Internet Engineering Task Force
Internet-Draft
Intended status: Informational

Expires: September 24, 2026

F. Badii
Digital Medusa
Jo Levy
ARDC
S. McKenna
Sequentum

March 24, 2026

Best Practices for Responsible Web Data Collection
draft-farzdusa-webbot-datacollection-00

Abstract

The IETF develops standards and protocols to make the internet work better, adhering to principles of openness and decentralization. Industry best practices and protocols for automated web data collection have long existed, but have not been documented at the IETF.

For decades, researchers, universities, journalists, public interest groups, and commercial entities have used automated tools to access and collect public web data (sometimes referred to as data scraping, web crawling, or text and data mining) for a wide range of uses [I-D.farzdusa-aipref-enduser]. Examples of these uses include extraction of pricing information for market intelligence or to create a consumer price index, comparative real estate analysis to support underwriting of loans and mortgages, webpage archiving to preserve human knowledge, preserving government websites to hold political powers accountable, journalist research and reporting, and university and scientific research. Recently, innovations in artificial intelligence have significantly increased the automated collection of public web data, creating tensions between the use of AI to equalize and increase access to knowledge and the disruption of existing Internet models, including non-profit repositories that face increased demands for access and businesses that profit from free web access to human viewers.

This document lists a set of technical best practices that are prevalent across industries for the automated collection of public web data. It provides protocols for how automated tools access and collect publicly available web data, including volume control, transparency, documentation, and access, that can be implemented by any automated data collector. It applies principles of net neutrality to the collection of data, providing uniform guidance regardless of the identity of the data collector or website operator, the location of the collection, or the applicable legal jurisdiction.

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 September 24, 2026.

Table of Contents

1. Introduction and Terminology

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.

"Data Collectors": Persons or entities that access websites through automated devices to collect, scrape, harvest, or crawl information.

"Public Web Data": Data published on the world wide web that is not behind a restricted access log-in.

"Publicly available information": Information that a person, organization, or data collector has a reasonable basis to believe is lawfully made available to the general public.

2. Collection Limited to Public Web Data

Data collectors MUST only access web data that is not behind a paywall or restricted access log-in, or non-public data that the Data Collector has received explicit approval to collect. Data Collectors MUST NOT circumvent a restricted access log-in or paywall.

3. Identification and Transparency

Data Collectors SHOULD provide sufficient information to allow website owners to contact them to report abuses, such as a dedicated email address or link on a published webpage, or an identifiable

User Agent name or email address. Data Collectors MUST NOT affirmatively misrepresent their identities by using the name, email, or URL of another Data Collector without authorization.

4. Rate Limits

Multiple methods of rate limitation are available to protect the health of internet locations and help prevent Distributed Denial of Service (DDoS) and Denial of Service (DoS) attacks. Data Collectors SHOULD select one or more rate limitation methodologies, taking into account factors such as whether monitoring indicates a degradation of service and the scope, breadth, and other parameters of the data collection. Rate limitation methods MAY include one or more of the following:

a. Use of a static or dynamic/random download delay:

   It is possible for a script to set a static (e.g., once every five seconds) or random (e.g., a random period between 2 and 7 seconds) delay between page requests.

b. Use of "auto throttle" and similar technologies:

   Many open-source libraries for web data collection offer functionality that will automatically adjust the frequency of page requests based on the current webserver load, for instance by inferring such load based on the latency between requests and responses.

c. Calculating average daily loads:

   A number of analytics firms offer data about website traffic that would allow a data collector to calculate the number of average page loads per day. A Data Collector MAY consider applying this data when determining the frequency at which the script accesses the website(s). For example, by ensuring that the percentage of page requests it makes remains below some threshold percentage of average daily page loads.

d. Collecting data during low-traffic timeframes:

   Consider limiting data collection to less busy times based on website traffic data or inferred from the geographic location of the site and the site’s content. Applying randomness to script start times can minimize concurrent requests.

e. Limiting the number of concurrent requests:

   It is possible to keep a script from issuing additional requests until the web server responds to outstanding requests. Consider keeping the number of outstanding concurrent requests below a predefined limit.

f. Crawling at "human speed":

   If a human collecting the data by copying and pasting would load a new page once every five seconds, consider limiting scripts to a similar rate.

g. Incorporating "speed bumps":

   Similar to applying "human speed," consider including speed bumps that pause the script at certain intervals (e.g., the script pauses for 5 seconds after every 10 page loads).

h. Following the robots.txt "Crawl-delay" directive:

Website owners can specify in their robots.txt file the number of seconds that a script should wait between successive page loads.

5. Preference Signals

a. Robots.txt:

Data collectors MAY exercise discretion in deciding whether to search for the presence of a robots.txt file and, if located, whether to follow the robots.txt instructions. Data collectors MAY decide to follow certain types of robots.txt instructions (e.g., Crawl-delay), or to follow robots.txt for certain websites but not others. To promote transparency, Data Collectors SHOULD retain documentation of whether robots.txt was read and followed and, if so, under what circumstances.

b. Access Requirements:

When Public Web Data is made available to the general public only after completion of a specific action, and no pre-authorization is required, Data Collectors MAY complete the action and access the Public Web Data.

6. Log Retention

a. Query Log Content:

Data Collectors SHOULD maintain query logs for each data collection that include, at a minimum:

i. A precise date and time of the query.

ii. The target URL (the domain/URL to which the query is directed) in standard format.

iii. The source IP address (the IP used to send the request) in standard format.

iv. A unique query identifier.

b. Query Log Retention:

Retention periods for query logs SHOULD be based upon the type and frequency of the query, and SHOULD include at a minimum:

i. Live log retention sufficient to support timely responses to ongoing data collections.

ii. Archived log retention for a minimum of 3 years.

7. Out of Scope

The following topics are explicitly out of scope for this document:

1. Restating, embedding, or enforcing legal or quasi-legal restraints on the use of collected data. Laws applicable to any particular data can vary based on the specific data collected, the identity of the party collecting the data, the timing of collection and data use, the location of the data collection, and the existence of contractual agreements beyond the parameters of machine-readable computer code.

2. Operational guidance on how persons or entities that collect data manage their business, organizations, or data handling practices after Public Web Data is accessed and collected from the internet. This document does not cover best practices for Acceptable Use Policies for the automated collection of data within organizations, governance practices for handling personal information, complaints, or reports, internal operations for determining whether particular data includes copyrighted or copyrightable information and if so, the legality of copying that data, or data and SOC2 compliance practices for handling certain data after collection. For guidance on these matters, please see ARDC's Technical Standards and Governance Guidelines, Section 2.

8. Security Considerations

TODO

9. IANA Considerations

This document has no IANA actions.

10. References

10.1. 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/info/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/info/rfc8174>.

10.2. Informative References

[I-D.farzdusa-aipref-enduser] Badii, F., Bailey, J., and J. Levy, "AI Preferences Signaling: End User Impact", Internet-Draft draft-farzdusa-aipref-enduser-00, 2024, <https://datatracker.ietf.org/doc/html/draft-farzdusa-aipref-enduser-00>.

Authors' Addresses

Farzaneh Badii
Digital Medusa
Email: farzaneh@digitalmedusa.org

Jo Levy
Alliance for Responsible Data Collection
Email: josaxe@yahoo.com

Sarah McKenna
Sequentum
Email: sarah.mckenna@sequentum.com