



Unlocking Transparency in Platforms' Content Moderation Activities

**DIGITAL
SERVICES ACT**

Introducing *dsa_tdb*, a Python Package for Analyzing the Digital Services Act Transparency Database



FOSDEM '25

Overview of the DSA

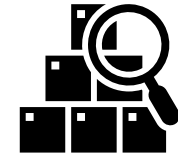
The Digital Services Act aims at creating a safer digital space:



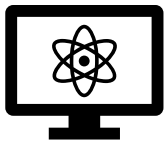
Transparent
Terms and Conditions



Consumer protection
and personal rights



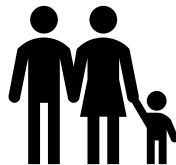
New transparency and
data access provisions



Explain, risk assess
& test algorithms
(Rec. Sys.)



Content
moderation
policies



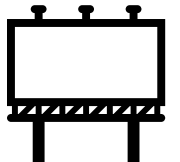
Protection of
minors
(no targeting)



Notice and
action to report
illegal content



Seller
information



Advertisement
details

Transparency features of the DSA

Transparency Reports



Bi(annual) statistics on content moderation, incl. accuracy, speed & human resources


Transparency Database



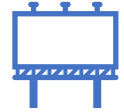
near real time content moderation seismometer

Terms & Conditions



 **OPEN TERMS ARCHIVE**
Clear and Transparent Language of the T&C

Ad Library



Repository of the ads hosted by the platform

Risk Assessments



Analysis of algorithmic risk factors

Independent Audits



Test of algorithmic systems

Data Access



Study of systemic risks

Whistleblower tool

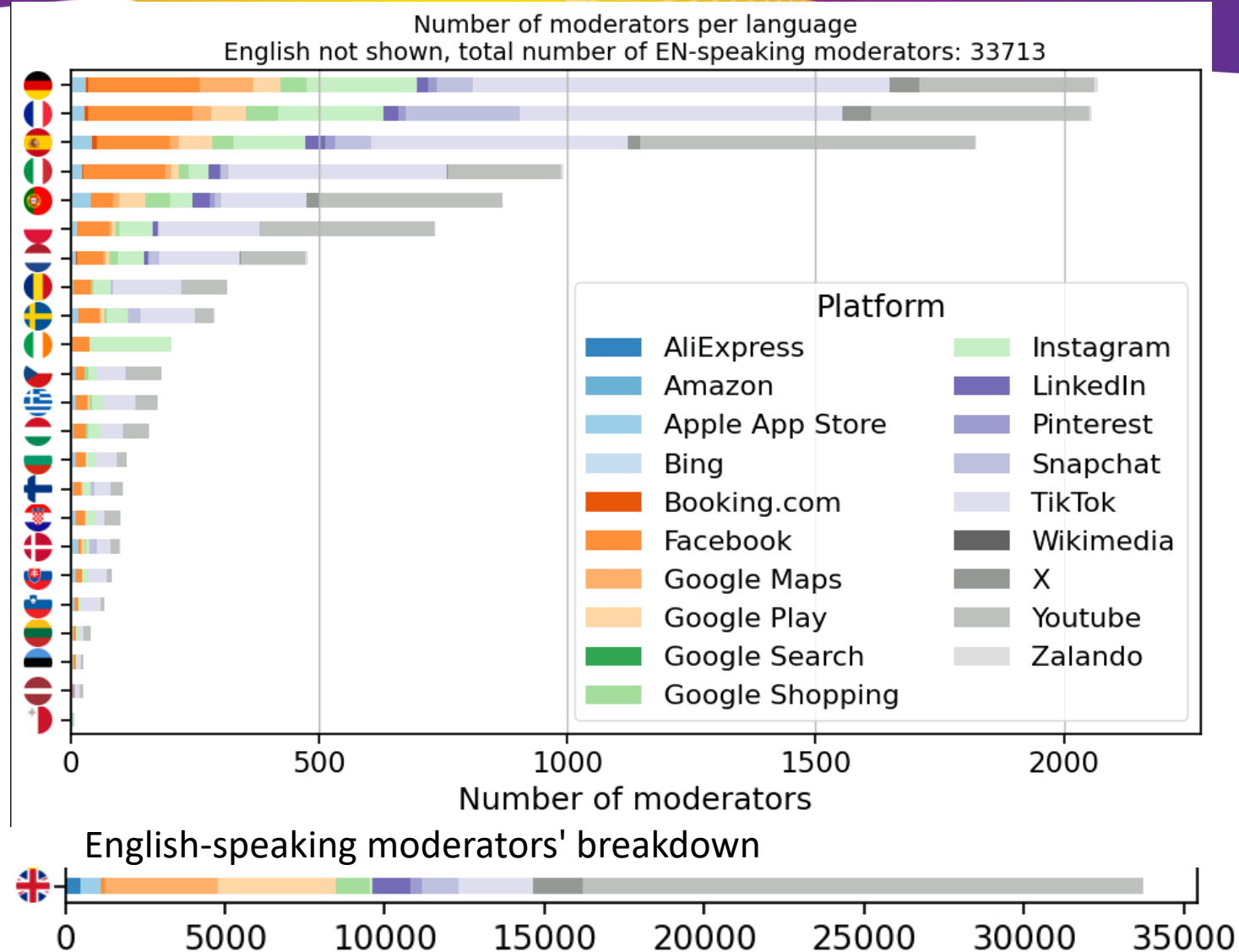


Employee and stakeholders can anonymously report bad practices and infringements

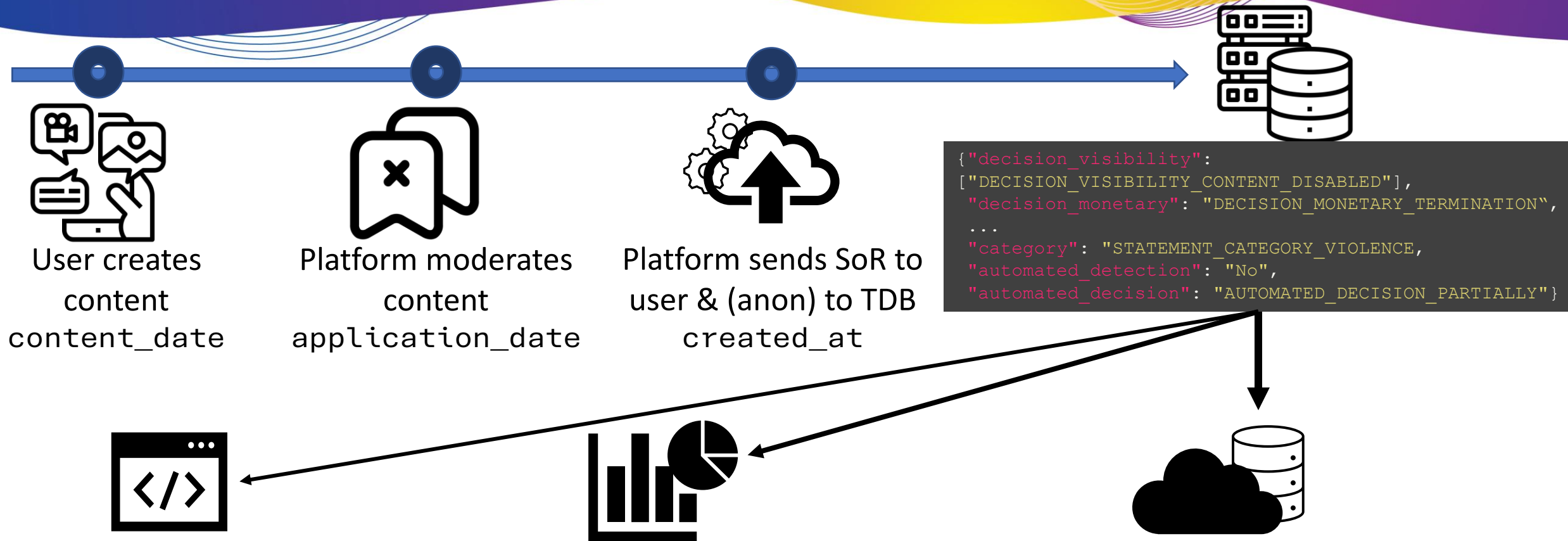
Transparency Reports

- Monthly Active Users (MAU)
- Content moderation Human resources
- Accuracy of content moderation and appeal

So far, 3 rounds of VLOPs Transparency Reports (Nov 23, May and Nov 24)



Transparency DB: data lifecycle



From website search

- Near Real time (24h delay)
- Limited to 10'000 (1'000 csv)
- **Only last 6 months**

Online dashboard

- Near real time (24h delay)
- **Aggregate view of the data**
- **No export / no customisaton**

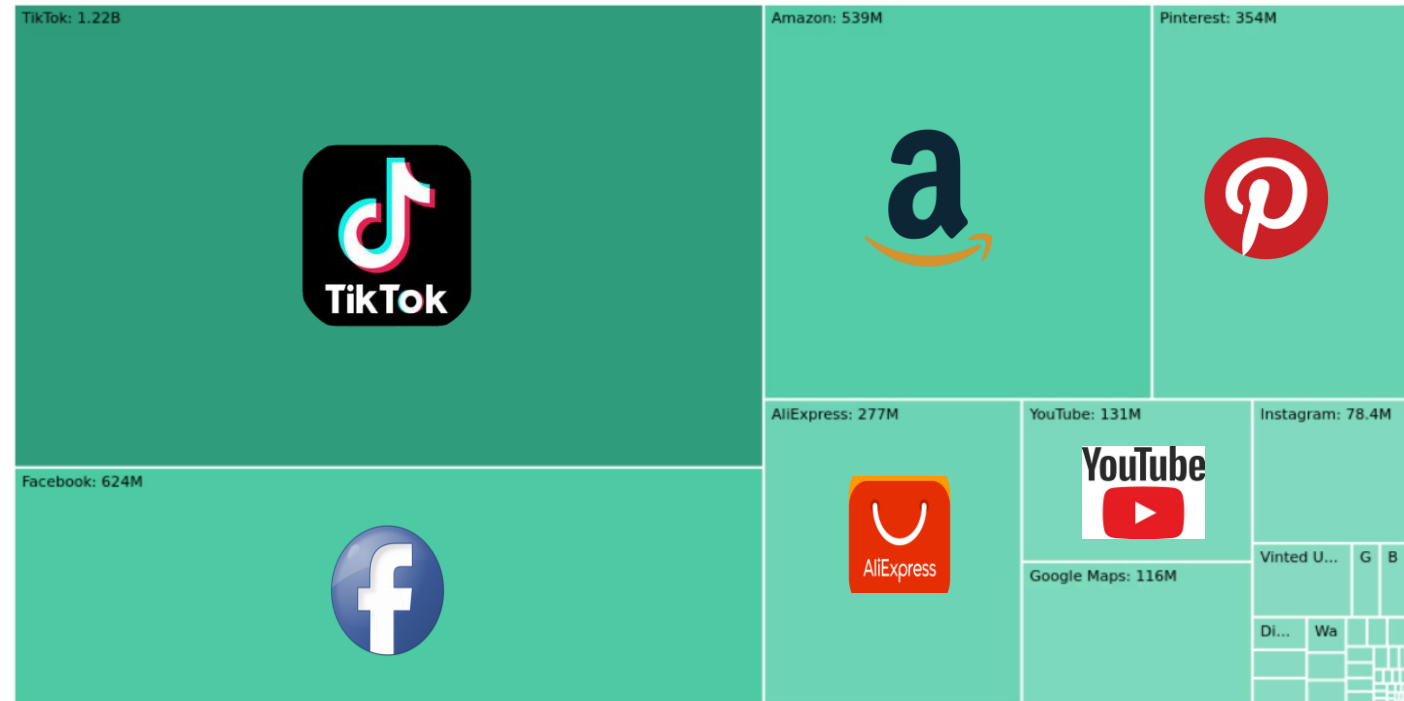
Daily dumps

- Near real time (24h delay)
- **Contain CSV with full SOR details**
- **Large files + pre-/post-processing**

Transparency DB: a massive dataset

The Transparency database is **large** (and *dsa-tdb* cannot do miracles).

- 1TB of **full daily dump** files
- 1.8GB of the aggregated dataset.
- ~5-10GB per day of caching when analysing.



dsa-tdb: how to use it

Three venues:

- `pip install --index-url https://code.europa.eu/api/v4/projects/943/packages/pypi/simple dsa_tdb`



- *Docker/Podman container*



- *Superset dashboards*



+ Online documentation:

The screenshot shows the documentation for the `dsa_tdb.etl` module. It includes a search bar, a table of contents, and detailed documentation for the `dsa_tdb.etl.daskAddColumns(df: DataFrame) → DataFrame` function. The function description states: "Unfortunatel dask does not support global state as it is lazy and using the submit trick slows things downs. See [here](https://dask.discourse.group/t/using-dataframe-apply-in-a-loop/949). This function adds the columns to the dask dataframe." The parameters are listed as `df (dd.DataFrame)` - The dask dataframe, and the return type is `dd.DataFrame`. Below this, the `dsa_tdb.etl.loadDataset` function is documented with its parameters and return type.

dsa-tdb: API and cli/package

- API interface (download, filter, aggregate)
- CLI interface (download, filter, aggregate)
- + Interactive

Prepare This is the **Prepare** endpoint. Use it to download the daily dumps and chunk them in a parquet or csv format. You can specify where to save the daily dumps, the platform to download (use `global` to download all of them), the version of the daily dumps, the start and end date for the preparation, the platforms to exclude (a comma separated list of platform names), etc. See the single value documentation for details.

POST /prepare/ Prepare


Prepare the daily dumps. Provide the output directory to store the prepared data. Specify the start and end date for the preparation (format YYYY-MM-DD), the platform name and version of the daily dumps.

The prepared data will be chunked in a parquet or csv format in the output directory following the structure: `output_dir/{platform_name}_{version}/daily_dumps_chunked/`

The checksum files will be stored in the parent daily chunked folder (do not delete them).

Parameters Cancel

Name	Description
output_dir string (query)	The root output directory to store the data. The files will be saved in the <code>platform__version</code> subfolders. <input type="text" value="/data/tdb_data"/>
platform_name string (query)	The platform name. <input type="text" value="global"/>
version string (query)	The version of the daily dumps (full or light). <input type="text" value="full"/>



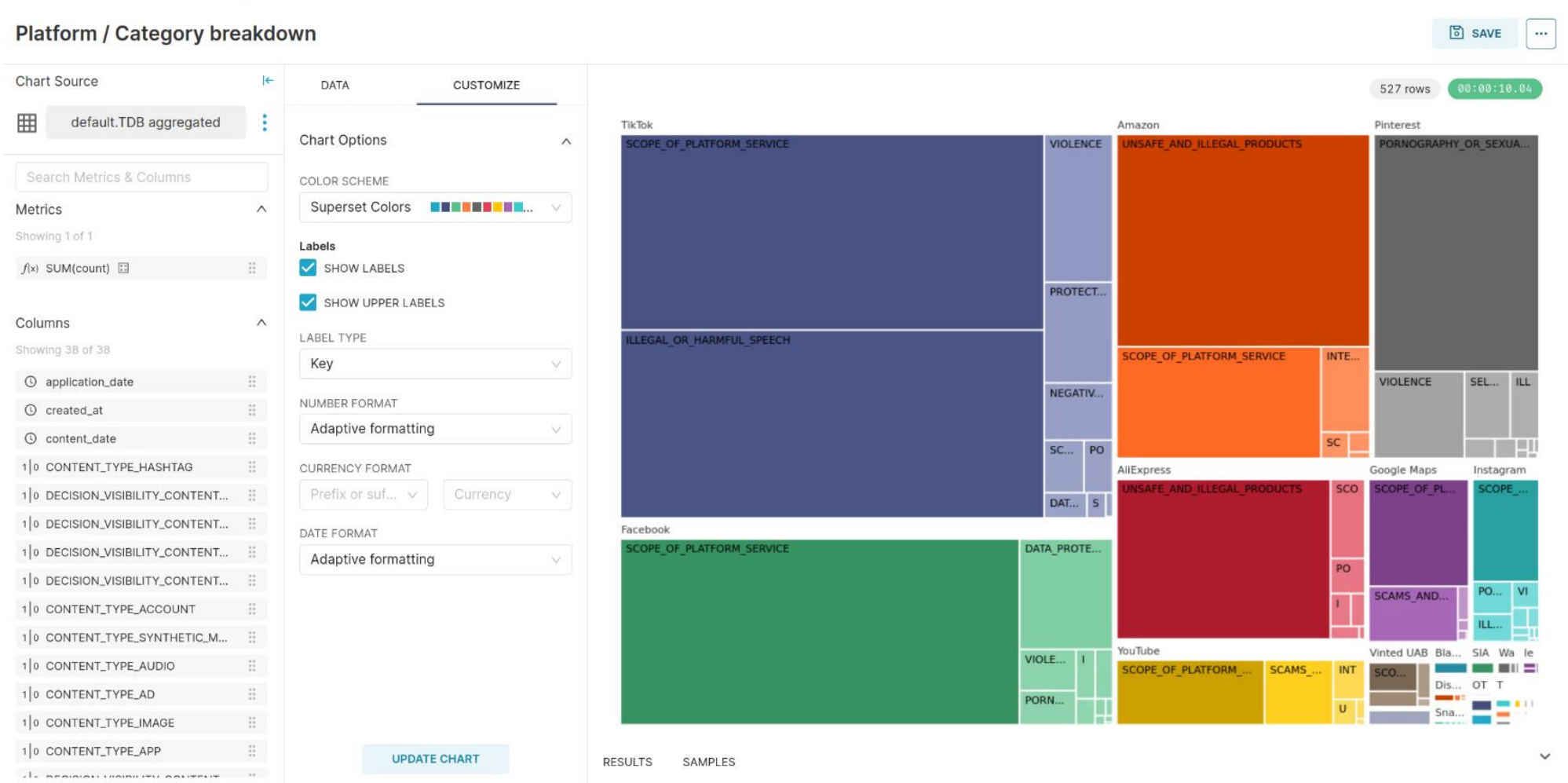
```
$ dsa-tdb-cli preprocess -p global
$ dsa-tdb-cli filter -c config.yaml
$ dsa-tdb-cli aggregate -c config.yaml
```



Stay around and join the workshop for a demo and additional details!

dsa-tdb: dashboard

Breakdown of platform and category



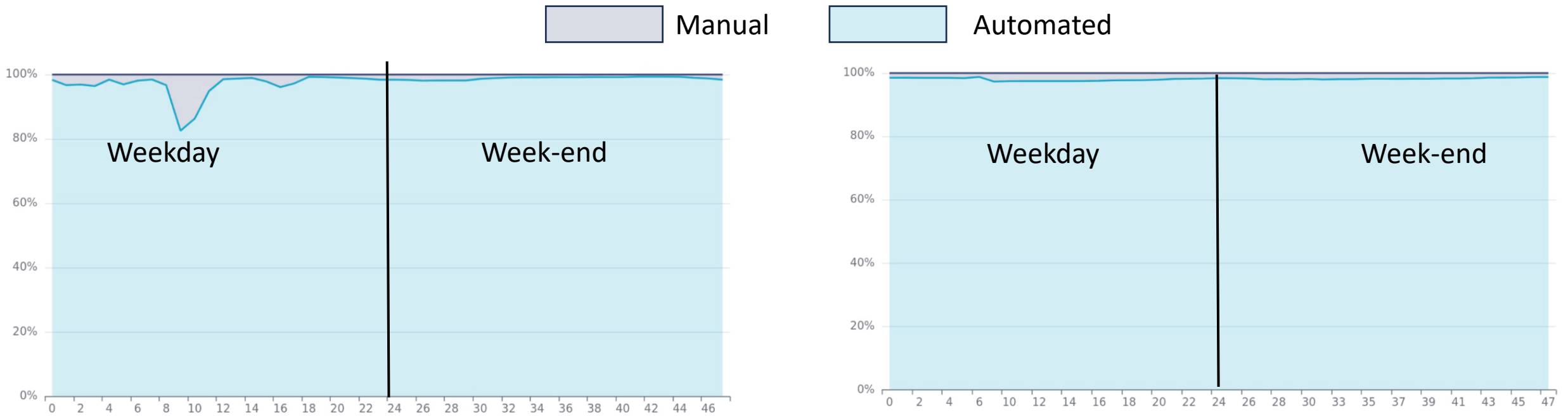
dsa-tdb: dashboard

Daily submission volume by platform



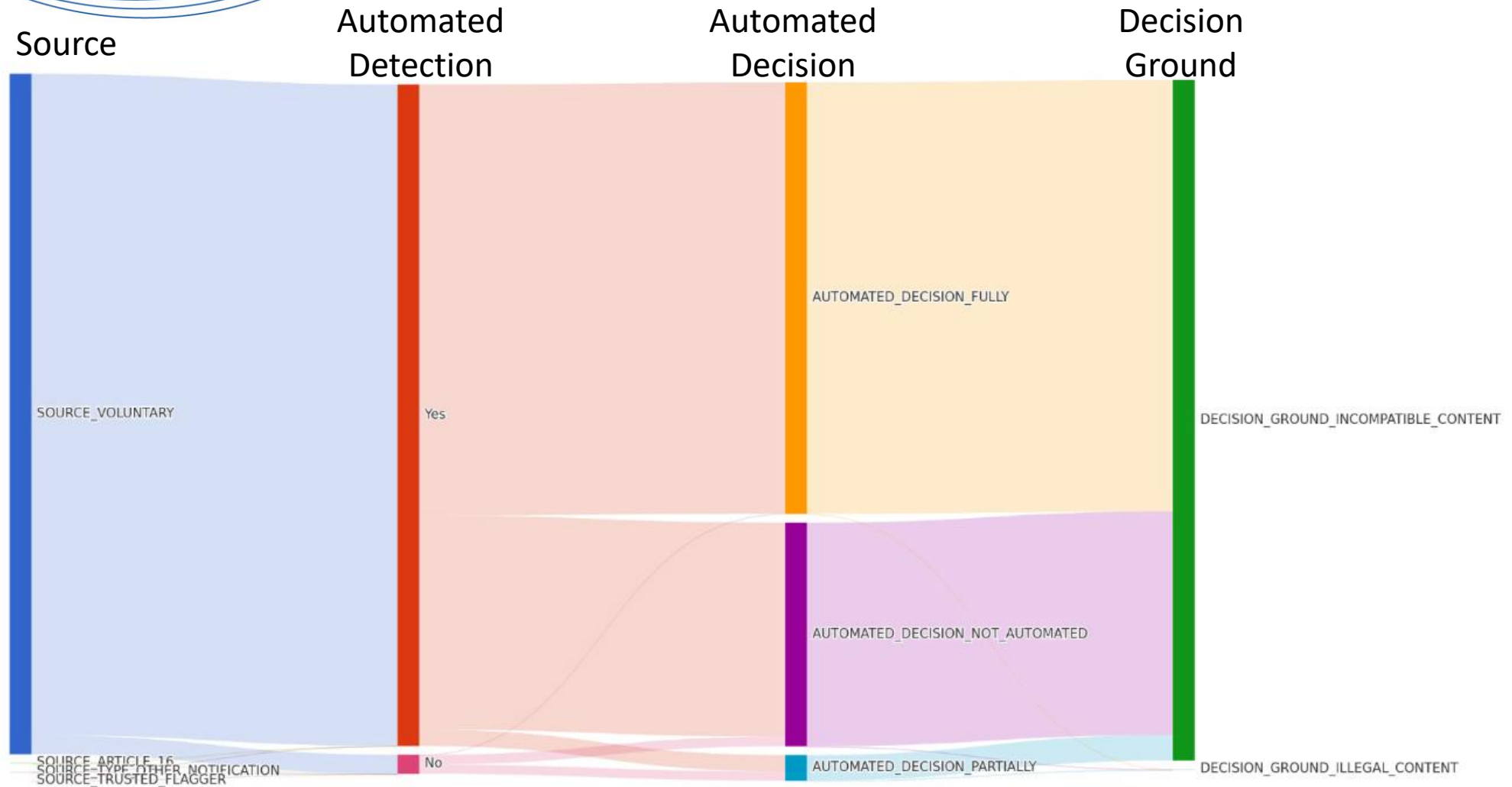
dsa-tdb: dashboard

Automated or manual content detection



dsa-tdb: dashboard

Flow of: Source



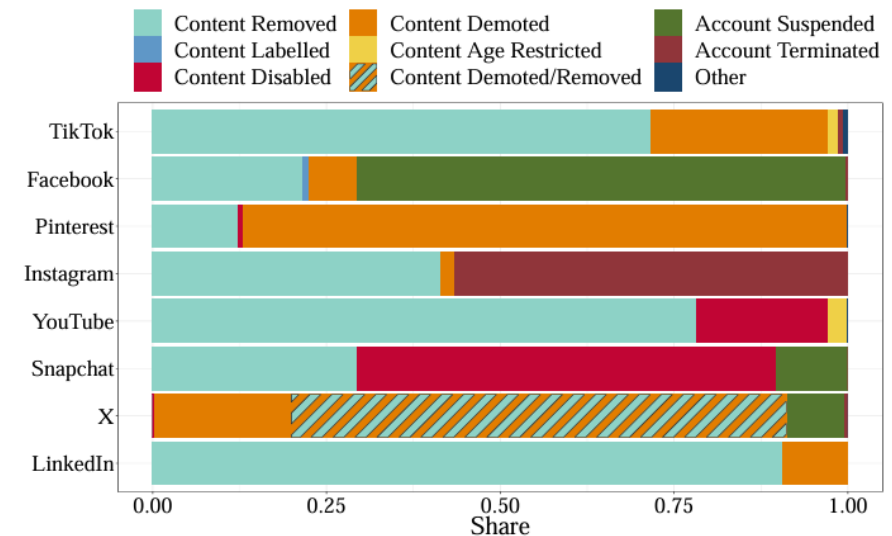
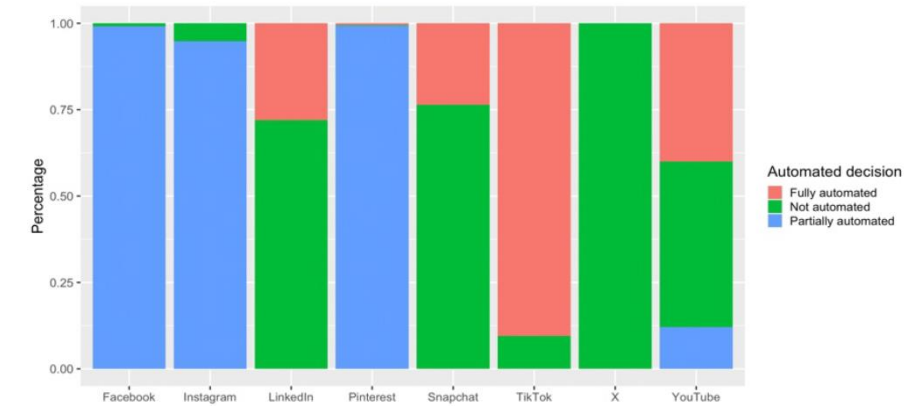
Community driven & research work on the Transparency DB

- Adaptation of the schema of the Database to the Reports in July 2025
 - Output of an open consultation between CSOs, researchers and companies
 - Optimizes reporting and aligns the transparency reporting provisions
 - Adds a product identifier (EAN-13) to track the spread of illegal products online

• Flourishing research community

- Kaushal, R., et al., Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. Preprint at <https://doi.org/10.48550/arXiv.2404.02894> (2024).
- Drolsbach, C. & Pröllochs, N., Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database. Preprint at <https://doi.org/10.48550/arXiv.2312.04431> (2023).
- Dergacheva, D., et al. One Day in Content Moderation: Analyzing 24 h of Social Media Platforms' Content Decisions through the DSA Transparency Database. (2023) doi:[10.26092/elib/2707](https://doi.org/10.26092/elib/2707).
- Platforms overwhelmingly use automated content moderation, first DSA transparency reports show – Lab Platform Governance, Media and Technology (PGMT). <https://platform-governance.org/2023/platforms-overwhelmingly-use-automated-content-moderation-first-dsa-transparency-reports-show/> (2023).
- Trujillo, A., Fagni, T. & Cresci, S. The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media. Preprint at <https://doi.org/10.48550/arXiv.2312.10269> (2023).
- Miller, G. Tracking the First Digital Services Act Transparency Reports | TechPolicy.Press. *Tech Policy Press* <https://techpolicy.press/tracking-the-first-digital-services-act-transparency-reports> (2023).

Dergacheva et al.: One Day in Content Moderation



Drolsbach, C. & Pröllochs, N.: Content Moderation on Social Media in the EU

DSA Transparency DB



Backend

github.com/digital-services-act/transparency-database



Website

transparency.dsa.ec.europa.eu



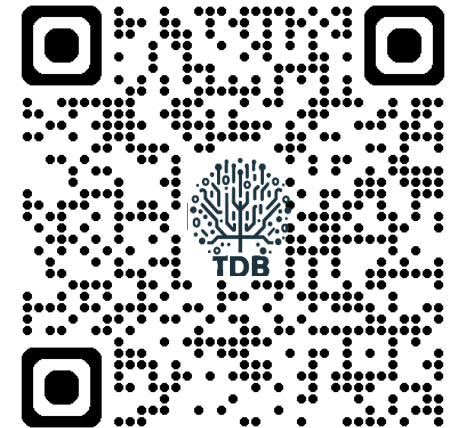
Package

code.europa.eu/dsa/transparency-database/dsa-tdb

Open source -> Try it, open issues and
[pull-requests are welcome!](#)



dsa-tdb



Stay around and join the workshop!

Thank You