March 2026

# The Difficult Task of Categorizing Machine Use of the Web

A Creative Commons Issue Brief: Backgrounders on topics related to AI & the commons

## Introduction

Over the past thirty years, humans have used automated programs to systematically navigate, access, and make use of web content, to do things like build search indexes or create archives. Although not entirely uncontested, this machine use was governed by relatively informal norms and standards.

Now, machines (sometimes referred to as "bots" or "crawlers") are being used to access the web to train and enable the functioning of large AI models. Machines like this now account for a large proportion of web traffic. We've described the significant technical, economic, and cultural impacts of this change in the paper *From Human Content to Machine Data*.

In this new age, simply defining the different forms of machine use of web content now in play—let alone managing it—is proving challenging. This issue brief describes why a common vocabulary for machine use is needed but proving difficult to achieve, and where definitions are currently being debated.

## Why a Common Vocabulary for Machine Use is Needed

In response to large AI models' extractive relationship with web content, new approaches to governing machine access are emerging.

Pay-to-crawl, for example, involves websites using technical systems to automate compensation for when their digital content—such as text, images, and structured data—is accessed by machines. Other approaches include preference signalling, data documentation standards (such as C2PA and Dataset Cards), platform settings (such as Bluesky's User Intents for Data Reuse), licenses (such as Copyfair Licences), and principled frameworks (such as the Alliance for Responsible Data Collection).

All of these approaches are predicated on being able to distinguish between different forms of machine use. When a machine's purpose is unclear, websites and

other publishers face a difficult decision. They can block the machine and risk not showing up in search engine results, or allow it and risk content being used in unwanted ways (such as being used to train a generative AI model). And when concepts like “training”, “indexing”, or “summarization” are used vaguely or inconsistently by publishers, the machines being targeted are unlikely to interpret permissions correctly, or act on them reliably.

Without meaningful categorization, attempts to govern AI's impact on the web also risk becoming harmful to the public interest. Measures to block large-scale, extractive data processing by commercial AI firms could end up shuttering off valuable machine use for archival, accessibility tools and academic research (including those recognized by limitations and exceptions to copyright). The result would be a binary landscape, where all machine use will get treated as either allowed or not allowed, despite the wide spectrum of legitimate, nuanced purposes it serves.

A coherent, shared vocabulary for describing and categorizing machine use is the foundation from which governance mechanisms can be made precise, enforceable, and aligned with the public interest.

## Where Different Forms of Machine Use Are Being Defined

Pay-to-crawl systems are being built using their own implicit definitions and categorizations of machine use. Cloudflare’s Pay Per Crawl, for example, currently recognizes “ai-train”, “ai-input”, and “search” as distinct categories of machine use. It recently updated the default settings for 3.8M websites that use its services, with “search” defaulting to yes, “ai-train” to no, and “ai-input” blank, indicating a neutral position.

Really Simple Licensing (RSL) is a new open standard designed to enable web publishers to define machine-readable licensing terms for their content. It adopts a different, hierarchical categorization of machine use to Cloudflare: automated processing in general (“all”), a subcategory of use by AI systems (“ai-all”), further subcategories of use by AI systems (“ai-train”, “ai-input”, and “ai-index”), and indexing for search (“search”).

Known Agents, a tool for monitoring machine visitors, adopts a more differentiated view. It recognises and keeps a list of 15 forms of “agent”, including “Achivers”, “Scrapers”, “Search Engine Crawlers”, “AI Search Crawlers”, “AI Assistants”, and “Fetchers”.

In early 2025, the Internet Engineering Taskforce (IETF) chartered the AI Preferences Working Group (AIPREF WG). The purpose of the AIPREF WG is to formalize a machine-readable vocabulary for AI usage preferences and specify how to attach those preferences to content over the web. However, recent reporting suggests that the group's progress has stalled, on account of divergent views among participants on how to disambiguate machine use into distinct categories. The group's draft vocabulary currently includes just two defined forms of machine use: "Foundation Model Production" and "Search".

The IETF has recently chartered a new Web Bot Auth Working Group (webbotauth WG). This group is related to the AIPREF WG in that it sets out to standardize methods to cryptographically authenticate automated clients. According to reporting of the group's first meeting, participants also disagreed on what kinds of machines should be differentiated, what machine behaviors should be encouraged or discouraged, and whether authentication was possible given the diversity of purposes and actors involved.

## Why Categorizing Machine Use is Difficult

Categorizing machine use of the web is difficult because the same act—in broad terms, an automated request to a URL—can involve radically different purposes, technical methods, economic stakes, and cultural implications.

### Divergent purposes and underlying motivations

Machine use has long served a wide array of purposes, including research, archival, indexing for search, translation, spam filtering, harmful language detection, and improving accessibility. Developing large AI models involves using automated programs to gather data for model training and undertake retrieval-augmented generation (RAG), and, more recently, enable "agentic" browsing and task execution. Even taken in isolation, these purposes are difficult to categorize neatly; each of them is internally diverse and can be seen as overlapping with others.

### Different technical methods, quickly changing

The same purpose for machine use can be implemented through different technical methods. This includes crawling, scraping, and text-and-data-mining (TDM), as well as machine-to-machine Application Programming Interfaces (APIs) and automated feeds like RSS. The technical landscape is evolving quickly, with new protocols

emerging for dynamically connecting AI systems with web content and services, such as the Model Context Protocol (MCP).

### Diverse operators

Machine use of digital content can be carried out by individual people, technology firms, non-profit research labs, startups, government agencies, or cultural heritage organizations, each with different obligations, capabilities, and incentives. A purpose such as "AI model training" may be viewed vastly differently when undertaken by a large technology firm for commercial exploitation versus a national library with a civic responsibility.

### Definitions are never neutral

A major fault line in discussions about definitions of machine use revolves around whether categories should be narrowly tailored to AI-related uses or integrate long-standing uses like search and archival. Some argue that separating "AI" from other machine uses is artificial; others insist that the unique nature and scale of large AI models' use of digital content warrants distinct treatment (and are the reason definitions are needed).

One person's significant difference is another's trivial detail. Definitions will ultimately encode normative stances. Decisions about what to foreground or collapse together will privilege some interests, rights, and values over others.

## Considerations

Large AI models have forced a reckoning in terms of machine access to the web. New approaches to governing machine access to digital content could have a role to play in supporting a healthier, more sustainable relationship between AI and the web. However, as we've already described with pay-to-crawl, these approaches must be built using nuanced definitions of machine use, otherwise they risk treating all machine uses the same and shuttering off access to content in the public interest. It's therefore vital that organizations with a commitment to preserving and growing the digital commons, rather than only those with an interest in commodifying it, have a voice where these definitions and categorisations are being developed.