Network Working Group
Internet-Draft
Intended status: Informational
Expires: 8 January 2026

G. Illyes
Independent
M. Kuehlewind
Ericsson

7 July 2025

# Crawler best practices
# draft-illyes-aipref-cbcp-00

## Abstract

This document describes best pratices for web crawlers.

## Discussion Venues

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

Source for this draft and an issue tracker can be found at https://github.com/garyillyes/cbcp.

## 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/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 8 January 2026.

## Copyright Notice

Table of Contents



## 1. Introduction

Automatic clients, such as crawlers and bots, are used to access web resources, including indexing for search engines or, more recently, for new artificial intelligence (AI) applications like training models. As crawling activity increases, automatic clients must behave appropriately and respect the constraints of the resources they access. This includes clearly documenting how they can be identified and how their behavior can be influenced. Therefore, crawler operators are asked to follow the best practices for crawling outlined in this document.

To further assist website owners, it should also be considered to create a central registry where website owners can look up well-behaved crawlers. Note that while self-declared research crawlers, including privacy and malware discovery crawlers, and contractual crawlers are welcome to adopt these practices, due to the nature of their relationship with sites, they may exempt themselves from any of the Crawler Best Practices with a rationale.

## 2. Recommended Best Practices

The following best practices should be followed and are already applied by a vast majority of large-scale crawlers on the Internet:

1. Crawlers must support and respect the Robots Exclusion Protocol.

2. Crawlers must be easily identifiable through their user agent string.

3. Crawlers must not interfere with the regular operation of a site.

4. Crawlers must support caching directives.

5. Crawlers must expose the IP ranges they are crawling from in a standardized format.

6. Crawlers must expose a page that explains how the crawled data is used and how it can be blocked.

### 2.1. Crawlers must respect the Robots Exclusion Protocol

All well behaved-crawlers must support the REP as defined in Section 2.2.1 of [REP] to allow site owners to opt out from crawling.

Especially if the website chooses not to use a robots.txt file as defined by the REP, crawlers further need to respect the X-robots-tag in the HTTP header.

### 2.2. Crawlers must be easily identifiable through their user agent string

As outlined in Section 2.2.1 of [REP] (Robots Exclusion Protocol; REP), the HTTP request header ’User-Agent’ should clearly identify the crawler, usually by including a URL that hosts the crawler’s description. For example:

User-Agent: Mozilla/5.0 (compatible; ExampleBot/0.1; +https://www.example.com/bot.html).

This is already a widely accepted practice among crawler operators. To remain compliant, crawler operators must include unique identifiers for their crawlers in the case-insensitive User-Agent, such as "contains ’googlebot’ and ’https://url/...’". Additionally, the name should clearly identify both the crawler owner and its purpose as much as reasonably possible.

### 2.3. Crawlers must not interfere with the normal operation of a site

Depending on a site’s setup (computing resources and software efficiency) and its size, crawling may slow down the site or even take it offline altogether. Crawler operators must ensure that their crawlers are equipped with back-out logic that relies on at least the standard signals defined by Section 15.6 of [HTTP-SEMANTICS], preferably also additional heuristics such as a change in the

relative response time of the server.

Therefore, crawlers should log already visited URLs, the number of requests sent to each resource, and the respective HTTP status codes in the responses, especially if errors occur, to prevent repeatedly crawling the same source.

Generally, crawlers should avoid sending multiple requests to the same resources at the same time and should limit the crawling speed to prevent server overload, if possible, following the limits outlined in the REP protocol. Additionally, resources should not be re-crawled too often. Ideally, crawlers should restrict the depth of crawling and the number of requests per resource to prevent loops.

Crawlers should not attempt to bypass authentication or other access restrictions, such as when login is required, CAPTCHAs are in use, or content is behind a paywall, unless explicitly agreed upon with the website owner.

Crawlers should primarily access resources using HTTP GET requests, resorting to other methods (e.g., POST, PUT) only if there is a prior agreement with the publisher or if the publisher’s content management system automatically makes those calls when JavaScript runs. Generally, the load caused by executing JavaScript should be carefully considered or even avoided whenever possible.

## 2.4. Crawlers must support caching directives

[HTTP-CACHING] HTTP caching removes the need of repeated access from crawlers to the same URL.

## 2.5. Crawlers must expose the IP ranges they use for crawling

To complement the REP, crawler operators should publish the IP ranges they have allocated for crawling in a standardized, machine-readable format, and keep this information reasonably up-to-date (i.e., should not be outdated for more than 7 days).

The object containing the IP addresses must be linked from the page describing the crawler, and it must also be referenced in the page’s metadata for machine readability. For example:

<link rel="help" href="https://example.com/crawlerips.json">

### 2.6. Crawlers must explain how the crawled data is used and the crawler can be blocked

Crawlers must be easily identifiable through their user-agent string, and they should explain how the data they collect will be used. In practice, this is usually done via the documentation page linked in the crawler's user agent. Additionally, the documentation page should include a contact address for the crawler owner.

The webpage should also provide an example REP file to block the crawler and a method for verifying REP files.

## 3. 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.

## 4. Security Considerations

TODO Security

## 5. IANA Considerations

This document has no IANA actions.

## 6. Normative References

[HTTP-CACHING]
Fielding, R., Ed., Nottingham, M., Ed., and J. Reschke, Ed., "HTTP Caching", STD 98, RFC 9111, DOI 10.17487/RFC9111, June 2022, <https://www.rfc-editor.org/rfc/rfc9111>.

[HTTP-SEMANTICS]
Fielding, R., Ed., Nottingham, M., Ed., and J. Reschke, Ed., "HTTP Semantics", STD 97, RFC 9110, DOI 10.17487/RFC9110, June 2022, <https://www.rfc-editor.org/rfc/rfc9110>.

[REP] Koster, M., Illyes, G., Zeller, H., and L. Sassman, "Robots Exclusion Protocol", RFC 9309, DOI 10.17487/RFC9309, September 2022, <https://www.rfc-editor.org/rfc/rfc9309>.

[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 acknowledge.

## Authors' Addresses

Gary Illyes
Independent
Email: synack@garyillyes.com

Mirja K端hlewind
Ericsson
Email: mirja.kuehlewind@ericsson.com