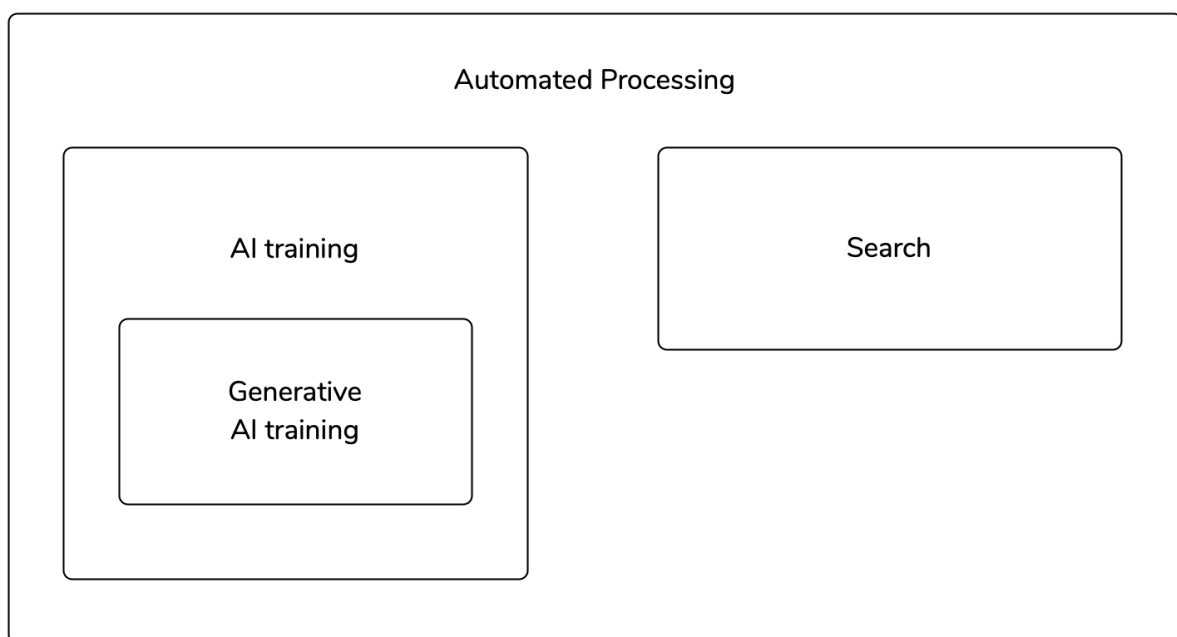


## Evolution of the AI preference vocabulary during the Zürich interim meeting

---

This note summarizes how the AI-preferences vocabulary currently being developed by the [IETF AI-Pref working group](#) evolved during the Zürich interim, focusing on the rationale for shifting from input-based to output-based categories and the implications for training and search-related control.

The working group went into the Zürich Interim meeting with the following proposal for a vocabulary (See [Section 4. Vocabulary Definition of draft-ietf-AIpref-vocab-03](#)):



This vocabulary definition remained structurally similar to the definition in the original draft. It contained a top level category (Automated Processing), two definitions related to use of assets for AI training (AI training and the Generative AI training sub category) and one definition related to use of assets by trained models/systems (Search).

Prior to the interim meeting, WG participants had filed issues that made it clear that all of these categories were disputed to a certain degree. The issues filed ranged from issues challenging individual definitions ([Issue #173](#), [Issue #155](#), [Issue #157](#)) to issues suggesting a replacement of almost all of the categories ([Issue #149](#)) and proposals for new categories ([Issue #150](#)). In addition, participants had also filed issues questioning the hierarchical structure and presence of a top-level category ([Issue #170](#)).

Trying to find ways to address these issues consumed most of the time of the interim meeting. This discussion consisted of two larger threads: (1) a thread dealing with the need for / desirability of a top-level or catch-all category and (2) a thread dealing with the structure of the categories, initially focusing on the non-training-related categories (search, substitutive use, and display-level categories).

The discussions related to (1) ultimately remained inconclusive with no clear preference for or against a top-level category emerging. This question remains unresolved after the meeting and will need to be revisited by the working group in the future. Given this, the presence of a top-level category is neither assumed nor excluded in the scenarios discussed below (hence the dotted line in the visualizations below).

## Shift towards output-based categories

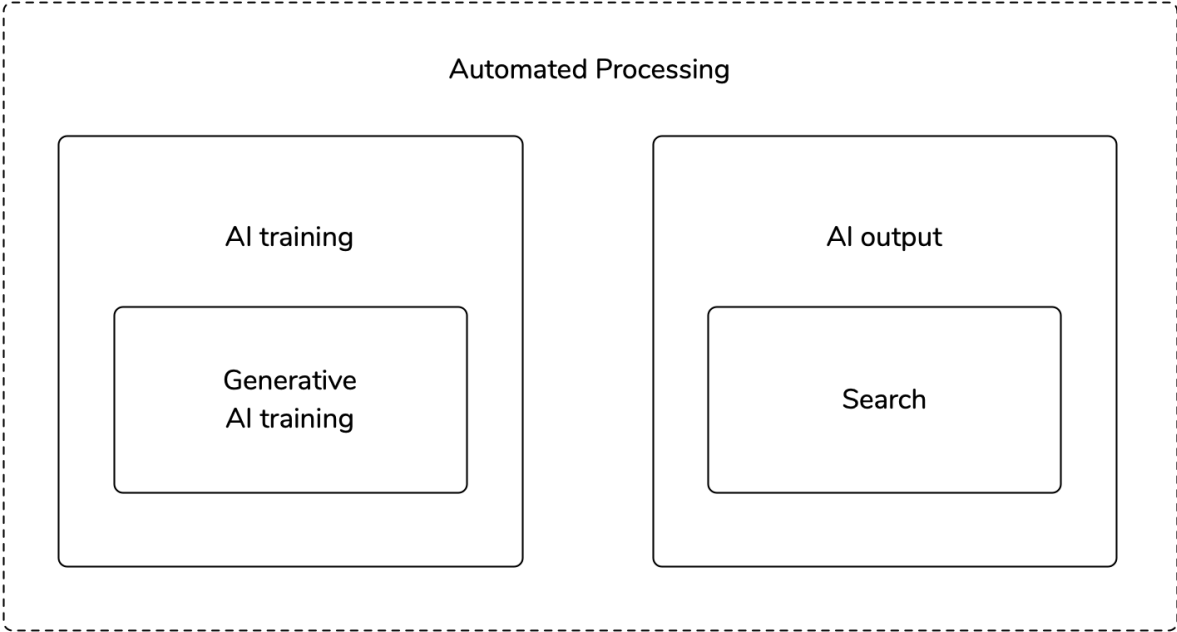
The vocabulary included in [draft-ietf-Alpref-vocab-03](#) assumed a simple dichotomy between the use of assets for AI training and the use of assets as input for a trained system. This dichotomy was based on an understanding of the underlying technology that assumes that it is possible to clearly differentiate between two distinct phases of the AI lifecycle: a training phase (including fine-tuning and other forms of post-training) and a deployment phase where assets are used as input to the model (context, grounding, RAG). The working group had previously identified two possible use cases related to the deployment phase for possible inclusion in the vocabulary. One related to “traditional” (a.k.a. 10 blue links) search and one related to inference (broadly the use of assets as input to a model that generates output based on them).

Discussions in the WG generally assumed that most content owners intended to allow search, while many content owners wanted to disallow other inference-time uses.

A number of participants argued that trying to define categories for the deployment phase by how assets were used by AI systems was fundamentally flawed and that the vocabulary should instead focus on how elements of assets used as input for AI systems were *displayed*. Such an approach, it was argued, would allow for making more meaningful differentiations between use cases such as “traditional” search and use cases that are being regarded as substitutive such as summaries. This approach was advocated by proposals coming from both participants representing AI developers who also offer traditional search engines.

During the 2nd day of the meeting, an ad-hoc design team was formed to explore the possibility of a hybrid approach that would combine the idea of display-based preferences (subsequently referred to as “output-based”) with the existing vocabulary.

This resulted in the following proposal with the existing training categories on the left and two new output-based categories (AI output and Search) on the right:



In this proposal (see [this commit](#) for the proposed changes) the AI output category is defined as:

Using an asset in an AI-based system in the generation of outputs that are presented to clients of that system. This does not apply to any assets that are directly provided as inputs. This includes the output of search results. This includes outputs that are presented to human users and outputs that are presented to automated clients<sup>1</sup>.

The search category is then defined as a sub category of the AI output category:

The Search category of use is a refinement of the AI output usage, with the addition of the following two conditions:

- A reference to the location that the asset was obtained is presented as part of the output.
- Only excerpts of material that is drawn verbatim from the asset can be presented as part of the output.

---

<sup>1</sup> The mention of “outputs that are presented to automated clients” in this definition does not reflect a consensus within the ad hoc drafting group, and is an unresolved question with some members of the ad hoc group arguing in favour and some against. The same arguments in favour and against also resurfaced in the subsequent discussion of the full group.

With both these conditions, a preference to allow Search usage enables the presentation of links and titles in what is considered “traditional” search results.

The use of assets for Search is a proper subset of AI Output usage.

The intent of this approach was to limit these preferences to uses of assets that result in output being presented to a user (as opposed to purely internal uses of assets). The overall AI output category would cover any such use, while the `search` sub category would only cover cases where the output resembles traditional search result pages. This setup was intended to enable content owners to opt-out of all relevant uses of their assets by trained AI systems while giving them the ability to signal that assets can be used for display in “traditional” search results.

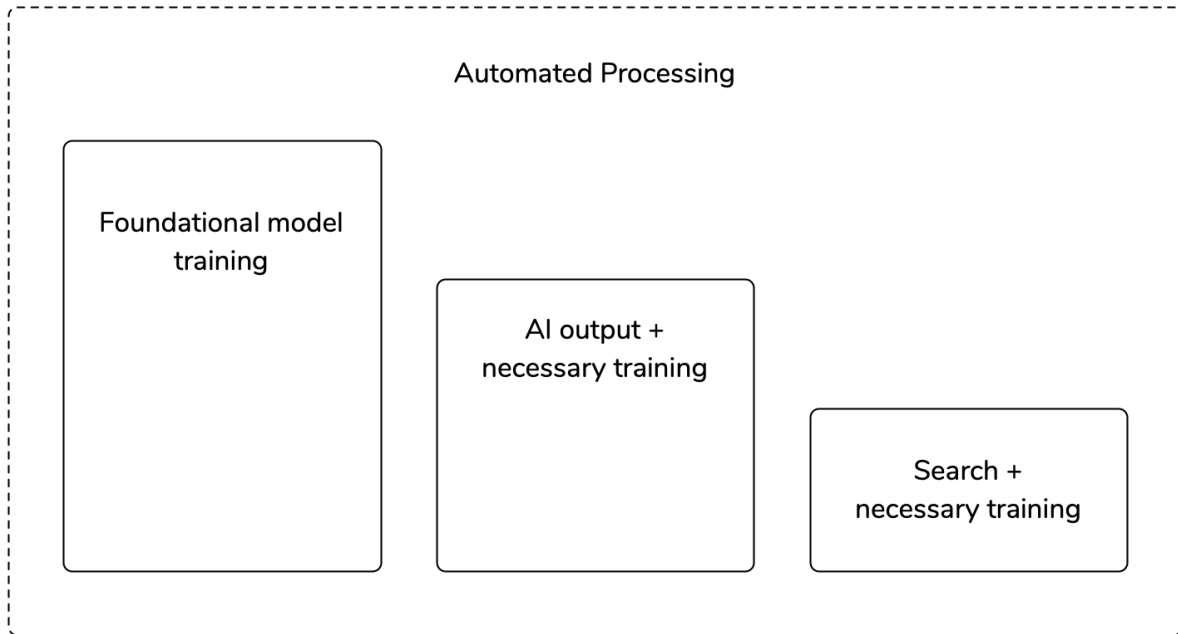
## Bridging training and deployment

During the discussion of this new approach another issue emerged. Participants suggested that indicating `search=y` would be insufficient to enable the inclusion of assets in “traditional search” interfaces as long as this was not accompanied by a permission to train models for this purpose. This suggestion was confirmed by representatives from one of the search engine providers participating in the discussion who noted that “traditional” search (as implemented today) has a strong need for AI training on the content in the search index to produce effective results.

This insight led to another breakout session by the ad-hoc design team that examined the distinction between the training and deployment phases. The outcome of this discussion was a new proposal that consists of three categories. One on the training side (`Foundational model training`) and two on the deployment side (`AI output` and `Search`<sup>2</sup>) that combine output derived definitions with the training necessary to achieve that output:

---

<sup>2</sup> During the discussion in Zürich the “search” category was referred to with a placeholder term (“[Lederhosen](#)”). This reflects an unresolved disagreement about naming the category. Some participants preferred to name categories based on an objective description of what they do, not on an interpretation of their purpose. Others wanted the label “search” to make the vocabulary easier to use, but this was criticized as creating a false sense of certainty because people mean very different things by „search“.



This model does away with the clear distinction between training and deployment phases. The two output categories (that are defined in the same way as in the previous model) now also include “whatever model training is necessary to produce the models that are used in the generation of these outputs”.

While the model is presented above (as it was during the workshop) as three independent categories, there was considerable discussion during the meeting if the AI output and Search categories should be nested in the same way as in the previous model. This discussion was not resolved during the meeting. The general approach was seen as a step in the right direction and no fundamental objections were made during the discussion. There was also only very little discussion about the change from two training related categories (AI training and Generative AI training) to a single Foundational model training category although some participants question the reason for the change in terminology here. Generally it will remain to be seen if this model allows for a granular enough expression of preferences that addresses the needs formulated by key stakeholders.

Members of the ad-hoc group also noted that the vocabulary could later be extended with additional output-based refinements (such as those suggested in [Issue #149](#)), but that these were explicitly left out of scope for the group’s discussion.

## Intent behind the new approach

This new approach tries to bridge a number of different concerns that seemed more difficult to reconcile under the original model contained in [draft-ietf-Alpref-vocab-03](#):

- The introduction of the `AI output` category on the deployment side gives content owners the ability to express preferences related to the use of their works by deployed AI systems, without affecting uses of content for purely internal purposes that do not result in an output to users of the system. This addresses concerns raised by some AI model developers in response to earlier attempts to define a category dealing with uses of assets by deployed AI systems.
- The definition of the `search` category in terms of the type of output it creates allows for indicating a preference about “traditional” search that is based on how search results are presented to users without limiting how the search results are produced. This addresses concerns raised by some AI model developers in response to earlier attempts to define a search category, who had pointed out that use of assets as input for AI systems has been an essential part of providing search applications for a long time already.
- The inclusion of model training that is necessary to produce the models that are used in the generation of the outputs in the definition of each of the deployment side categories ensures that preferences can be expressed and acted upon independent from any training side preferences.
- The consolidation of training stage categories into a single `Foundational model training` category provides content owners with a single preference to allow or disallow a broad range of training uses — note that the exact scope of this category has been left undefined by the ad hoc design team, which makes it impossible to assess if the scope of the new preference is acceptable to content owners and AI developers. Acceptability may also hinge on whether the vocabulary retains an overarching top-level preference, especially if `Foundational model training` is defined more narrowly than the current `AI training` category.

## Reception and outlook

This approach was presented in the final phases of the meeting and will be vetted by stakeholders across the spectrum. The introduction of the `AI output` and `Search` categories address some of the longstanding concerns voiced by content owners. At the same time the acknowledgement that training and deployment cannot be fully separated and that some training permissions are necessary to allow meaningful use by search applications are likely to be questioned by some participants.

On the positive side the fact that the three categories proposed in the updated model broadly mirror the three categories proposed by Cloudflare in [draft-romm-aipref-contentsignals-00](#), suggest that they might broadly align with the types of uses that website owners seek to control / express preferences about.