Technical and Governance Guidelines for Responsible Data Collection

Submitted by the Alliance for Responsible Data Collection (ARDC), Jo Levy, The Norton Law Firm, jlevy@nortonlaw.com

Who is ARDC?

The Alliance for Responsible Data Collection ("ARDC") is an inter-industry alliance of thought leaders from businesses, non-profits, and academia aligned on a mission to establish responsible data collection standards and guidelines that:

1. Provide data collectors with guidance on best practices on how collections are carried out.
2. Offer third parties a reliable means to assess the responsible sourcing of data.
3. Preserve open access to public internet data and prevent data monopolization.

Participants in ARDC represent diverse data usage models and share the common goal of ensuring open access to public data within a trusted framework. Our discussions have included contributions from Author's Alliance, Bright Data, Common Crawl, OpenAI, Sequentum, Stanford CodeX, and others.

ARDC Draft Guidelines.

The ARDC Draft Guidelines are a work-in-progress resulting from robust discussions about minimum thresholds for data collection best practices, including technical standards for data collection and governance guidelines that foster transparency and accountability. Views expressed may not reflect the individual practices of any one participant. ARDC is sharing this draft to suggest a framework, reiterate the need for broad participation, and seek feedback from the IETF workshop.

Philosophy.

Long before large language models burst onto the public stage, data collectors crawled, scraped, and extracted data from publicly available internet sites for a wide range of legitimate uses. Common historical uses include market intelligence, competitive analysis, investment analysis, website archiving, and monitoring hate speech. A description of popular uses for mass extracted data is attached as Exhibit A.

The sudden explosion of LLMs and generative AI has led to a global focus on regulation of AI models and systems. While many initiatives focus on regulation of the development and use of AI (see, e.g., the EU AI Act and the U.S. Executive Order), a number of initiatives have proposed restrictions on the use of automated tools to collect publicly available internet data. This approach is overly broad and is likely to harm the public interest by preventing legitimate uses of publicly available data. The ARDC Guidelines are designed to mitigate the risk of such unintended consequences by providing voluntary standards designed to protect websites from harm, drive transparency and accountability, and allow variation and choice based on the individual use at issue. An added advantage is the ability to democratize access to data and level the playing field for organizations of all sizes to have access to public internet data.

The ARDC welcomes IAB and IETF comments, inputs, and feedback on the Draft Technical Standards for Public Internet Data Collection and the Draft Governance Guidelines for Public Internet Data Collection, below, and welcomes the opportunity to participate in the upcoming conference.

Benefits of Responsible Data Collection Guidelines

- Promote responsible behavior among data collectors.
- Enhance transparency and accountability.
- Create a flexible and adaptive framework.
- Provide a means for data users to identify responsibly sourced data sets.
- Prevent overly broad regulations that may choke legitimate uses of scraped data.
- Prevent public data monopolization.

Draft Technical Standards for Public Internet Data Collection

ARDC Guideline 1

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 important guidelines for data collection governance. As such, all ARDC guidelines must be applied holistically. These standards build upon previous work by the FISD-SIA-Alternative Data Council on Web Data Collection Considerations.

1.1 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.

1.2 Domain health monitoring. Domain health monitoring entails monitoring of website responsiveness to prevent degradation of services of the target internet location. Use domain health monitoring to identify degradation that correlates with the relevant platform traffic and implement rate limits based on time slices and geo-locations to remediate detected degradation and prevent its recurrence.

1.3 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 1.2, as well as the scope, breadth, and other parameters of the data collection. Rate limitation methods may include:

1.3.1 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 dynamic (*e.g.* random period between 2 and 7 seconds) delay between page requests.

1.3.2 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.

1.3.3 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 should 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.

1.3.4 **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.

1.3.5 **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.

1.3.6 **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.

1.3.7 **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).

1.3.8 **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.

1.4 **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 honor the request. Data collectors may decide to honor certain types of robots.txt directives (e.g., crawl-delay), or to honor robots.txt in certain websites, but not others. To promote transparency, the ARDC Governance Guidelines for Public Internet Data Collection provide guidelines for documenting the approach to robots.txt files in data collections.

1.5 **Log retention.**

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

- A precise data and time of the query, in ISO 8601 format (YYYY-MM-DDTHH:MM:SSZ).
- 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.

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

- 90 Days: Immediate queries.
- 1 Year: Medium speed queries.
- 5 Years: Deep archive queries.

Draft Governance Guidelines for Public Internet Data Collection

ARDC Guideline 2

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 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 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, conducting periodic reviews of data collection practices, compliance with the ARDC Technical Standards, and compliance with these Governance Guidelines.

Last rev. 07/29/2024

(Continued on next page.)

Exhibit A: Description of Data Collection Use Models

eCommerce: eCommerce businesses utilize public web data to gain real time information about pricing, supply chain, comparative analysis, and customer sentiment. By continuously monitoring public websites, product reviews, and social media discussions, companies can adapt their strategies in real-time. This enables them to remain competitive by adjusting prices, improving product offerings, and enhancing customer satisfaction.

Finance: In the finance sector, public web data is used to aggregate reliable investment information, identify new investment opportunities, and conduct detailed analyses of companies within a portfolio. This data includes current stock prices, financial news, regulatory filings, and social media sentiment, providing investors with a comprehensive view of the market.

Market research: Market researchers rely on public web data to gain valuable insights into their audience, consumer trends, competitive activities, and market sentiment. This data is crucial for strategic decision-making and market analysis, enabling companies to forecast trends and understand market dynamics.

Research: Public web data is a vital resource for academic and scientific research, offering a rich pool of information for exploratory studies and empirical research across all disciplines like health sciences, environmental studies, and technology. Researchers utilize this data to test hypotheses, analyze trends, and derive insights that facilitate a deeper understanding of complex issues and fosters innovation in research methodologies and outcomes. While some proposed standards differentiate between nonprofit and for-profit research, the line between for-profit and nonprofit is murky and moveable. Many for-profit organizations rely on public web data for scientific and technological developments that benefit the greater good, while some non-profits (e.g., universities) monetize their non-profit research for financial gain.

Data for AI: Public web data is indispensable for training AI models, including natural language processing (NLP) systems, predictive analysis tools, and large language models (LLMs). This data serves as the foundation for the accuracy and reliability of these applications. By feeding AI systems vast amounts of real-world information, they can learn, adapt, and make accurate predictions.

Data for good: Public web data is also a necessity for non-profit organizations, governmental bodies, and academic institutions who rely on it for value-driven research and projects. It can be used to identify and report online hate speech, archive public websites, understand and improve gender disparities in the workforce, and develop policies to mitigate the dangers social media poses to today's youth.

Last rev. 07/27/2024