Federated Moderation Wishlist

IFTAS is researching quick wins for Mastodon/Activitypub moderation tooling that can improve trust and safety, and reduce inauthentic activity.

Laundry List

Add anything you can think of

Feature/Function	Likelihood of/Barriers to Adoption	Level of Effort to develop
Audit logs for domain, IP and server blocks add/edit/remove		
Bulk import toxic IP blocks, toxic email domain blocks, or third party tool to do so over API		I'd like to be able to auto-import third party lists
Federating (push) or allowing pull for findings so multiple servers can "subscribe" to things like a blocklist on a particular server, IP blocks, email domain blocks from a sentinel server e.g. iftas.org or fedifence.social - this imagines a trusted service that can be		

queried by any server.		
Relays/fedi feeds for moderation logs, “X account/instance silenced/banned by Y instance” - machine readable could be used to grow next generation of connected moderation tools or part of the moderator workflow		
Bubbling/Allowlisting/Trusted Networks - allow servers to trust other servers, can share blocks, and moderators.		
Standard name exclusions for usernames (tech terms and offensive terms) eg https://gist.github.com/theskumar/54be20713e53d418bf02 or https://gist.github.com/jamiew/1112488	High	Low
Spam/inauthentic account reductions, eg		

captcha and IP trust score at signup, option to integrate third party service for this sort of thing (eg ipqualityscore, akismet)		
Shareable or subscribable member blocklists, end user to end user. Hotly debated https://github.com/mastodon/mastodon/issues/10304		
CSAM scanning, could be automated/all or only on reported images (“send to IFTAS” eg)	High	High
Malicious link checking (eg https://www.ipqualityscore.com/threat-feeds/malicious-url-scanner)		
Plugins to allow misinfo/disinfo scanning and annotation (“this content has been flagged as x or y”)		
"request edit/redraft" tool, where a post is hidden until the		

person complies with, say, adding a content warning or removing an offensive word		

Prioritising/First 90

What low-hanging fruit can the group begin to specify that can lead to clear Issues being added to Mastodon GitHub for community discussion and implementation? This group can prioritise based on

1. Likely to be well-received by broad community
2. Minimally invasive to Mastodon codebase
3. Have a significant impact for a large proportion of Mastodon servers

What can we do to explore federating moderation, either through protocol (push) or subscription (pull)?

Feature/Function	Notes/Comments