Opt-Out Vocabulary

1.0 {revremark} : 2024-07-26

1. Purpose

The purpose of this document is to provide a common vocabulary that can be used for machine-readable rights reservations / opt-outs by rightholders who wish to restrict the use of their works and other subject matter for the purpose of AI/ML training and other forms of Text and Data Mining (TDM).

The elements of the vocabulary can be used to describe in a standardized way the types of uses that a rightholder may wish to restrict (or allow), thereby ensuring that rights reservations / opt-outs can be communicated, processed and stored in a consistent and interoperable manner. The vocabulary is agnostic to the technical implementations of opt-out systems and is designed to ensure that opt-out information can be effectively exchanged between different systems.

2. Definitions

Asset

A digital file or stream of data containing protected works or other subject matter, usually with associated metadata.

Rightholder

A person or organization that owns the legal rights to something. In this context, holders of copyright and related rights.

AI/ML Training

The pre-training, training, and fine-tuning of AI models.

3. Vocabulary Structure

The vocabulary consists of the overarching TDM (Text and Data Mining) category and a number of specific use cases that can be addressed independently. The overarching category `TDM` corresponds to the definition of Text and Data Mining in Article 2(2) of the European Union Directive on Copyright in the Digital Single Market.

4. Proposed Vocabulary

The following categories are defined for use in the opt-out vocabulary:

`TDM`

Text and Data Mining. The act of using one or more assets in the context of any automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations for purposes other than scientific research.

> **NOTE** For the purposes of this vocabulary, TDM is not intended to cover the indexing of content for the purpose of online search. This use case is addressed by the existing Robots Exclusion Protocol.

`Generative AI Training`

The act of training AI/ML models that are capable of generating text, images, and other synthetic content using one or more assets as input.

`AI Training`

The act of training AI/ML models that are NOT capable of generating text, images, and other synthetic content using one or more assets as input. Examples include models that can be used for classification or object detection.

This list of categories may be expanded in the future, should a consensus emerge between stakeholders, to include categories that address additional use cases as they emerge. In addition to these categories defined in the vocabulary, it is also expected that some systems implementing this vocabulary may extend this list with additional categories for their particular needs. Systems referencing the vocabulary must not introduce additional categories that encompass existing categories defined in the vocabulary.

4.1. Relationship

The `TDM` category is the overarching category that includes the two categories related to AI/ML training, as they are considered to be forms of `TDM`. As such, when a rights holder reserves the rights to `TDM`, they also opt out of these categories. AI model developers processing opt-outs must therefore interpret an opt-out from `TDM` to also mean an opt-out from `Generative AI Training` and `AI Training`. All categories other than `TDM` can be addressed independently to affect only the specific use case covered by that category.

The figure below shows the relationship between the currently defined categories:

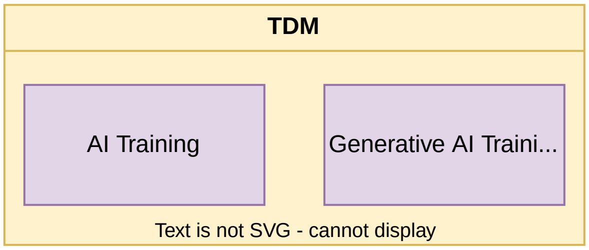

Figure 1. Relationship between the categories

5. Usage

The vocabulary may be used by declaring that an opt-out system or entity expressing or processing opt-outs uses the terms defined in Proposed Vocabulary in accordance with how they are defined in this document.