providing additional information about TDM licenses and has the following content elements:

- Targets: identifiers (or labels) that refer to collections of content elements.
- Contact: contact details where information on TDM licensing can be requested.
- URL: a link to a file with human-readable information on content licensing possibilities.
- Permissions: describing what type of TDM is allowed or not. For each collection (as identified by a target and possibly further described in the textual licensing file), the two recommended permissions are:
    - Ask consent for TDM if the DSM Directive's Article 3 is not applicable.
    - If the DSM Directive's Article 3 is applicable, allow for TDM.

Sample tdm-policy Machine-readable JSON

```
{
 "@context": [
  "http://www.w3.org/ns/odrl.jsonld",
  "http://www.w3.org/2006/vcard/ns",
  {
   "tdm": "http://www.w3.org/ns/tdmrep/",
   "stm": "http://www.stm-assoc.org/standards/ns/tdmrep/",
   "odrl:leftOperand": {"@type": "@id"},
   "odrl:rightOperand": {"@type": "@id"},
   "vcard:hasTelephone": {"@type": "@id"},
   "vcard:hasEmail": {"@type": "@id"}
  }
 ],

  "uid": "https://publisher.com/policies/tdm/1",
  "@type": "Offer",
  "profile": "http://www.w3.org/ns/tdmrep",
  "assigner":
  {
     "uid": "https://publisher.com",
     "vcard:fn": "Name",
     "vcard:nickname": "N Name",
```

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org


```
      "vcard:hasEmail": "mailto:tdm@name.com",
      "vcard:hasAddress":
      {
         "vcard:street-address": "Street plus number",
         "vcard:postal-code": "Postal code",
         "vcard:locality": "City",
         "vcard:country-name": "Country"
      },
      "vcard:hasTelephone": "tel:+1234567890",
      "vcard:hasURL": "https://www.publisher.com/licensing/v1.html"
   },
   "permission":
   [
      {
         "target": "https://publisher.com#tdm",
         "action": "tdm:mine",
         "constraint":
         [
            {
               "odrl:leftOperand":"purpose",
               "operator": "neq",
               "odrl:rightOperand": "stm:eu-dsm-article3"
            }
         ],
                  "duty":
         [
            {
               "action": "obtainConsent"
            }
         ]
      },
      {
         "target": "https://publisher.com#tdm",
         "action": "tdm:mine",
         "constraint":
         [
            {
               "odrl:leftOperand":"purpose",
               "operator": "eq",
               "odrl:rightOperand": "stm:eu-dsm-article3"
            }
         ]
      }
   ]
}
```

Glossary

- **Digital Object Identifier (DOI):** A DOI is a digital identifier of an object, any object — physical, digital, or abstract. The DOI system has been standardized through the International Standards

ADDR Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands
TEL +44 18 6533 9321
FAX +44 18 6533 9822
www.stm-assoc.org

Organization, as ISO 26324, Digital Object Identifier System. The Standard was approved in November 2010, published in May 2012 and revised in 2022. The Standard specifies the syntax and the functional components of the DOI System, plus the general principles for the creation, registration and administration of DOI names. The Standard is careful not to stipulate any specific technologies. DOI Home Page

- **EPUB:** The EPUB format is an open standard for e-books created by the International Digital Publishing Forum (https://idpf.org/). EPUB is designed for reflowable content, that can adapt its presentation to the reader device, although EPUB now also supports fixed-layout content.
- **HTML:** HTML is the World Wide Web's core markup language. Originally, HTML was primarily designed as a language for semantically describing scientific documents. Its general design, however, has enabled it to be adapted, over the subsequent years, to describe a number of other types of documents and even applications. Current Specification Standard
- **JSON:** JSON (JavaScript Object Notation) is a lightweight data-interchange format that is easy for humans to read and write and for machines to parse and generate. It is based on a subset of the JavaScript Programming Language Standard ECMA-262 3rd Edition - December 1999. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers.
- **Open Package Format (OPF)**: standard that defines the mechanism by which all components of an electronically published work including metadata, reading order and navigational information are packaged into a publication. Open Packaging Format (OPF) 2.0.1 v1.0 (idpf.org)
- **Portable Document Format (PDF):** A file format developed by Adobe Systems that can be used to distribute formatted output, including text and graphics, from a variety of applications to users working on a variety of platforms. Archivists.
- **Rightsholder:** Person or organisation that owns the legal rights to something, in the case of this recommendation this pertains to Web or portable document (e.g., PDF) resources Wiktionary.
- **Publisher:** Person or organisation that makes Web or portable document (e.g., PDF) resources available to the public.
- **Text and Data Mining (TDM)**: per Article 2(2) of the DSM Directive, TDM means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations;
- **TDM Actor:** Person or organisation practising TDM (on Web or portable document – PDF – resources in our case).
- **TDM Licence:** Description of the terms and conditions by which a TDM Actor can process a given Web or portable document (e.g., PDF) resource.
- **TDM Policy:** Description of the kind of TDM Licences a TDM Actor may obtain from a Rightsholder.
- **TDM Rights:** Rights to process a Web or portable document (e.g., PDF) resource via TDM techniques, for a certain purpose (e.g., scientific research, commercial).
- **TDM (Web) Resource:** Identifiable thing available on the Web Wikipedia. Web resources are located using URLs.
- **Web page:** Web resource formatted in HTML.