ARDC June 26 Workshop Report
Keeping Public Data Public: Confronting Challenges, Constructing Solutions.

A. Executive Summary

The Alliance for Responsible Data Collection (ARDC) Workshop on *"Keeping Public Data Public: Confronting Challenges, Constructing Solutions"* brought together legal experts, technologists, researchers, and representatives from major tech firms and public interest organizations to address growing restrictions on access to publicly available internet data, improve and accelerate adoption of ARDC's Technical Standards and Governance Guidelines for Responsible Data Collection ("ARDC Standards"), and examine how robots.txt and AI preference signals are increasingly used to assert control over web data, even where legal rights may be unsettled or non-existent.

Presentations by leading legal scholars highlighted the tension between emerging data restrictions and foundational internet values like openness and transparency. Participants expressed broad support for the continued development and adoption of ARDC Standards and identified next steps including outreach to NIST, development of a machine-readable "Croissant" vocabulary for scrapers, and formation of new working groups to promote responsible data practices.

ARDC invites all stakeholders to get more involved by joining ARDC, participating in ARDC working groups, enlisting your organization in the ARDC Steering Committee, planning the next ARDC workshop, and continuing to advocate for open, equitable, and responsible access to public web data.

B. Introduction

1. The ARDC Workshop

On June 26, 2025, the Alliance for Responsible Data Collection, or ARDC, hosted a one-day workshop in San Francisco on Keeping Public Data Public: Confronting Challenges, Constructing Solutions. Representatives from non-profits, academia, large tech companies, AI developers, start-ups, and companies that provide tools for public web data collection engaged in brainstorming, dialogue and small working group discussions to address growing efforts to restrict open access to publicly available internet data. Participants included lawyers, technologists, engineers, researchers, CEOs, and founders from Google, Microsoft, Open AI, Meta, Intel, Git Hub, Mozilla, Electronic Frontier Federation, Common Crawl, Author's Alliance, Bright Data, Sequentum, Zyte, Stanford, MIT, Emory Law School, ZealStrat, BuildETH, The Norton Law Firm, Stobbs, Amazon Key and more.

2. ARDC Background

ARDC was founded on a simple premise: open access to public internet data is critically important to the world. For decades, businesses of all sizes have relied on public internet data to understand the market, develop new products, and gain customer insights. Scientists and researchers rely on public internet

data to study human behavior, develop new technologies, and benefit society as a whole. Non-profit organizations rely on public web data to monitor for illegal and unethical behavior and to hold governments and others accountable. For decades, the free flow of information on the Internet was seen as one of its greatest values.

But over time, efforts to restrict open access to internet data have increased.

ARDC's mission is to preserve open access to public internet data by creating a trusted framework for responsible scraping and crawling. Central to this mission are ARDC's Technical Standards and Governance Guidelines for Responsible Data Collection ("ARDC Standards"). These standards fill a critical gap by providing specific guidance on HOW to responsibly scrape and crawl the internet.

3. ARDC Workshop Objectives

Workshop objectives included

(1) feedback, support, and an action plan to accelerate adoption of ARDC's Standards and Guidelines for Responsible Data Collection,

(2) feedback and an action plan related to the increased development and use of preference signals by website owners or others to prevent the use of public web data for AI training or inference, and

(3) developing a community of stakeholders with diverse interests but a shared belief in the value of an open internet with equal access to publicly available internet data.

C. The State of The Internet: Legal, Regulatory, & Industry Responses to AI

Matt Sag, the Jonas Robitscher Professor of Law in Artificial Intelligence, Machine Learning, and Data Science at Emory University Law School and Dr. Robert Mahari, Associate Director of CodeX, the Stanford Center for Legal Informatics, discussed the empirical evidence of increasing restrictions on access to and use of publicly available internet data and the implications from a social and legal perspective.

As AI becomes the primary interface for information and AI summaries reduce the visitors to the source websites, the value proposition underlying the relationship between web crawlers and website owners is shifting. Before generative AI, the primary values underlying the internet were the free flow of knowledge and open access to information. These values accommodated the autonomy of rightsholders and an ecosystem that supported creative work. Robots.txt was used to enable the infrastructure to function properly, not to declare, establish, or enforce legal rights. Today, efforts to use preference signals like robots.txt to control the future use of published internet data suggests the emergence of conflict between underlying social values of open access to information, autonomy, and the creation of ecosystems that incentivize creators.

Participants raised concerns about the mismatch between the party asserting a preference signal and the actual underlying rightsholder. Studies of license assertions have established a large number of mismatches where the license asserted does not exist, or no longer exists, or is held by a different entity. Similarly, attempts to use preference signals to establish copyrights or an exception to copyright under Article 4 of the EU DSM Directive will require additional layers of proof to establish that the party

asserting the preference has a legitimate right to do so. The potential for copyright overreach will undermine the usefulness and effectiveness of preference signals.

Preference signals are one way to create voluntary norms or standards regarding the copying and use of web data. With preference signals, the website owner or developer issues the signal and relies on others to follow it.

A second type of voluntary standards and norms, the ARDC Standards, addresses the manner in which web crawlers and scrapers interact with websites and target domains. The ARDC Standards provide a baseline for responsible data collection practices to assure target domains are not harmed by public web data collection and documentation is created to provide transparency and accountability.

Croissant for Crawlers and Scrapers is one method of documentation currently under development. It will provide a machine readable vocabulary that will record the date, start, and stop time of any crawling or scraping and provide other critical information.

D. Workshop 1: ARDC Standards

During breakout sessions, five separate tables of participants studied the standards to identify any showstoppers, missing elements, missing stakeholders, concerns, and recommended next steps.

1. **Definition of Public Web Data**.

Participants questioned what was meant by "public web data." Does "public web data" simply mean anything that is not behind a paywall or restricted-access log-in? What about information that was mistakenly made public by incorrectly selecting a privacy setting? Or data whose status changed from public to private or vice versa?

ARDC Steering Committee Members explained that, as used in the Standards, public web data is intended to mean data that is accessible to the public without going behind a paywall or restricted access log-in. The way the data (manually or automated) is accessed does not alter whether it is public web data or not.

Limiting collection to public web data is a fundamental threshold for responsible data collection. Other standards and guidelines may also apply. For example, under the ARDC Governance Guidelines, an entity engaged in data scraping should have an Acceptable Use Policy that defines and limits acceptable purposes for data collection within that organization. Each organization's acceptable use policy will vary. A large university whose professors and students engage in a wide range of research may have a different Acceptable Use Policy than a small or medium business seeking pricing information on a particular line of products, an investment firm seeking market data to use for trading analyses, or a non-profit serving as a watchdog of potential hate speech.

In the future, ARDC will provide industry-specific guidelines for different types of data collections.

2. **Data Protection**

Even though data collected under the ARDC Standards will be limited to publicly available information, it may still contain personal or sensitive information. Some participants felt the ARDC Standards should address how scraped data is stored, anonymized, and processed in line with regulations such as GDPR and CCPA.

3. **User Agent Identity v Anonymity**

Participants at multiple tables raised questions around whether a data scraper or crawler should have to identify itself and its purpose. Requiring bot identification would allow website owners to engage with responsible bots, building trust in automated systems.

On the flip side, participants noted that "robots have rights, too." In other words, human users are entitled to privacy rights and to maintain their anonymity while on the internet. Bots are operated by persons or organizations and should be afforded this same protection. Participants suggested outlining scenarios where bot anonymity would be important. An example was shared of a bot collecting political speech posted by a government entity.

Several participants noted that residential proxies are core to their business model and, while not all collections require or use residential proxies, many do use them in order to avoid discriminatory treatment.

Some participants suggested that the key is not necessarily identifying the data collector's identity in real time, but in providing a way for the website owner to hold the data collector accountable for any harm to or abuse of the domain.

One participant suggested the technology used for PPID's (or publisher provided ID's) could be used to provide traceability without disclosing user identities. In other words, the PPID would result in a two-step system where the domain owner may not know the actual identity of the web crawler or scraper in real-time, but it would have the PPID which would enable them to determine the identity of the crawler or scraper.

Another table suggested using a carrot rather than (or in addition to) a stick approach. That is, if the data collector did provide its identity, or otherwise identified itself as a responsible bot or certified ARDC data collector, it would be entitled to a "fast pass" that would allow it to collect the data faster.

4. **Rate limits**

Several participants appreciated the section on rate limits. A suggestion was made that these need to be "battle-tested" to determine whether they are feasible. Most participants felt that crawling at "human speed" was not practical in most situations. ARDC Steering Committee Members explained that the intent was for data collectors to choose which rate limiting measures to adopt in different scenarios, not to require implementation of all. In some scenarios, automated data collections may need to proceed at

"human speed" in order to assure no "material non-public information" is inadvertently collected and used for stock market trading.

5. **Domain Health Monitoring & Bidirectional Communication**

Participants appreciated the section on domain health monitoring, but some noted that not all data collectors have the ability or visibility to continually monitor target domain health. Organizations seeking to minimize impact to the target domain often rely on website traffic analysis tools such as Similarweb and set their own internal measures for staying below a particular percentage of the typical website's traffic at the time of collection.

Participants noted that librarians and providers of open source content would like to maintain the openness of their databases and libraries but they have been hit with a rapid rise in requests from crawlers and scrapers that have strained their systems. What can we do to help them stay open?

One working group observed that, currently, there is no real-time communication between the data scraper or crawler and the website or domain it is accessing. That is, websites don't have an automated way to communicate in real-time whether the scraping or crawling is, in fact, causing any harm or otherwise impacting functionality. Other participants noted that, under ARDC Governance Guidelines, data collectors provide 24/7 abuse reporting.

6. **Next Steps for Adoption of Standards**.

There was general support for continuing the standardization efforts. Participants agreed that despite the evolving nature of AI, developing clear, globally applicable standards that can evolve with new technology and new laws is important.

The group discussed the challenges of driving broader adoption of ARDC Standards and identified several next steps: (1) creating an elevator pitch and concise slides for use when talking to standards bodies; (2) working with NIST to incorporate or point to ARDC Standards; (3) identifying and reaching out to other standards bodies such as IEEE's standards for AI System certification.

ARDC has been in touch with NIST in the past, before the change in administration, and had received positive feedback.

ARDC will be forming a working subgroup to drive adoption of ARDC Standards with standards bodies. Any entity or individual interested in helping drive further adoption of ARDC Standards, preparing the materials, and reaching out to NIST and other standards organizations, is encouraged to contact Jo Levy at jlevy@nortonlaw.com.

In addition, a separate group is working on creating a Croissant vocabulary for crawlers and scrapers that would enable a machine-readable format to document the key aspects of the ARDC Standards, such as date and time of crawling or scraping (which will be required in California in 2026 under AB 2013, the AI Training Data Transparency Act) and other query data.

Any entity or individual interested in helping create and drive adoption of Croissant for Web Data, please contact Greg Lindahl at greg@commoncrawl.org.

E. Workshop 2: AI Preference Signals & Robots.txt

The workshop on AI Preference Signals and Robots.txt was particularly timely. Creative Commons published its proposal for preference signals the day before the ARDC Workshop and the IETF will be discussing its proposed vocabulary for AI-preferences to be added to the robots.txt vocabulary in July.

1. Copyright Law, AI Training, and the Shift to Output Monitoring

Participants generally agreed that there is a mismatch between copyright law and generative AI. Not all data that is copied for generative AI is copyrightable. Even if some publicly available web data is copyrightable, the entity declaring an AI opt-out may not be the legal or legitimate holder of the copyright. Persons or entities that seek to enforce copyrights to which they have no right may themselves be violating copyright laws. Moreover, the application of exceptions and defenses to copyright infringement vary widely based on the jurisdiction and facts of each situation.

Building off of Professor Matt Sag's comments about global trends on copyright and generative AI, participants observed that lawmakers have signaled support for more formalized exceptions for AI training. In line with longstanding principles (like those applied to Xerox machines), participants reflected that copyright law has historically supported development of new technology when there are many positive, non-infringing uses of the technology. Participants observed a trend of judicial rulings and regulatory comments that focus their copyright analyses on the use and impact of the AI outputs—not on whether copyrighted data was accessed to create the AI model. This focus was viewed as a welcome development, particularly since restricting public access would stifle innovation and impact equitable access to human knowledge.

A participant noted that copyright law does not currently protect an artist's style and suggested that treating a distinctive style as trade dress under the Lanham Act might be more appropriate than applying copyright law to such claims. Participants also noted the need to create incentives to promote human expression while still embracing AI for productivity.

2. Internet Standards and AI Preference Signaling

a. Declarant.

Current versions of AI-signal preferencing use the term "Declarant" rather than "rights holder" to acknowledge the signaler might be the site admin, not the legal owner of the content. The group discussed questions such as, Who gets to signal preferences: web host? content platform? creator? How can crawlers and scrapers interpret this vocabulary consistently across different jurisdictions? How can we ensure these declarations don't unfairly restrict access or innovation?

b. Rights v. Sense of Entitlement. The group discussed the risk of false declarations of rights, legal ambiguity, and uncertainty by Declarants who are not rights holders. When preference signals are not tethered to legal rights, they act based on a sense of entitlement while lacking clarity on the source of the entitlement.

c. Attribution and Authority Verification.

One suggestion was to use general labels (e.g., web host, CDN, admin) to guide crawler behavior without needing legal ownership verification. Others suggested preference declarations should include attribution and authority verification, not just surface-level metadata. One participant referred to this as "transposability." In addition to representations about the identity of the Declarant, participants suggested information about jurisdiction be included to facilitate assessment of future uses in line with the Declarants' assertions and applicable laws.

Participants suggested that the burden should be on the Declarant to disclose their identity and basis for any opt-out preference. If the Declarant does not disclose their identity and a valid basis for their opt-out, there would be a presumption of open access. Declarant authentication might be within the file itself, or maintained in a separate registry with a link in the robots.txt or llm.txt or other file.

d. **Definitions of AI Training, AI Inference and AI.**

Participants generally agreed that the lack of clear definitions of the terms AI training, AI inference, and AI, coupled with the rapid evolution of AI capabilities and technology, creates subjectivity and ambiguity into any signal preferences using those terms. Participants stressed the need for AI preferences to be semantically meaningful and technically credible.

e**. Content Source Quality and Reliability.**

Participants expressed concern over the quality and reliability of content sources for AI development and stressed the need for mechanisms to audit the sources of AI training data content.

f. **TDM and robots.txt**

Participants noted that there is a shared interest amongst internet users and website providers for search to work. Creating a system of "preference signals" for AI under TDM is too broad. It creates a new legal standard without the due process and legislative scrutiny required under democratic systems.

g. **Beyond AI: Preference Signals and Scope Creep**

Participants discussed that even though current proposals for preference signals are limited to the use of public web data for AI, adoption of preference signals may lead to scope creep to use cases beyond AI. The pace at which the copying of publicly available web data for AI is growing and the sense of entitlement to control downstream uses of publicly shared information even where no copyright applies, suggests that the use of preference signals may expand beyond AI, copyright, or existing legal frameworks.

F. Other Workshop Topics

Participants brainstormed and voted on additional workshop topics for future workshops.

1. How to turbo charge open access to internet data (the flip side of how to prevent restrictions on open access to public web data). (17 votes)
2. Revenue sharing models for AI training data and public web data scraping. (Could some form of revenue sharing be used to prevent open source libraries from closing down? How do we prevent more websites erecting paywalls?) (9 votes)

3. Developing empirical research on public internet data (use, collections, & restrictions) (6 votes)
4. Ensuring high quality data. (How do we ensure the data collected is robust, diverse, and appropriate for the intended use?) (6 votes)
5. Developing Vocabulary for Data Collection Standards (5 votes)
6. How to promote adoption of ARDC Standards (5 votes)

G. Next Steps

- ARDC Workshop participants, please join ARDC membership and consider ARDC Steering Committee Membership to play a larger role in the future of responsible data collection.
- Reach out to ARDC to join the working group on Standards Adoption to engage with NIST, the AI Safety Institute, and the LINUX Foundation on AI & Data for adoption of ARDC Standards.
- Join the ARDC working group on Croissant for Web Data to translate ARDC standards to machine-readable format.
- Participate in the next ARDC Workshop. Planners and sponsors welcome!