CARIAD

Consensus Aggregated Retractable IFTAS Allowlist Denylist

Version 1.0 March 1st, 2024

Table of Contents

IFTAS

Summary

This document describes a domain denylist (“CARIAD”) IFTAS intends to curate and make available via the FediCheck service.

CARIAD is intended to provide new service providers a basic, first step in defederating the domains already widely defederated by high volume service providers. IFTAS community members, advisors, and respondents to our Needs Assessment Survey have asked for a shared knowledge of the most commonly blocked domains. The CARIAD list reflects the observable domain blocks affecting roughly 45% of all Mastodon accounts.

CARIAD cannot and will not provide safety for all possible use cases. Many communities will require additional research and resources to provide more comprehensive domain federation management. Specifically, CARIAD does not protect LGBTQ, BIPOC, BAME or other marginalised communities. Additional research will be required to better serve your membership, we recommend reviewing the Denylist Resources section below.

CARIAD is not intended to be a long term solution. It is highly recommended that new service providers first use CARIAD or a similar “minimum necessary” list to begin their domain curation activities, then explore additional resources.

Motivation

Newcomers to Mastodon service provision are often unaware of the full extent of the large number of servers with which they will immediately federate content once installation is complete. A known number of these federating servers are operated by bad faith actors, leading to documented cases of newly-created servers being overwhelmed with hate speech, illegal content, network or service abuse, and spam.

Denylist assistance is an oft-requested feature and the lack of knowledge or at-install support for denylists leads to documented cases of new administrators being overwhelmed by hate speech, trolling, network or service abuse, and spam. This highlights the need for an effective, early denylist approach to curtail the most obvious vectors for abuse and inauthentic behaviour.

As noted in The Atlantic Council’s Task Force for a Trustworthy Future Web report “Scaling Trust on the Web”:

> “*Federated spaces have many of the same propensities for harmful misuse by malign actors as mainstream platforms like Facebook and Twitter, while possessing few, if any, of the hard-won detection and moderation capabilities necessary to stop them. Each instance of a federated service can choose for itself what its governance approach will be. Community standards, content moderation, user reporting, and protecting against large-scale or coordinated campaigns of harassment or disinformation—even within an*

individual instance—require a broad array of technical, institutional, financial, and logistical competencies that federated spaces are not currently designed to support."

Securing Federated Platforms: Collective Risks and Responses[1] notes that

> *"…shareable or centralized denylists—that is, lists of instances believed to be malicious or harmful that can be blocked en masse by instance administrators and moderators—are a useful first step for knowledge-sharing among community members, while alleviating burdens on moderators to curate and block instances individually. Initial implementations of shared instance denylists could readily extend to a critical gap identified in our analysis: an inability to exchange content moderation decisions and threat information across instance boundaries."*

The high-volume providers that CARIAD observes have demonstrated that they block domains for:

- Network and service abuse - spam, malware, malicious activity;
- Illegal (in their respective jurisdiction) content - this may include (but is not limited to) sale of illicit drugs or weapons, child sexual abuse material, terroristic content, advocation of national socialism;
- Inactive domains - domains that are no longer in services but may become available for sale or otherwise transferred and re-used maliciously.

[1] Roth, Y., & Lai, S. (2024). Securing Federated Platforms: Collective Risks and Responses. *Journal of Online Trust and Safety*, *2*(2)

CARIAD 1.0 Specification

The CARIAD denylist is observational in nature, intended to be a reflection of the defederation decisions made by the largest service providers. This creates clear and obvious bias, and FediCheck allows all users to fully review the CARIAD list, IFTAS labels, and source domain notes before use.

Intended Use

CARIAD is intended to provide new service providers a basic, first step in defederating the domains already widely defederated by high volume service providers. IFTAS needs assessment respondents, community members and advisors have asked for a shared knowledge of the most commonly blocked domains. The CARIAD list reflects the observable domain blocks in place covering roughly 40% of all Mastodon accounts.

CARIAD cannot and will not provide safety for all possible use cases. Many communities will require additional research and resources to provide more comprehensive federation management. It is not intended to be a long term solution, and resources are listed below for service providers seeking to enhance their denylist to further curate the domains with which they federate.

Source Inclusion Criteria

CARIAD combines data from two sources:

1. The IFTAS Do Not Interact list, a manually reviewed list of domains labelled by harm.
2. A curated aggregation of the blocks in place on high volume Mastodon service providers. To be eligible for observation, each service provider is reviewed for the following criteria:
    a. Have been in service for at least 6 months;
    b. Have at least 3,000 monthly active users;
    c. Themselves have a demonstrable set of domain blocks already in place;
    d. Not themselves appear on the CARIAD list.

Bias

The inclusion criteria skew to a mix of North American and Western European service providers, with English as the predominant primary language. This biases the list in favour of white, global north speech and prejudices, and as such should be used only by new service providers who are comfortable reflecting the aggregated views of white, global north service providers.

Additionally, the sources represent the largest service providers, who have generally favoured preserving relationships over blocking speech and content, and therefore are less likely to take action against domains that others may consider worthy of blocking. However, large service providers are also less hesitant to block a small service with few accounts, which may lead to unintended aggregation of this bias.

As of March 2024, CARIAD observes service providers that represent roughly 45% of all known Mastodon accounts.

Domain Inclusion Criteria

To be visible in FediCheck or the domain audit file, the domain must be blocked by at least 51% of observed sources, or be present on the IFTAS DNI list. As new entries become eligible for inclusion, they are reviewed by IFTAS staff and advisors. If approved for inclusion, entries are listed with the majority recommendation. Domains listed on the IFTAS DNI list are included at the IFTAS severity level, regardless of observed sources.

The database and its associated lists will be reviewed quarterly by IFTAS advisors to ensure the criteria are working as intended, and not causing harm to any community.

Domain Exclusion Criteria

As and when domains fail to meet inclusion they will be delisted, and FediCheck will remove the listing from any subscribed server.

Appeals

In order to appeal a listing, a request must be sent from an address at the listed domain (eg: abuse@example.com) to our contact address below, with evidence that the issue has been resolved. We may verify the address by sending back a confirmation message asking for a response.

Delisting requests must be sent to the delisting email address, written in English language, in text form: delist-cariad (a) iftas (.) org

Requests are typically investigated and processed within three business days.

All delistings are free of charge.

Listing Longevity

As an observational list, all listings are observed and reported. Each individual listing will remain on the list for as long as a listing is visible on any of the sources and meets the requisite threshold.

Access

Access to the CARIAD list will be free and available via the FediCheck service. No payment is required for use of the list, nor for listing or delisting requests.

Please note: access via FediCheck is not available to (1) domains that are approved for inclusion, and (2) domains that are observed sources.

An audit file that lists domains approved for inclusion is available for review.

Versioning

This policy will be versioned, and lists published subject to this policy will bear the version identifier.

Public Feedback

Members of the public may make enquiries about the list, or raise issues, using the following methods:

- Using the IFTAS Connect Web site
- Using the IFTAS Denylists Working Group chat https://matrix.to/#/#wg-denylists:matrix.iftas.org

Further Reading

Denylist Resources

CARIAD is intended to be a new service provider's first step in obtaining recommendations for limiting or defederating third-party services. It is highly recommended that service providers research additional resources to understand the threats, mitigations, and approaches to managing domain federation. The following links may be of help in this regard.

Denylist Curators

The following curators maintain denylists as well as writing on the subject.

- https://seirdy.one/posts/2023/05/02/fediverse-blocklists/
- https://writer.oliphant.social/oliphant/the-oliphant-social-blocklist
- https://gardenfence.github.io/
- https://thebad.space/

Denylist Tools

- GitHub - eigenmagic/fediblockhole: A tool for automatically syncing Mastodon admin domain blocks

- GitHub - ineffyble/mastodon-block-tools: An attempt to list as many different projects/tools/scripts related to Mastodon & fediverse block management as possible

Server Administration Communities

Online communities are a good way to learn from experienced service providers and community managers.

- https://matrix.to/#/#space:matrix.iftas.org - IFTAS moderator community chat
- https://matrix.to/#/#mastodon_admin:matrix.org - Community admin chat
- https://matrix.to/#/#mastodon_moderation:matrix.org - Community moderation chat
- https://matrix.to/#/#local-moderators-hub:matrix.org - Locality-based service providers
- https://discord.gg/jxEMxvF7 - The official Mastodon chat space

Academic Research

These papers may be helpful in understanding the benefits and dangers of shared lists.

- Mansoux, A., & Roscam Abbing, R. (2020). Seven Theses on the Fediverse and the Becoming of FLOSS (pp. 124–140). Institute for Network Cultures and Transmediale. http://urn.kb.se/resolve?urn=urn:nbn:se:mau:diva-55221
- Zulli, D., Liu, M., & Gehl, R. (2020). Rethinking the “social” in “social media”: Insights into topology, abstraction, and scale on the Mastodon social network. New Media & Society, 22(7), 1188–1205. https://doi.org/10.1177/1461444820912533
- Gehl, R. W., & Zulli, D. (2022). The Digital Covenant: Non-Centralized Platform Governance on the Mastodon Social Network. Information, Communication & Society. https://hcommons.org/deposits/item/hc:49433/
- Rozenshtein, A. Z. (2022). Moderating the Fediverse: Content Moderation on Distributed Social Media. SSRN Electronic Journal. https://www.journaloffreespeechlaw.org/rozenshtein2.pdf
- Van Raemdonck, N. & Pierson, J. (2022) A conceptual framework for the mutual shaping of platform features, affordances and norms on social media Tijdschrift voor Communicatiewetenschap vol. 50 nr.4 pp.358-383 https://cris.vub.be/ws/portalfiles/portal/92575001/TRANSLATION_Conceptual_framework_for_interaction_of_platform_features_FINAL.pdf
- Marwick, A. E. (2021). Morally Motivated Networked Harassment as Normative Reinforcement. Social Media + Society, 7(2), 205630512110213. https://doi.org/10.1177/20563051211021378

FAQ

1. **What sources does the aggregation process draw from?**
   See Source Inclusion Criteria. IFTAS manually curates a list of service providers that are

deemed appropriate and representational. The source for CARIAD is IFTAS, not the observed sources.

2. **How are observed sources weighted?**
   No weighting is applied, other than the source inclusion criteria, which precludes all but the largest providers from participating.

3. **What happens when a CARIAD source blocks an instance because it appears on the CARIAD list?**
   A potential outcome of this activity is a number of domains may slowly progress to 100% observed agreement. This is an anticipated possible outcome that we will monitor and act on as needed. Of note, the observed sources at time of writing only agree on one single domain, and fewer than ten percent of all observed domain blocks are shared by more than half the sources. Source instances are not able to use FediCheck to import the CARIAD database.

4. **Does IFTAS perform manual review?**
   Each domain is reviewed for inclusion by IFTAS staff or advisors before being made available to FediCheck. In a 60 day pilot, IFTAS observed three new domains being added to the list; two for illegal content, one for spam.

5. **What are the reasons or labels?**
   The source reasons and IFTAS labels are visible in FediCheck. IFTAS uses a common vocabulary for the DNI list, and may, in the future, undertake to label additional domains and/or incorporate labels from trusted flaggers.

6. **Why “CARIAD”?**
   Cariad is the Welsh word for love. We believe helping create and preserve safety is an act of love.

Labels

For reference, the labels used by IFTAS are reviewable at
https://github.com/iftas-org/resources/tree/main/LABELS