



Nota Metodológica¹

Painel "Discord na DSA Transparency Database": fonte de dados, dataset e metodologia

Data: 8 de agosto de 2026

Objeto: Descrição técnica da DSA Transparency Database enquanto fonte de dados públicos; do dataset integral de statements of reasons (SoRs) do Discord nela contido (compreendendo o período de 09/07/2024 a 06/08/2026); da arquitetura do painel analítico construído sobre esse dataset; e da metodologia de aquisição, processamento e validação empregada.

Natureza: Documento metodológico e de infraestrutura de dados. A análise jurídico-regulatória do comportamento de reporte de plataformas é objeto de documento próprio e não é tratada aqui.

Artefatos associados: painel interativo (arquivo autocontido discord-dashboard.html e versão [publicada](#)), dataset compacto de auditoria (discord-compact.csv.gz), scripts de reprodução (fetch_discord.py, extract_process_discord.py, build_discord.py) e planilha de referência externa (site-listing.csv).

1. Apresentação

Este documento descreve a construção de um painel analítico independente sobre a integridade dos **statements of reasons** (SoRs) submetidos pela plataforma Discord à **DSA Transparency Database**, base pública mantida pela Comissão Europeia. O trabalho compreendeu quatro etapas: (i) estudo da fonte oficial e de seus canais de acesso; (ii) aquisição da série histórica completa de *dumps* (arquivos diários de exportação) da plataforma; (iii) processamento em fluxo (*streaming*) de 25,7 milhões de registros individuais, com validação adversarial independente dos agregados; e (iv) construção de um painel estático autocontido, publicável e reproduzível. Cada etapa está documentada nas seções seguintes, com as referências consolidadas na Seção 7.

¹ Documento concebido por especialista, produzido com auxílio de ferramentas de Inteligência Artificial Generativa e Agêntica, seguindo instruções parametrizadas sob supervisão humana. Revisado e validado por Thiago Tavares. Em caso de dúvida, escreva para comunicacao@safernet.org.br

2. A DSA Transparency Database

2.1 Base jurídica e finalidade

O Regulamento (UE) 2022/2065 (Digital Services Act -DSA) [1] obriga os provedores de plataformas online a fornecer ao destinatário afetado uma **fundamentação** ("statement of reasons", art. 17) sempre que impuserem restrições a conteúdos ou contas, e a submeter essas decisões à Comissão Europeia, "sem demora injustificada", para inclusão em **base de dados pública** legível por máquina (art. 24(5)). A obrigação aplica-se, desde **17 de fevereiro de 2024**, a todas as plataformas online que não se qualifiquem como micro ou pequenas empresas [15]; as plataformas de muito grande porte (VLOPs) já estavam sujeitas a ela desde o fim de agosto de 2023 (quatro meses após a designação de 25 de abril de 2023), embora a base só tenha entrado em operação semanas depois. A base DSA Transparency Database [2] foi lançada pela Comissão Europeia em **25 de setembro de 2023** [15] e recebe submissões contínuas via API ou formulário web; sua finalidade declarada é permitir o escrutínio público das decisões de moderação de conteúdo e o acompanhamento da difusão de conteúdos ilegais e nocivos on-line.

2.2 A unidade de registro: o statement of reasons

Cada registro da base corresponde a uma decisão restritiva individual, descrita por um esquema de 38 campos [3], dos quais se destacam:

- quatro grupos de decisão, não excludentes entre si no esquema: restrições de visibilidade do conteúdo (remoção, desabilitação de acesso, rebaixamento, restrição etária, restrição de interação, rotulagem, outra); restrições monetárias; suspensão ou encerramento da provisão do serviço (parcial ou total); e suspensão ou encerramento da conta;
- o fundamento da decisão (*decision_ground*): ilegalidade do conteúdo ou incompatibilidade com os termos e condições, cada qual com campos próprios de fundamentação em texto livre;
- a categoria da violação alegada (*category*) e *keywords* de especificação (*category_specification*);
- o tipo de conteúdo atingido (*content_type*, multivalorado, com campo livre *content_type_other*);
- a origem da detecção (*source_type*): iniciativa voluntária da plataforma, notificação do art. 16, notificação de *trusted flagger* (art. 22) ou outro tipo de notificação;
- o grau de automação da detecção (*automated_detection*) e da decisão (*automated_decision*: totalmente, parcialmente ou não automatizada);
- datas de aplicação da restrição (*application_date*) e de submissão à base (*created_at*), além do escopo territorial, idioma do conteúdo, identificador interno (*puid*) e campos acessórios.

Em 1º de julho de 2025, a terminologia das categorias foi revista para harmonização com as regras de relatórios de transparência do DSA [10]; registros anteriores preservam, nos dumps, a nomenclatura vigente à época da submissão (fato relevante para séries históricas - ver Seção 3.3, item 5). Na mesma data, o esquema dos dumps ganhou a coluna *content_id_ean*, totalizando as atuais 38 colunas.

2.3 Arquitetura da base e canais de acesso

A aplicação oficial é um sistema Laravel (PHP), de código aberto [7], operado pela DG CONNECT. Conforme o repositório público, o armazenamento primário combina banco relacional (MySQL na documentação; a exportação de dumps lê de conexão PostgreSQL, tabela `statements_beta`) com índice Elasticsearch para busca e agregações; a infraestrutura roda em nuvem AWS (Laravel Vapor), com autenticação institucional EU Login (CAS) para submissão. Rotinas agendadas diárias geram os produtos públicos de dados: às 06:00 CET, a cadeia de exportação dos dumps diários (lotes encadeados de jobs: exportação CSV em partes de até 1 milhão de registros → zips agrupados → checksums SHA-1 → cópia para *object storage* → registro na página de download); entre 06:05 e 06:15, o congelamento dos agregados diários; a base viva aplica política de retenção de aproximadamente 6 meses (rotina de remoção de `statements` com mais de 181 dias), de modo que o acervo histórico integral reside exclusivamente nos dumps [9].

O público acessa os dados por cinco canais:

1. Dashboard oficial [2]: relatório Microsoft Power BI incorporado ("publish to web") em página institucional, com seis páginas de visualizações agregadas e atualização diária;
2. Busca de `statements` individuais, com filtros, limitada ao período de retenção da base viva;
3. Data Download [5]: dumps diários zipados por plataforma e globais, em versões full (esquema completo) e light (sem os campos de texto livre longos nem escopo territorial), além de Parquet para a versão global; URLs estáveis por nome de arquivo (`sor-{slug}-{AAAA-MM-DD}-{full|light}.zip`), servidas por redirecionamento a URLs pré-assinadas de 60 minutos;
4. Agregados diários (`aggregates-{AAAA-MM-DD}.{csv|json}`): contagens por combinação de dimensões categóricas, produto que alimenta o dashboard oficial;
5. Research API [13]: consulta programática (linguagem de consulta do Elasticsearch — DSL — ou SQL) mediante credenciamento, limitada ao período de retenção.

A Comissão publica ainda o pacote Python `dsa-tdb` [6], ferramental oficial para download, conversão e agregação dos dumps em escala

.

2.4 Licenciamento

Os dados são disponibilizados sob Creative Commons BY 4.0 [8], com citação canônica pelo DOI 10.2906/134353607485211 [11].

3. O dataset do Discord

3.1 Identificação e cobertura

Parâmetro	Valor
Plataforma (razão social na base)	Discord Netherlands B.V.
Identificador interno (platform_id)	59
Slug dos arquivos	discord-netherlands-bv
Primeiro arquivo diário	09/07/2024
Último arquivo considerado	06/08/2026
Arquivos diários adquiridos	759 (100% da série publicada)
Volume bruto adquirido	1.228 MB (zips <i>full</i>)
Registros únicos (dedup. por UUID)	25.674.022
Duplicados descartados	134
Dias com atividade	706 (de 716 desde o primeiro SoR)

O Discord ingressou na base em 09/07/2024, mas os primeiros 43 arquivos diários estão vazios: o primeiro statement of reasons data de 21/08/2024 — um único registro naquele dia. A partir de então, o reporte é praticamente contínuo (apenas 10 dias sem envios em 716).

3.2 Propriedades estatísticas centrais

Os valores abaixo foram verificados por recontagem independente (Seção 5.4).

Volume e trajetória. Média de 36,4 mil SoRs por dia ativo. Rampa de crescimento até o pico de janeiro/2025 (2.436.276 SoRs no mês), decaimento gradual ao longo de 2025 (2,3 milhões em fevereiro; 2,1 em março; 2,0 em abril; vale de 1,36 em maio; 1,9 em junho; 1,8 em julho; 1,08 em agosto), queda abrupta de 65% em setembro/2025 (374.737, ante agosto), vale em novembro/2025 (287.341) e retomada parcial em 2026 (entre 503 mil e 815 mil por mês). Nove dos doze maiores dias da série concentram-se entre agosto/2024 e março/2025 e, na maioria deles, uma única categoria responde por mais de três quartos do volume (em três, a categoria líder fica entre 51% e 59%): o recorde é 31/01/2025, com 510.564 SoRs (87,4% de proteção de menores); em 24/08/2024 — três dias após o primeiro envio — foram 303.039 SoRs, 100% de proteção de menores.

Categorias. Proteção de menores lidera com 10.624.047 SoRs (41,4%), seguida de golpes e fraude (6.504.481; 25,3%) e "comportamento não consensual" (2.556.989; 10,0%, categoria da terminologia antiga). A *keyword* específica mais frequente é material de abuso sexual infantil (4.374.012 SoRs), seguida de compartilhamento não consensual de imagens (2.159.031) e phishing (2.135.556).

Natureza das decisões. O Discord nunca combina grupos de decisão: cada statement aciona exatamente um. A ação dominante é a suspensão da provisão do serviço (18.419.086; 71,7%), quase sempre parcial (18.061.343 vs. 357.743 totais); seguem-se as ações sobre conteúdo (4.205.804; 16,4% apenas remoção, 4.084.877, e rotulagem, 120.927) e sobre a conta (3.049.132; 11,9%, encerramento em 93,5% dos casos, 2.851.885, nenhum com prazo de término). **Não há uma única restrição monetária na série.** O alvo declarado é uma conta de usuário em 85,4% dos casos: o campo

content_type_other contém um único rótulo — "User Account" — em todos os seus 21.928.338 preenchimentos.

Fundamentação. 100% das decisões invocam incompatibilidade com os termos e condições; em 53,6% (13.754.906) a plataforma marca que o conteúdo também seria ilegal. O campo de fatos e circunstâncias (decision_facts) contém um único texto, idêntico byte a byte, nas 25.674.022 decisões; o fundamento contratual (incompatible_content_ground) usa 27 formulações distintas, correspondentes às regras das Community Guidelines — a mais citada: "You broke Discord's rules regarding Child Safety content" (8.867.309).

Origem e automação. 99,85% das decisões decorrem de iniciativa voluntária (25.635.785); 38.237 (0,15%) de notificações do art. 16; **zero de trusted flaggers** ou de "outro tipo de notificação". A automação é binária: detecção automática implica decisão totalmente automatizada (7.819.031 casos, com apenas 381 exceções), detecção não automática implica decisão humana em 97,5% dos casos (17.412.476), e não existe nenhuma decisão "parcialmente automatizada" na série. No agregado, 32,2% das decisões são totalmente automatizadas (8.261.165).

Tempestividade do reporte. O intervalo entre a aplicação da restrição e a submissão à base é zero dias na mediana, no percentil 90 e no percentil 99 (99,3% das decisões reportadas no mesmo dia; máximo observado: 29 dias).

Campos sistematicamente vazios ou constantes. Idioma do conteúdo: nunca informado. Tipo de conta: nunca informado. Escopo territorial: os 30 países do Espaço Econômico Europeu (EEE), em bloco, em 100% dos registros. Identificadores puid: 25.674.009 valores distintos (apenas 13 repetidos, todos em duplicidade simples).

Ritmo operacional. Padrão de expediente: pico às quartas-feiras e às 15h (horário do servidor), fim de semana com metade do volume dos dias úteis.

3.3 Qualidade dos dados e anomalias documentadas

1. **Linha duplicada na listagem oficial:** a página de download exibe duas linhas idênticas para 01/06/2025 (mesma contagem, 26.037) — duplicidade de registro na interface, sem efeito sobre os arquivos.
2. **Subcontagem na listagem oficial em 08/05/2026:** a listagem informa 3.305 SoRs para o dia, mas o próprio arquivo diário correspondente contém **12.454** registros únicos. A diferença (9.149) explica integralmente a divergência entre o total da listagem (25.664.873) e o total real dos arquivos (25.674.022).
3. **Registros na fronteira do dia:** 191 SoRs têm created_at divergente da data do arquivo em que foram publicados; na comparação diária com a listagem oficial, isso se manifesta como 15 pares de dias adjacentes com diferenças simétricas (+N/−N) que se anulam.
4. **Duplicatas entre arquivos:** 134 UUIDs aparecem em mais de um arquivo diário; foram contados uma única vez.

5. **Transição de terminologia (01/07/2025):** categorias da nomenclatura antiga (por exemplo, "NON_CONSENSUAL_BEHAVIOUR" e "SCOPE_OF_PLATFORM_SERVICE") cessam em junho/2025 e não são convertidas retroativamente nos dumps; séries por categoria devem tratar o ponto de quebra explicitamente.

4. O painel: arquitetura do template e conteúdo

4.1 Princípios de projeto

O painel é um arquivo HTML único e autocontido (~0,3 MB), sem dependências externas de qualquer natureza (scripts, fontes, CSS ou dados remotos), atendendo ao requisito de portabilidade (funciona offline) e de compatibilidade com a política de segurança de conteúdo do ambiente de publicação. Os dados agregados são injetados no próprio arquivo como bloco JSON inline; todos os gráficos são SVG gerados por JavaScript próprio (sem bibliotecas de visualização).

4.2 Sistema visual

- **Temas claro e escuro** definidos por *tokens* CSS (custom properties), com resposta tanto à preferência do sistema operacional (*prefers-color-scheme*) quanto ao seletor manual do ambiente de publicação (*data-theme*), este com precedência;
- **Paleta categórica de cinco posições em ordem fixa**, validada computacionalmente nos dois temas para deficiência de visão de cores (simulação de protanopia/deutanopia pelo modelo de Machado, Oliveira & Fernandes, 2009 [12], com distância mínima ΔE em espaço OKLab), banda de luminosidade, piso de croma e contraste contra a superfície; onde o contraste de uma série é inferior a 3:1 no tema claro, a mitigação exigida (rótulos de valor visíveis e vista em tabela) está presente;
- **Marcas finas e chrome recessivo**: barras ≤ 24 px com extremidade arredondada, linhas de 2 px, grades em linha-fina sólida, rótulos de valor seletivos; segmentos empilhados separados por vãos de 2 px na cor da superfície;
- **Interatividade não excludente**: *crosshair* e *tooltip* em todas as séries; cada cartão possui alternância **gráfico** $\rightleftharpoons$ **tabela**, de modo que nenhum valor dependa do *hover* para ser lido (requisito de acessibilidade).

4.3 Conteúdo analítico

O painel organiza-se em: faixa de indicadores (total, primeiro envio, alvo predominante, automação, proteção de menores); série diária completa com marcação de períodos sem envio; volume mensal; categorias e sua evolução mensal (100% empilhado); combinações de grupos de decisão; decisões sobre contas (incluindo presença/ausência de prazo); ações sobre conteúdo; matriz de automação e série mensal de decisões totalmente automatizadas; origem da detecção; *keywords*; tipos de conteúdo e rótulos livres; formulações contratuais citadas; histograma de atraso; ritmo semanal/horário; maiores

dias com categorias dominantes; e um cartão de lacunas de metadados. O rodapé documenta fonte, licença, resultado da verificação independente e comandos de reprodução.

5. Metodologia usada durante a execução do projeto

5.1 Reconhecimento da fonte

1. Estudo do site oficial, da documentação da API [3], da FAQ [4] e do repositório de código da base [7], estabelecendo: o esquema de 38 campos, os padrões de URL estáveis dos dumps, o mecanismo de redirecionamento a URLs pré-assinadas e a rotina diária de geração dos arquivos.
2. Identificação do dataset alvo na página de download [5] (filtro `platform_id=59`): faixa de datas publicada (09/07/2024–06/08/2026) e sondagem preliminar da ordem de grandeza dos arquivos (43 KB no primeiro dia; no acervo completo, mediana de ~1,2 MB e faixa usual de ~0,4 a ~3,2 MB conforme a época).
3. Raspagem da listagem oficial completa (16 páginas) para constituir referência externa independente: por dia, a contagem de SoRs informada pelo site (`site-listing.csv`, 759 dias após deduplicação de uma linha repetida na origem).

5.2 Aquisição

Script `fetch_discord.py`: download paralelo (12 conexões) dos 759 zips *full* pela rota estável do site, com validação da assinatura do formato ZIP (bytes iniciais "PK"), três tentativas por arquivo, e *cache* incremental (execuções subsequentes baixam apenas o que falta). Fallback documentado para o bucket público legado nos casos em que a rota do site retorna 404 para datas antigas. Resultado: 759/759 arquivos, 1.228 MB.

5.3 Extração e agregação em fluxo

Script `extract_process_discord.py`, passada única sobre os 759 zips (estrutura aninhada: zip do dia → partes `.csv.zip` → CSVs), sem materializar os textos livres em disco:

- deduplicação global por UUID (identificador único universal de cada registro; 134 descartes);
- normalização de prefixos de enumeração e tratamento dos campos multivalorados (arrays JSON e concatenações `__` presentes nos dumps);
- acumulação de ~40 contadores (categorias, decisões, combinações, automação, origem, *keywords*, tipos, escopo, idioma, ritmos, histograma de atrasos, formulações de texto com teto de cardinalidade, cardinalidade de `puid`, séries diárias e mensais por dimensão);
- emissão simultânea de: `discord-daily.csv` (série diária), `discord-compact.csv.gz` (uma linha categórica por SoR, 33,6 MB — insumo de auditoria) e `data/discord-dashboard.json` (254 KB — agregados consumidos pelo painel).

Tempo total de processamento: da ordem de 15 minutos em estação de trabalho (14 núcleos, 48 GB), sem uso de dependências externas (apenas biblioteca-padrão Python).

5.4 Validação adversarial independente

Quatro agentes de verificação independentes (rotinas automatizadas assistidas por IA), instruídos a escrever seus próprios procedimentos de contagem sem consultar o código original, executaram em paralelo:

- A — Recontagem integral sobre `discord-compact.csv.gz`: total, categorias, matriz de automação, combinações, decisões de conta, origem, 25 totais mensais, atrasos e escopo — 22/22 verificações idênticas;
- B — Recontagem bruta amostral sobre 8 zips originais relidos do zero (incluindo casos extremos: dia com 0 SoRs, dia com 1 SoR, dia recorde) — 10/10 idênticas;
- C — Confronto externo: (i) série diária vs. listagem oficial raspada: 728 dos 759 dias idênticos; 30 dias divergem em pares adjacentes simétricos que se anulam; 1 dia diverge por subcontagem da própria listagem (Seção 3.3); (ii) três datas confrontadas com os agregados diários oficiais da base: coincidência exata nas três;
- D — Consistência interna do JSON publicado (17 identidades de soma entre séries, dimensões e metadados) — 17/17 satisfeitas.

Resultado consolidado: 49/49 recontagens internas exatas; divergências externas integralmente atribuídas e documentadas em favor dos arquivos-fonte.

5.5 Construção e publicação

Template preenchido por injeção do JSON (`build_discord.py`); verificação em navegador (console limpo, ambos os temas, alternâncias gráfico/tabela, *tooltips*); publicação como página autocontida [14] e entrega do arquivo local. A narrativa dos cartões usa exclusivamente números que passaram pela validação da Seção 5.4.

5.6 Reprodução

```
python3 fetch_discord.py           # aquisição incremental dos dumps
python3 extract_process_discord.py # extração + agregação em fluxo
python3 build_discord.py           # geração do painel autocontido
```

6. Limitações

1. O dataset reflete o que a plataforma declara à base — não é medida independente da moderação efetivamente praticada; ausências (por exemplo, zero *trusted flaggers*) são fatos do reporte, não necessariamente da operação.
2. A agregação diária usa a data do arquivo em que o registro foi publicado; 191 registros situados na fronteira do dia divergem em ± 1 dia da data de submissão.
3. Campos de texto livre foram tratados por formulações distintas com teto de cardinalidade; no caso concreto, nenhum teto foi atingido (27 formulações contratuais; 1 texto de fatos).

4. A comparação com a listagem oficial depende de raspagem de páginas HTML sujeitas a mudança de estrutura.
5. Dados posteriores a 06/08/2026 não estão cobertos; a reexecução dos scripts incorpora novos dias de forma incremental.

7. Referências

- [1] União Europeia. **Regulamento (UE) 2022/2065** do Parlamento Europeu e do Conselho, de 19 de outubro de 2022 (Digital Services Act). EUR-Lex: <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
- [2] Comissão Europeia. **DSA Transparency Database**. <https://transparency.dsa.ec.europa.eu/> (dashboard oficial em /dashboard)
- [3] Comissão Europeia. **DSA Transparency Database — API Documentation** (esquema completo do statement of reasons). <https://transparency.dsa.ec.europa.eu/page/api-documentation>
- [4] Comissão Europeia. **DSA Transparency Database — Questions and Answers (FAQ)**. <https://digital-strategy.ec.europa.eu/en/faqs/dsa-transparency-database-questions-and-answers>
- [5] Comissão Europeia. **DSA Transparency Database — Data Download**. <https://transparency.dsa.ec.europa.eu/explore-data/download> (listagem do Discord: ?platform_id=59)
- [6] Comissão Europeia. **dsa-tdb — ferramental oficial em Python para os dados da base**. <https://dsa.pages.code.europa.eu/transparency-database/dsa-tdb/index.html>
- [7] Comissão Europeia (DG CONNECT). **Repositório de código-fonte da DSA Transparency Database** (GPLv2). <https://github.com/digital-services-act/transparency-database>
- [8] Comissão Europeia. **DSA Transparency Database — Data License** (CC-BY 4.0). <https://transparency.dsa.ec.europa.eu/page/data-license>
- [9] Comissão Europeia. **DSA Transparency Database — Data Retention Policy**. <https://transparency.dsa.ec.europa.eu/page/data-retention-policy>
- [10] Comissão Europeia. **Harmonised transparency reporting rules under the Digital Services Act now in effect** (mudança de terminologia de 01/07/2025). <https://digital-strategy.ec.europa.eu/en/news/harmonised-transparency-reporting-rules-under-digital-services-act-now-effect>
- [11] Comissão Europeia. **DSA Transparency Database — citação canônica**. DOI: 10.2906/134353607485211. <https://doi.org/10.2906/134353607485211>
- [12] MACHADO, G. M.; OLIVEIRA, M. M.; FERNANDES, L. A. F. **A Physiologically-based Model for Simulation of Color Vision Deficiency**. IEEE Transactions on Visualization and Computer Graphics, v. 15, n. 6, p. 1291–1298, 2009.
- [13] Comissão Europeia. **DSA Transparency Database — Research API**. <https://transparency.dsa.ec.europa.eu/page/research-api>

[14] Painel "Discord na DSA Transparency Database" (este projeto): <https://claude.ai/code/artifact/3966473a-3a97-407a-96a1-b2bd17065097> — scripts e dados de auditoria em dsa-dashboard/discord/.

[15] Comissão Europeia. **Digital Services Act: Commission launches Transparency Database** (notícia oficial de lançamento, 26/09/2023). <https://digital-strategy.ec.europa.eu/en/news/digital-services-act-commission-launches-transparency-database>

Documento elaborado com apoio de agentes de processamento e verificação automatizados assistidos por IA; todos os valores citados foram objeto da validação descrita na Seção 5.4.