

IEFT AI preferences working group - update in the work on the vocabulary (August 2025)

On 17–18 July, the [IETF AI Preferences Working Group](#) held a two-day design team meeting to advance work on the two drafts within its charter: The draft for a vocabulary to express preferences and a separate draft specifying an attachment mechanism (via robots.txt and HTM header files). The aim was to bring both drafts as close as possible to a state suitable for a final Working Group Call for Consensus, which will take place on the mailing list at a later date.

For the vocabulary draft—co-edited by Paul and based on our original proposal—the meeting largely validated the overall approach. While the chairs noted that there appeared to be a reasonable level of consensus, two issues remained points of contention:

1. Hierarchy: whether an overarching TDM category should exist to encompass all others.
2. Inference category: whether to include an `ai-inference` category alongside the existing `ai-training` and `search` categories.

Most of the London meeting’s discussion focused on the inference category. Despite the chairs’ conclusion that consensus was sufficient, the discussion left the impression that some concerns around this category had not been fully resolved.

Various participants voiced concerns that an `ai-inference` category covering all uses of assets by a trained AI system would be too broad. A subset of participants raising these concerns specifically highlighted the fact that such a category could potentially impact users’ rights under exceptions and limitations to copyright or under other legal regimes such as accessibility legislation. Participants highlighting such concerns pointed to the possibility that adherence to non-binding expressions of preferences (such as “no” to `ai-inference`) might become a legal requirement through future legislation or regulation referencing preferences expressed in conformance with the vocabulary.

On the other hand, a number of participants expressed a strong conviction that the vocabulary should contain an `ai-inference` category, mainly on the grounds that uses of existing assets by AI systems are something that publishers and other declaring parties want to control, and that such uses are expected to be central to emerging business models. In addition, some participants noted that certain existing standards in this field, such as CWAG and IPTC-Plus, already provide mechanisms to express preferences related to uses at inference time, and that including such a category would therefore be important to ensure compatibility with established approaches.

In the end, these two positions proved to be irreconcilable. Based on the discussion, the editors decided to maintain the inclusion of a broad `ai-inference` category in the draft (subsequently renamed `ai-use`).

All in all, the impression at the end of the London design team meeting was that the group was close to consensus, even though the discussion on inference/use had been somewhat rushed and not fully conclusive.

Post London draft and Madrid WG session

Based on the discussion at the London meeting the editor produced an [updated vocabulary draft \(v02\)](#) as input for the Madrid working group meeting. This draft contained the following vocabulary definition:

None

4. Vocabulary Definition

This section defines the categories of use in the vocabulary.

Figure 1 shows the relationship between these categories:

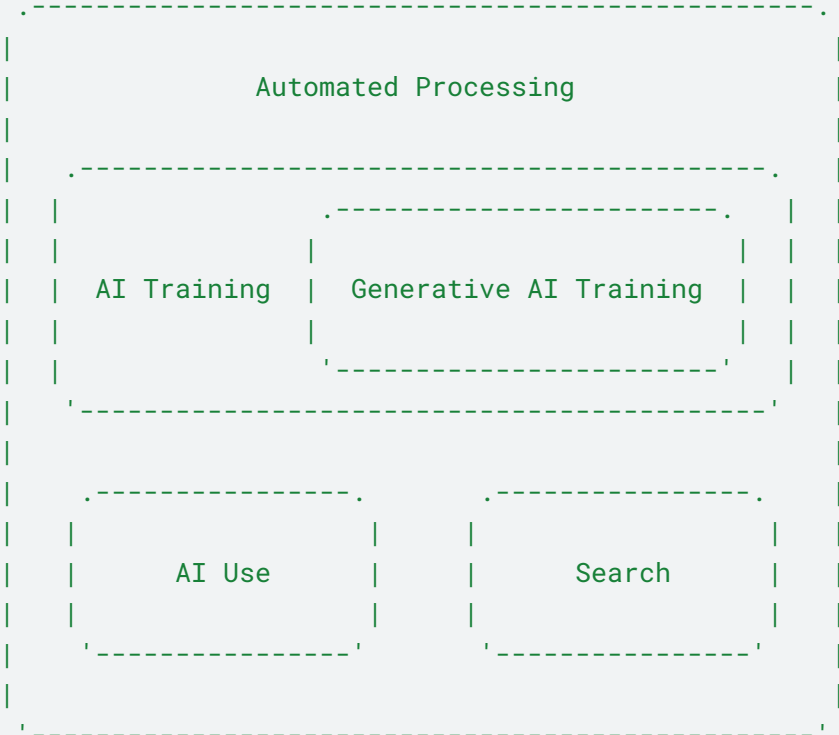


Figure 1: Relationship Between Categories of Use

4.1. Automated Processing Category

The act of using one or more assets in the context of automated processing aimed at analyzing text and data in order to generate information which includes but is not limited to patterns, trends and correlations.

The use of assets for automated processing encompasses all the subsequent categories.

4.2. AI Training Category

The act of training machine learning models or artificial intelligence (AI).

The use of assets for AI Training is a proper subset of Automated Processing usage.

4.3. Generative AI Training Category

The act of training general purpose AI models that have the capacity to generate text, images or other forms of synthetic content, or the act of training more specialized AI models that have the purpose of generating text, images or other forms of synthetic content.

The use of assets for Generative AI Training is a proper subset of AI Training usage.

4.4. AI Use Category

The act of using one or more assets as input to a trained AI/ML model as part of the operation of that model (as opposed to the training of the model).

The use of assets for AI Use is a proper subset of Automated Processing usage.

4.5. Search Category

Using one or more assets in a search application that directs users to the location from which the assets were retrieved.

The purpose of defining a distinct Search category is to allow preferences to be expressed about search applications, independent of other categories of use. A distinct Search category allows for preferences specific to search applications, even if the use of AI is involved in their implementation.

The use of assets for Search is a proper subset of Automated Processing usage.

Despite the discussion, the vocabulary definition in this draft was essentially the same as in the drafts that preceded the London design team meeting. The main differences concern the naming of some of the categories and some clarifying language added to the [search](#) category. By contrast, the definitions of [TDM](#) (now [Automated processing](#)), [AI-training](#) and [Generative AI training](#) have been stable since the original draft.

The updated draft was presented and discussed at the [AI Preferences WG session at IETF 123](#) on the 21st of July, where most of the discussion again revolved around the [ai-use](#) category.

Post Madrid discussion

After the Madrid meeting, the discussion about the [ai-use](#) category continued on the mailing list, with a number of vocal participants arguing against its inclusion on the grounds that it is considered to be overly broad. These participants have generally proposed limiting the current version of the vocabulary to the categories related to the training of AI models, which in their view should be the main focus of the work.

In response, the Chairs have pointed out that the charter of the working group covers “the expression of preferences about how content is collected and processed for Artificial Intelligence (AI) model development, deployment, and use,” and that inference-time preferences are thus in scope for the working group.

However, this has not allayed concerns, with additional participants pointing out that, as AI becomes more widely used, a broad [ai-use](#) category poses a danger of applying to a very wide set of interactions with online content. These participants argue that, if regulators or legislators introduce requirements to comply with the proposed vocabulary—as, according to them, the EU has done via the Code of Practice for AI model providers—the presence of an [ai-use](#) category in the vocabulary would give publishers and rightsholders the ability to govern uses of publicly available content that have traditionally been outside their control (for good reasons). This line of argument also discards the counter-argument that the vocabulary that is being developed is merely intended to express preferences that can be overridden/ignored by implementing parties.

Another thread of discussion on the mailing list has focussed on strengthening the existing language in the draft that deals with compliance and reasons for ignoring expressions of preference. This has resulted in the inclusion of additional language in the “Respecting Preferences” and “Applicability and Legal Effect” sections of the draft:

None

Respecting Preferences

Specification conformance does not encompass whether preferences are actually respected during data processing. A data processor MAY choose to respect preferences that it has discovered, according to:

- * an understanding of the nature of the processing being performed and how it corresponds to the usage categories where preferences have been expressed, and

* the applicable legal context; see [Applicability and Legal Effect](#).

Usage preferences can be overridden through express agreements between relevant parties.

There are also many situations where other priorities could override any usage preferences. Priorities that could justify ignoring preferences include -- but are not limited to -- free expression, safety, education, scholarship, research, preservation, interoperability, and accessibility.

A choice to ignore a preference could be explicitly permitted in law or be based on the judgement of particular individuals or organizations.

The following lists examples of cases where other priorities could override specific preferences:

* People with accessibility needs, or organizations working on their behalf, might ignore a preference to disallow Automated Processing in order to access automated captions or generate accessible formats.

* A cultural heritage organization could ignore a preference to disallow Automated Processing in order to provide more useful, reliable, or discoverable access to historical web collections.

* An educational institution could ignore a preference to disallow AI Training in order to enable scholars to develop or use tools to facilitate scientific or other types of research.

* A website that permits user uploads could ignore a preference to disallow Automated Processing in order to develop or use tools that detect harmful content according to established terms of use.

None

[Applicability and Legal Effect](#)

This document provides a set of definitions for different categories of use, plus a system for associating simple preferences to each (allow, disallow, or no preference; see [Statements of Preference](#)).

The categories of use that are defined as part of the vocabulary are not always clearly applicable or inapplicable to a particular system

or application. The universe of possible systems is far more complex than any simple vocabulary is capable of describing. That means that some discretion could be involved in deciding whether a preference applies.

The expression of preferences might activate regulatory or legal consequences, which has implications for entities that consume those preferences. Their interpretation of the meaning of different terms could have legal ramifications. Different jurisdictions could reach subtly different conclusions about the applicability of each category of use to specific applications.

It is the responsibility of those that process affected assets to understand the legal implications of their use of digital assets.

This includes understanding:

- * obligations regarding how preferences are obtained (in particular, which methods of associating preferences with content are expected to be understood),
- * the specific uses to which assets are put,
- * how preferences apply to the those uses, and
- * how relevant jurisdictions might interpret those preferences.

These considerations will depend on jurisdiction and the details of the system.

However the strengthening of the language in these two sections has not had a noticeable effect on the opposition to the inclusion of the **ai-use** category in the discussion on the mailing list.

There has also been some more discussion on the mailing list related to the definition of the **search** category. Here the main concerns continue to be that publishers are looking for a category that allows them to indicate that irrespective of other preferences that apply to their assets, they want these assets to be included in traditional search applications. Other participants have argued that the concept of 'traditional' search applications is increasingly meaningless since almost all search applications make use of AI in various stages of deployment (and have been doing so for many years). To address these changes, the editors have proposed an updated (albeit much more verbose) definition of the **search** category:

None

4.5. Search Category

Using one or more assets in a search application that directs users to the location from which the assets were retrieved.

Search applications can be complex and may serve multiple purposes. Only those parts of applications that direct users to the location of an asset are included in this category of use. This includes the use of titles or excerpts from assets that are used to help users select between multiple candidate options.

Preferences for the Search category apply to those parts of applications that provide search capabilities, regardless of what other preferences are expressed. Though search applications often employ AI and so might otherwise be governed by AI Use preferences, preferences regarding AI Use are overridden by preferences for the Search category.

Parts of applications that do not direct users to the location of assets, such as summaries, are not covered by this category of use.

The use of assets for Search is a proper subset of Automated Processing usage.

On the 8th of August the Working group chairs posted the following message to the AI-pref mailing list:

In Brussels, we had strong agreement in the room to start with a simple framework that had terms for a high-level, widely encompassing preference (formerly 'tdm' and currently 'all') along with preferences for AI training (both generative and otherwise). The chairs believe that there is consensus around these items, and plan to confirm this explicitly soon.

We also discussed the possibility of adding a preference for 'search' to assure that someone opting out of AI would not also accidentally opt out of Web search listings, even if the search engine uses AI. Discussion of this continues -- especially around the inclusion of 'summary' search results using AI -- but the chairs believe that discussion is progressing well, and that we have a good chance of reaching consensus after some additional discussion and, potentially, adjustments.

In addition, we agreed to discuss adding a preference for 'inference' (now 'use'), and the editors incorporated a proposal for that term into the vocabulary document. At the time, many (including our editors) expressed concern about the need for considerable discussion and work in this area to make it suitable for

standardisation. That has proven true; as chairs, we are finding it difficult to see a path to consensus in the foreseeable future.

To enable the group to focus on shipping its deliverables, then, we are instructing the editors to remove the 'use' term from the document for now. Please note that this does not preclude continued discussion, nor does it prevent us from coming to consensus on including it or an alternative approach (such as the purpose-focused approach that has been alluded to), either in this document or in subsequent work.

In addition they also scheduled an additional interim meeting for the working group on 30/9-2/10 in Zürich.

After the removal of the **ai-use** category, the vocabulary definition in [the working draft](#) looks like this:

None

4. Vocabulary Definition

This section defines the categories of use in the vocabulary. Figure 1 shows the relationship between these categories:

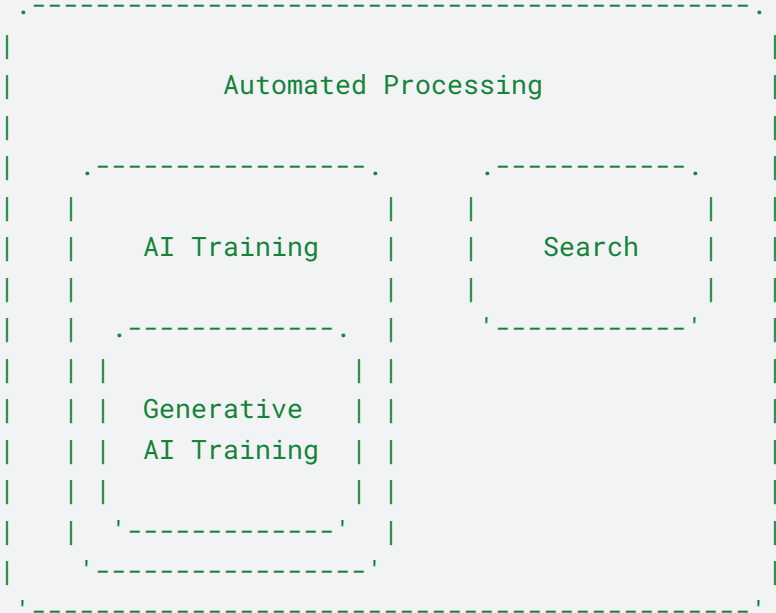


Figure 1: Relationship Between Categories of Use

4.1 Automated Processing Category

The act of using one or more assets in the context of automated processing aimed at analyzing text and data in order to generate

information which includes but is not limited to patterns, trends and correlations.

The use of assets for automated processing encompasses all the subsequent categories.

4.2 AI Training Category

The act of training machine learning models or artificial intelligence (AI).

The use of assets for AI Training is a proper subset of Automated Processing usage.

4.3 Generative AI Training Category

The act of training general purpose AI models that have the capacity to generate text, images or other forms of synthetic content, or the act of training more specialized AI models that have the purpose of generating text, images or other forms of synthetic content.

The use of assets for Generative AI Training is a proper subset of AI Training usage.

4.4 Search Category

Using one or more assets in a search application that directs users to the location from which the assets were retrieved.

Search applications can be complex and may serve multiple purposes. Only those parts of applications that direct users to the location of an asset are included in this category of use. This includes the use of titles or excerpts from assets that are used to help users select between multiple candidate options.

Preferences for the Search category apply to those parts of applications that provide search capabilities, regardless of what other preferences are expressed.

Parts of applications that do not direct users to the location of assets, such as summaries, are not covered by this category of use.

The use of assets for Search is a proper subset of Automated Processing usage.

Assessment of the current situation

At the time of writing (mid-August), the situation is considerably more complex than it appeared after the London design team meeting, and it seems there will be at least some delay in wrapping up the work on the vocabulary (i.e., moving to a Working Group Last Call).

The discussions at and following the London meeting have resulted in a draft that, at its core (the vocabulary definition), remains very much in line with the original. They have also produced various improvements that seem largely uncontroversial at this point and that address the majority of the issues identified during the Brussels meeting in April of this year, which preceded the actual drafting work. These include important additional safeguards, in the form of new language clarifying the relationship between expressed preferences and rights and obligations arising from various legal frameworks (see above).

The fact that it has been more difficult to achieve consensus on the categories related to [search](#) and [ai-training](#) is in line with our expectations.

Search category

As expressed by the chairs in their recent message, it still seems possible at this stage to achieve consensus on including a search category in the vocabulary. The ability to maintain a presence in traditional search applications, even when opting out of TDM/AI training, has been a key requirement from rightholders since the start of our discussions on implementing the EU copyright framework. We believe it is important to include such a category in the IETF working group's initial deliverable.

While it is evident that even 'traditional' search is increasingly powered by AI systems, this does not currently prevent separating the use of assets by a deployed search system (indexing and retrieval) from the use of those assets to train or improve the system. As long as this distinction is possible, the vocabulary should allow rightholders to express such a preference.

(Removal of) Inference category

The (for now) removal of the inference category makes sense given the current state of discussion. It is indeed difficult to see a path to consensus on this issue, which is also challenging to define correctly. There are several issues at play:

- At the highest level, there is a real risk that this category would be too broad, potentially "giving site owners fine-grained control over any interaction with their sites, whether that's for interoperability, modifying the content for user needs, or anything else," as long as some form of AI is involved. As multiple participants have pointed out, such control could conflict with users' rights under copyright, fundamental rights, and more specific legal regimes such as accessibility

requirements. It would conceptually expand publishers' control beyond the scope previously enabled by copyright.

- We have so far considered the vocabulary to be intended for the expression of non-binding preferences that can be ignored or overridden, as a safeguard against such a scenario. As outlined above, the current draft includes substantially improved language to this effect. However, the effectiveness of this safeguard is being questioned by some participants, who point to the specter of regulators or lawmakers introducing requirements to comply with expressions of preference, thereby turning them into legal obligations. In this context, they cite the inclusion of a requirement to comply with robots.txt in the European Union's Code of Practice for GPAI model developers. This concern overstates the scope of the relevant provision, which is explicitly limited to the purpose of "training of [...] general-purpose AI models." As such, the Code of Practice would not give additional legal weight to preferences related to the use of copyright works by already trained AI models. If anything, the CoP example shows that regulators are capable of scoping references to technical specifications in ways that preserve existing trade-offs between different sets of rights.
- Publishing the initial version of the vocabulary without a category related to inference-time uses would also risk undermining one of its core purposes: enabling semantic interoperability between different mechanisms for expressing AI-related preferences. As mentioned above, a number of existing systems (such as CWAG/C2PA and IPTC-plus) already allow for the expression of inference-related preferences. Excluding such a category would limit the vocabulary's usefulness for the providers of these systems.

Overall, publishing the initial version of the vocabulary without a category related to inference-time uses would largely align with positions taken by AI developers. There is a risk that allowing preferences related to training while excluding those related to inference-time uses could undermine the credibility of the overall effort. From the perspective of rightholders, the ability to signal preferences on inference-time uses of their works—where they perceive some leverage—is likely seen as more important than signalling preferences related to training, where many consider that "the damage has already been done."

In practice, such an outcome would likely incentivize declaring parties to opt out of the overall automated processing category while opting in for search. This would counteract the underlying objective of encouraging targeted expressions of preference. It seems unlikely that such an outcome would benefit the overall ecosystem.

Robots.txt as an attachment mechanism

There have also been voices on the mailing list questioning the overall wisdom of using robots.txt as a vehicle for conveying AI-related preferences. The concerns expressed in this context partially overlap with those about regulatory interventions raised in discussions on

the [ai-use](#) category. Fundamentally, these voices have warned against using the IETF as a platform for exporting what is perceived as a European Union–specific approach to regulating online speech. It is unclear to what extent this criticism can influence the working group’s overall outcome, since the current draft is well within its chartered purpose.