AI Preferences
Internet-Draft
Intended status: Informational
Expires: 25 March 2026

K. Madhavan
F. Canel
J. Gimbel
S. Cooper
Microsoft Corporation

21 September 2025

A Vocabulary for Controlling Usage of Content Collected by Search and AI Crawlers

draft-madhavan-aipref-displaybasedpref-01

Abstract

This document proposes a standardized vocabulary to express preferences for usage of digital content collected by Search and AI crawlers. This vocabulary allows for the creation of structured declarations about restrictions or permissions for use of content retrieved by such systems.

About This Document

This note is to be removed before publishing as an RFC.

The latest revision of this draft can be found at https://kmadhavan-msft.github.io/i-d-ietf-aipref-displaybasedpref/draft-madhavan-aipref-displaybasedpref.html. Status information for this document may be found at https://datatracker.ietf.org/doc/draft-madhavan-aipref-displaybasedpref/.

Discussion of this document takes place on the AI Preferences Working Group mailing list (mailto:ai-control@ietf.org), which is archived at https://mailarchive.ietf.org/arch/browse/ai-control/. Subscribe at https://www.ietf.org/mailman/listinfo/ai-control/.

Source for this draft and an issue tracker can be found at https://github.com/kmadhavan-msft/i-d-ietf-aipref-displaybasedpref.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Copyright Notice

Table of Contents

1. Introduction

This document defines a common vocabulary of terms for search and AI systems that process digital content. The primary purpose of this vocabulary is to enable machine-readable expressions of preferences about using digital content collected by Search and AI crawlers.

The terms defined by the vocabulary can be used to describe, in a standardized way, the types of uses that a declaring party may wish to explicitly restrict or allow. Preferences are then expressed as a grant or denial of permission concerning each of the types of use defined in the vocabulary. This ensures that preferences can be communicated, processed, and stored in a consistent and interoperable manner.

The vocabulary or the preferences that might be expressed do not proscribe how automated processing systems obtain or act on preferences. Separate documents will describe how preferences might be associated with digital content. It is designed to ensure that preference information can be exchanged between different systems and consistently understood. A reader will also find that this document identifies existing implementations of certain vocabulary elements, helping readers connect these concepts to current preferences supported by most search engines and AI solutions. The authors anticipate removing the references to existing implementations in the final version.

Expressing preferences is without prejudice to applicable laws including the applicability of exceptions and limitations to copyright.

2. Conventions and Definitions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

This document uses the following terms:

Crawler: A crawler is an automated program that scans the web, collecting content (web pages, images, documents etc.) or availability status per URI scanned.

3. Statements of Preference

The vocabulary is a set of categories, each of which is defined to cover a class of usage for digital content. The section on Section 4 defines these categories in more detail.

A statement of preference is made about a specific digital content. Statements of preferences can assign preferences to each of the categories of use in the vocabulary.

A statement of preferences can express preferences about some, all, or none of the categories from the vocabulary. This can mean that no preference is expressed for a given usage category.

In the absence of a statement of preference, no preference is set.

3.1. Conformance

TODO Conformance

3.2. Applicability and Effect

This specification provides a set of definitions for different categories of use based on expressed display preferences.

This specification does not provide any enforcement mechanism for those preferences, and conformance to it does not encompass whether preferences are actually respected during data processing.

Preferences do not themselves create rights or prohibitions, either in the positive or the negative. Other mechanisms—technical, legal, contractual, or otherwise—might enforce stated preferences and thereby determine the consequences of following or not following a stated preference.

An entity that receives usage preferences MAY choose to respect those preferences it has discovered, according to an understanding of how the asset is used, how that usage corresponds to the usage categories where preferences have been stated, and the applicable legal context.

Usage preferences can be ignored due to express agreements between relevant parties, explicit provisions of law, or the exercise of discretion in situations where widely recognized priorities justify doing so. Priorities that could justify ignoring preferences include - but are not limited to - free expression, safety, education, scholarship, research, preservation, interoperability, and accessibility.

Because enforcement is not provided by this specification, the consequences of ignoring preferences could vary depending upon how a given legal jurisdiction recognizes preferences.

4. Vocabulary Definition

The following definitions apply to content collected by search and AI crawlers. It does not include user-initiated access of content. All these categories apply independently of each other with the most restrictive taking precedence in case all/some categories are present.

4.1. Indexing and Retrieval

The act of allowing or disallowing content collected by web crawlers from being indexed or retrieved for purposes of display. Such preference mechanism can also be applied for cases where digital content is not accessible. In existing implementations, access preferences are typically expressed via the NOINDEX statement set in HTTP header or meta tags.

4.2. Display text

The act of allowing or disallowing a reproduction of text content collected by a web crawler, except for the title if specified, from the whole or parts of the content to display portions of that content. In existing implementations preference on which text can be used for caption are expressed via the NOSNIPPET statement set in http header, HTML meta tags, or HTML tags properties (data-nosnippet).

4.3. Display text length

The act of limiting the number of characters as a textual display from content collected by a web crawler. In existing implementations quotation length preferences are expressed via the max-snippet statement set in http header or HTML robots meta tags.

4.4. Exact text match

The act of limiting text content to only an exact match if displaying text content from the document. If this preference is present, text content must be quoted as is or use avoided and an explicit link back to the source of the document used in that instance. One example of existing implementation of text quotation preferences is notranslate.

4.5. Image preview

The act of limiting usage and size of images. In existing implementations image preview preferences are typically expressed via the max-image-preview statement set in http header or HTML meta tags.

4.6. Video preview

The act of limiting usage and length of videos. In existing implementations video preview preferences are typically expressed via the max-video-preview statement set in http header or HTML robots meta tags.

4.7. Generative AI training

The act of using content in training general purpose AI models that have the intent 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. In existing implementations preferences are communicated via robots.txt or via http header or HTML robots meta tags.

5. Usage

The vocabulary is used by referencing the terms defined in the section on Section 4, directly or via mappings, in accordance with how they are defined in this document.

5.1. More specific instructions

A recipient of a statement of preferences that follows this model might receive more specific instructions in two ways: Extensions to the vocabulary might define more specific categories of usage. Preferences about more specific categories override those of any more general category.

Statements of preferences are general purpose, machine-readable statements that cannot override contractual agreements or more specific statements.

For instance, a statement of preferences might indicate that the use of a digital content is disallowed for Generative AI Training. If arrangements, such as legal or business agreements, exist that explicitly permit the use of that digital content, those arrangements are likely to apply, unless the terms of the arrangement explicitly say otherwise.

5.2. Usage category labels

Each usage category in the Section 4 is mapped to a short textual label. The table below (Table 1) tabulates this mapping.

Category	Label	Reference
Indexing and retrieval	index	Section 4.1
Display text	display-text	Section 4.2
Display text length	max-text-length	Section 4.3
Exact text match	match-text	Section 4.4
Image preview	max-image-preview	Section 4.5
Video preview	max-video-preview	Section 4.6
Generative AI training	train-genAI	Section 4.7

Table 1: Usage Category Labels

An important note about this process and format is that, if the same key appears multiple times, only the last value is taken. This means that duplicating the same key could result in unexpected outcomes.

5.3. Consulting a Preference Expression

A single preference expression can be evaluated for a usage category as follows: If the expression contains an explicit preference, that is the result.

Otherwise, no preference is expressed.

5.4. Combining Preferences

The application might have multiple preference expressions, obtained using different methods.

If multiple preference expressions are active, all preference expressions are consulted as described in the section on Applicability and Legal Effect (Section 6). This might result in conflicting answers.

Absent some other means of resolving conflicts, the following process applies to each usage category: If any preference expression indicates that the usage is restricted, the result is that the usage is restricted.

Otherwise, if any preference allows the usage, the result is that the usage is allowed.

Otherwise, no preference is set.

This process ensures that the most restrictive preference applies.

6. Applicability and Legal Effect

TODO

7. IANA Considerations

This document has no IANA actions.

8. Addendum - Explanatory Note

Category	Search Experience if preference set to disallowed	AI Tool Experiences (such as Chat experience) if preference set to disallowed
Indexing and Retrieval	Content is not used or linked in response to a user search query.	Content may not be used or linked in response to a user query. Eg: Response in Copilot to the query, “Tell me what the mayor of SF said last night at city hall?” may not retrieve and use a relevant SF Chronicle article to inform a user response if this preference is set to not allowed.
Display Text	When content is shown in response to a user query, only the	Content cannot be used as a direct input to generate an AI experience (such as an AI summary or overview) in response to a user query. When content is shown in response to a user query, only the title (if specified) and URL may be displayed.

	title (if specified) and URL.	Eg: Response in Copilot to the query, “Tell me what the mayor of SF said last night at city hall?” may only display the title and URL to a SF Chronicle article if that article is delivered in the response and it will not serve as a direct input for grounding, provided the whole document is set to no display.
Display Text Length	Any display that includes a portion of the content must comply with the specified character limit.	Any display that includes a portion of the content must comply with the specified character limit. Eg: Response in Copilot to the query, “Tell me what the mayor of SF said last night at city hall?” may use a SF Chronicle article for grounding purposes to generate a response, but any passage of the article that is included as part of that response must comply with any established character limit. The response may go beyond the passage from the content and include other statements or information whether observations derived from examining the article or not.
Exact Text Match	Any display that includes a portion of the content must only present the designated portions of the content.	Any display that includes a portion of the content must comply with the specified character limit. Eg: Response in Copilot to the query, “Tell me what the mayor of SF said last night at city hall?” may use a SF Chronicle article for grounding purposes to generate a response, but any passage of the article that is included as part of that response must only include characters from the designated portion of the content.
Generative AI Training	Any text included cannot be	Any text included cannot be used for training of Generative AI models.

	used for training of Generative AI models.	

Table 2: Search and AI Tool Behavior Examples

9. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

Acknowledgments

TODO

Authors' Addresses

K. Madhavan
Microsoft Corporation
Email: krishna.madhavan@microsoft.com

F. Canel
Microsoft Corporation
Email: fabrice.canel@microsoft.com

J. Gimbel
Microsoft Corporation
Email: jordangimbel@microsoft.com

S. Cooper
Microsoft Corporation
Email: sonia.cooper@skype.net