ARDC Alliance For Responsible Data Collection

Technical Standards and Governance Guidelines for Public Internet Data Collection

ARDC Guideline 1: Technical Standards

With the rapid expansion of online digital data, it is critical to establish responsible data collection standards that provide data collectors with guidance on best practices, provide third parties with a reliable means to assess whether the public web data they seek to use has been responsibly sourced, and protect public access to public data. The technical guidelines set forth below are part of a broader framework for responsible data collection that includes guidelines for data collection governance. As such, all ARDC guidelines must be applied holistically.

1.1 **Compliance with the law.** All data collection activities must comply with applicable laws.

1.2 **Public data access.** Data collection under these standards is limited to public internet data. Collection of data that is accessible only via a restricted access log-in is not included.

1.3 **Domain health monitoring**. Domain health monitoring entails continuous monitoring of website responsiveness during collection to prevent degradation of services of the target internet location. Use domain health monitoring to identify degradations that correlate with the relevant platform traffic and implement rate limits to remediate detected degradation and prevent its recurrence.

1.4 **Rate limits**. Multiple methods of rate limitation are available to protect the domain health of internet locations and help prevent DDoS/DoS (Distributed Denial of Service/Denial of Service) attacks. Selection of rate limit methodology should take into account factors such as whether domain health monitoring indicates a degradation of service as set forth in Guideline 1.3 as well as the scope, breadth, and other parameters of the data collection. Rate limitation methods may include one or more of the following:

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

1.4.2 **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.

1.4.3 **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 in which the script accesses the website(s). For example, by ensuring that the percentage of page requests it makes is under some percentage of average daily page loads.

1.4.4 **Collecting data during low-traffic timeframes**: Consider limiting data collection to less busy times based on website traffic data or inferred from geographic location of the site and the site's content. Applying randomness to script start times can minimize concurrent requests.

1.4.5 **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.

1.4.6 **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.

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

1.4.8 **Following robots.txt "Crawl-delay" directive**: Website owners can specify a number of seconds that a script should wait in between successive page loads set forth in their robots.txt.

1.5 **Robots.txt.** Data collectors generally retain discretion to choose whether to search for the presence of a robots.txt file and, if located, whether to follow the request. Data collectors may decide to follow certain types of robots.txt directives (e.g., crawl-delay), or to follow robots.txt in certain websites, but not others. To promote transparency, retain documentation of whether robots.txt was followed and, if so, under what circumstances.

1.6 **Log retention.**

1.6.1 **Query Log Content:** Maintain query logs for each data collection that include, at a minimum:

- A precise data and time of the query.
- The target URL(the domain/URL to which the query is directed) in standard format.
- The source IP (the IP used to send the request) in standard format.
- A unique query identifier.

1.6.2 **Query Log Retention:** Set retention periods for query logs based upon the type and frequency of the query including, at a minimum:

- Live log retention sufficient to support timely responses to ongoing data collections.
- Archived log retention for a minimum of 3 year.

Last rev. March 2025

ARDC Technical Standards and Governance Guidelines for Public Internet Data Collection

ARDC Guideline 2: Governance Guidelines

With the rapid expansion of online digital data, it is critical to establish responsible data collection standards that provide data collectors with guidance on best practices, provide third parties with a reliable means to assess whether public web data has been responsibly sourced, and protect open access to public data. These governance guidelines set forth below are part of a broader framework for responsible data collection that includes important technical guidelines for public internet data collection. As such, all ARDC guidelines must be applied holistically.

2.1 Acceptable Use Policies

Acceptable use policies should be used to confirm the data collectors' commitments to comply with applicable laws related to data collection. Acceptable use policies also define additional types of activities not permitted by the data collection organization, as well as any activities that are subject to internal review or approval processes and may vary amongst data collectors.

2.2 Documentation & Record Keeping

Before beginning a data collection project, data collectors should record key collection parameters, such as the purpose of the collection, the script(s) to be used, the domains to be queried, the timing and frequency of the queries, whether robots.txt will be honored and, if so, under what circumstance. Records of data collection projects should be correlated to the data collected, used to monitor for compliance with the ARDC Guidelines and Standards, and provided with the data collection if it is transferred or sold.

2.3 Reporting Mechanisms

Reporting should include the ability of web site owners to report potential "abuse" e.g., a dedicated email/page (i.e: abuse@XYZ.com or xyz.com/abuse-reporting) published for all.

2.4 Investigation & Response Process

Data collectors should maintain internal processes for investigating and responding to external reports of abuse, government or law enforcement inquiries, and requests for information under applicable laws or regulations, guided by the principles of transparency, cooperation, and expediency.

2.5 Compliance Oversight & Monitoring

Organizations that regularly engage in data collection should maintain a program to

oversee and monitor data collection processes and conduct periodic reviews of data collection practices, compliance with the ARDC Technical Standards, and compliance with these Governance Guidelines.

Last rev. March 2025