Opt-Out Vocabulary

1.0 DRAFT : 2024-07-18

Table of Contents

Abstract

Transparency of the data used to train AI models is a prerequisite for understanding how these models work. It is crucial for improving accountability in AI development and can strengthen people's ability to exercise their fundamental rights. Yet, opacity in training data is often used to protect AI-developing companies from scrutiny and competition, affecting both copyright holders and anyone else trying to get a better understanding of how these models function.

Open Future, is committed to advancing openness, transparency, and good governance in AI development. As part of this commitment, they have written a series of white papers on the topic of AI and rights including Considerations FOR opt-out compliance policies by AI model developers. In that paper, it describes the framework for machine-readable rights reservations required by Article 4(3) of the Copyright in the Digital Single Markets (CDSM) directive statement as:

> 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}** pieces of that 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. It is intended to be useful for both location-based as well as unit-based asset identifiers.

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. See Wiktionary.

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.

2. Use Cases

3. Proposed Vocabulary

3.1. Categories

The following categories are proposed for use in the Opt-Out Vocabulary, based on the set of use cases identified in the previous section:

Search and Discovery

The act of indexing the content (and/or metadata) of assets for the purpose of retrieval.

TDM

Text and Data Mining. The Copyright in the Digital Single Market (CDSM) Directive defines TDM as "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" (Article 2.2).

AI Training

The act of training an AI/ML (Artificial Intelligence/Machine Learning) model using one or more assets as input. This can include training for classification, object detection, as well as generative AI.

Generative AI Training

A form of `AI Training` where the AI/ML model being trained can generate new assets based on the training data.

In addition to the pre-defined categories, it is also expected that some systems may extend this list with additional categories for their particular needs.

3.2. Relationship

`Search and Discovery` is a category that is separate from the others, as it is not a form of TDM or AI Training. It is however reflected here as some opt-out systems (i.e., `robots.txt`) include this category.

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 rightsholder opts out of TDM, they are opting out of those other categories as well.

The figure below shows the relationship between the categories.

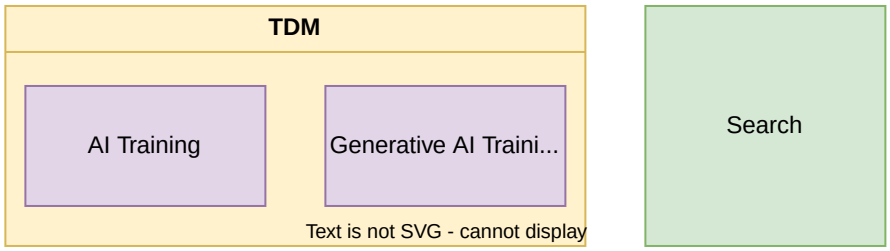

Figure 1. Relationship between the categories