Opt-Out Vocabulary

1.0 DRAFT : 2024-07-19

Table of Contents

Abstract

This document provides a common vocabulary that seeks to standardize machine readable rights reservations and other rightholder opt-outs from using works protected by copyright and related rights in the context of AI/ML training and other forms of Text and Data Mininig (TDM). It has been developed to adress one of the issues identified in this April 2024 policy brief by Open Future: Considerations for opt-out compliance policies by AI model developers. The paper dentifies four different concepts that require consensus among stakeholders in order to ensure that opt-outs / rights reservation can be expressed in a way that is effective, scalable, and able to meet the needs of both rights holders and AI model developers. the four concepts are (in bold):

> If you tell us what **{identifier}** you want to opt out from which uses **{vocabulary}** via these means **{infrastructure}** then we will do this **{effect of opt-out}**.

This document focuses on the **{vocabulary}** piece of this statement by describing a proposed opt-out Vocabulary that can be used to describe whether one or more assets may be used as part of a data mining or AI/ML training workflow. The purpose of the vocabulary is to provide a stable set of categories that can be used by rightholders to specify the scope of opt-outs / rights reservations they wish to make. A stable, well defined set of rights reservations will ensure that opt-out reservation can be exchanged across value change and that different systems to communicate, process and store opt-out & rights reservations can be designed in an interoperable way. The vocabulary is implementation

agnostic and can be implemented in a variety of opt-out schemes including `location-based` and `unit-based` schemes.

1. Terms and Definitions

Asset

A digital file or stream of data containing content and usually with associated metadata.

Rightsholder

Person or organization that owns the legal rights to something. In this context preliminary holders of copyright and related rights.

Location-based identifiers

A machine-readable location, such as a domain or URL, for use in identification of assets.

Unit-based identifiers

A machine-readable piece of information that identifies a single asset (unit), regardless of where the asset is located, either through the embedding of that information directly into the asset or via a separate database or registry.

AI/ML Training

The pre-training, training and fine-tuning of AI Models before their release.

2. Structure

The vocabulary consists of the overarching Text and Data Mining (`TDM`) category and a number or more specific use cases that can be independently adressed. The overarching `TDM` category is derived from the definition of Text and Data mining in Article 2(2) of the European Union’s Copyright in the Digital Single Market Directive. For the purposes of this vocabulary, TDM is not understood to cover indexing of content for the purpose of online search. This use case is addressed using a `location-based` schema such as the exiting Robots Exclusion Protocol.

Currently the proposed vocabulary contains two use cases that relate to the use of assets for AI/ML training purposes.

3. Proposed Vocabulary

3.1. Categories

The following categories are proposed 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 analysing 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 reserach.

AI Training

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

Generative AI Training

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

This list of categories may be expanded with categories addressing additional uses cases should such use cases 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.

3.2. Relationship

The `TDM` category is the overarching category that includes not only its own use cases, but also the various types of AI training, as they are considered to be forms of TDM. As such, if a rights-holder opts out of or reserves the rights to `TDM`, they are opting out of those other categories as well. The other categories can be independently addressed and affect only the specific use case covered by the category.

The figure below shows the relationship between the categories.

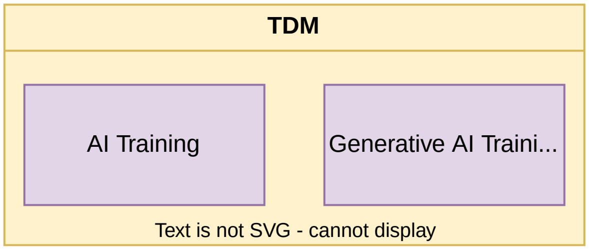

Figure 1. Relationship between the categories