



D2.2 : FRISCO Toolbox

Grant Agreement ID	101080100	Acronym	FRISCO
Project Title	Fighting teRrorISt Content Online		
Start Date	15/11/2022	Duration	24 Months
Project URL	https://friscoproject.eu		
Contractual due date	14/11/2023	Actual submission date	30/11/2023
Nature	DEM	Dissemination Level	PU
Author(s)	Adeline Kugler (Tremau), Pal Boza (Tremau), Antonis Koukourikos (NCSR-D), Pierre Sivignon (Civipol)		
Contributor(s)			
Reviewer(s)	Martina Manfredda (DLEARN), Rositsa Dzhekova (VPN)		



This project has received funding from the European Union's Internal Security Fund (ISF) programme under Grant Agreement No 101080100. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Document History *(including peer reviewing & quality control)*

Version	Date	Changes	Contributor(s)
v1.0	20/11/2023	First version	Tremau, NCSR-D
V1.5	29/11/2023	Quality control	VPN
V2.0	30/11/2023	Final version	Tremau

Executive Summary

The [*Fighting Terrorist Content Online \(FRISCO\)*](#) project is a project which received funding from the European Commission – Internal Security Fund under Grant agreement No 101080100 – and will be implemented by 8 partners from 6 different European countries between 2022 and 2024. FRISCO aims to support Hosting Service Providers (HSPs) to comply with the TCO Regulation. The Regulation (EU) 2021/784, addressing the dissemination of terrorist content online (TCO Regulation), entered into force in this context on the 7th of June 2021 and is applicable as of the 7th of June 2022. It sets out several specific measures that HSPs must implement to address the misuse of their services. The FRISCO project aims to help HSPs comply with the TCO regulation through 5 Work Packages (WPs).

Within the framework of WP2, a mapping report (T2.1) had been conducted during the first 6 months of the project to assess the needs and barriers faced by HSPs to be compliant with the TCO Regulation. The results highlighted a lack of awareness and resources from micro and small HSPs to be compliant with the TCO Regulation. As a response to the results of the mapping report, a toolkit (T2.2) has been developed responding to the needs and barriers in the report.

The first tool, a self assessment questionnaire, aims at helping small and medium HSPs understand their compliance with the TCO. This questionnaire is meant as a first step for HSPs to understand how their current internal processes for content moderation regarding terrorist content align with the TCO Regulation. The objective of this tool is to provide HSP with a compliance score, which helps them situate themselves in the path to full TCO compliance.

The second tool is an interactive process map that structures and describes the entire compliance process with the TCO Regulation and related duties for HSPs in a holistic way (i.e., from the exposure to terrorist content to the transparency reports). This tool, focused on HSPs' operational needs, provides a precise breakdown of the TCO Regulation, step by step, and is based on a chronological approach to compliance.

Finally, the third tool is a user-friendly trust and safety content moderation tool that addresses user-generated content related risks. The tool is based on Tremau's in-house solutions and tailored to HSPs' needs in relation to the TCO, thanks to the project's findings and resources. This tool is crafted to cater to the specific needs of HSPs, with the primary goal of optimising and streamlining content moderation workflows and processes.

Ultimately, these tools respond to the needs of HSPs regarding TCO compliance by offering comprehensive and tailored information to TCO compliance as well as providing guidelines and resources to enhance internal processes for better content moderation practices. Following the initial testing phase, refinement of these tools is scheduled over the coming months to elevate the overall user experience and make potential adjustments in light of the testing results.

Table of Contents

1	Introduction	7
2	Purpose and Scope	7
3	Approach for Work Package and Relation to other Work Packages and Deliverables	8
3.1	Methodology and Structure of the Deliverable	9
4	Tool n°1: Self-Assessment Questionnaire	10
4.1	Purpose and objective	10
4.2	Design and specifications	10
	Tool summary and requirements	10
	Technical Choices	11
4.3	Description: feature and functionalities	12
	Administration panel	12
	Online questionnaire	16
5	Tool n°2: Process map	18
5.1	Purpose and objective	18
5.2	Development process	18
5.3	Description: feature and functionalities	19
6	Tool n°3: Content moderation tool	20
6.1	Purpose and objective	20
6.2	Development process	20
6.3	Description: feature and functionalities	21
	Dashboard	21
	Reports List	21
	Appeal List	22
	User Management	22
	Policy Configuration	23
	Statement of Reasons	24
	Queues Configuration	25
	Moderation Actions	25
	LEA Portal Intake	26
	TCO transparency report	29
7	Access to FRISCO tools	29
8	Conclusions and next steps	30

List of Tables

Table 1: Question categories	9
------------------------------	---

List of Figures

Figure 1: Administration Panel - Questionnaire Overview	11
Figure 2: Question Categories List	12
Figure 3: Question Category Description	12
Figure 4: Question Panel	13
Figure 5: Question editor	13
Figure 6: Question creation form	14
Figure 7: Aggregate Dashboard	14
Figure 8: Responses panel	15
Figure 9: Questionnaire introduction	15
Figure 10: Presentation of a question with additional information attached	16
Figure 11: Questionnaire conclusion screen	16
Figure 12: Screenshot of dashboard	20
Figure 13: Screenshot of Report list	21
Figure 14: Screenshot of User Management feature	22
Figure 15: Screenshot of policy configuration feature	23
Figure 16: Screenshot of statement of reason feature	23
Figure 17: Screenshot of moderation action feature	25
Figure 18: LEA request submission form	28

List of Terms & Abbreviations

Abbreviation	Definition
HSPs	Hosting Service Providers
EU	European Union
TCO	Terrorist Content Online
FRISCO project	Fighting Terrorist Content Online project
WPs	Work Packages
LEAs	Law Enforcement Agencies

1 Introduction

Terrorist and other illegal content online is an increasing issue both from a security and public policy perspective. In today's complex, interconnected world, countering the spread of terrorist content online requires a multifaceted approach. One which recognizes the interdependence of global and digital phenomena and requires a combination of legislative, non-legislative and voluntary measures based on collaboration between authorities and Hosting Service Providers (HSPs).

The Regulation (EU) 2021/784, addressing the dissemination of terrorist content online (TCO Regulation), entered into force in this context on the 7th of June 2021 and is applicable as of the 7th of June 2022 and sets out several specific measures that hosting service providers exposed to TCO Regulation must implement to address the misuse of their services.

In this context, the objective of the [Fighting Terrorist Content Online \(FRISCO\)](#) project is to support HSPs to comply with the TCO Regulation, through:

1. Informing Hosting Service Providers and increasing their awareness of the Terrorist Content Online Regulation and their new obligations
2. Developing and validating tools, frameworks, and mechanisms to support hosting service providers in the implementation of the Terrorist Content Online Regulation
3. Sharing experience, best practices and tools to support the implementation of the Regulation.

The project activities are based on, first, the mapping of technical and human needs of target HSPs and their level of awareness of the Terrorist Content Online Regulation. Based on this, in WP2 tools, frameworks and mechanisms will be developed and validated. WP3 will enhance the capacity building of HSPs and WP4 will support raising their awareness in relation to compliance duties. Finally, WP5 will disseminate the results of the project.

As a result of the project, targeted hosting service providers will have a better understanding of what is terrorist content online and will be better prepared to deal with it and to comply with the Terrorist Content Online Regulation. This will lead to safer navigation online by reducing the risk of encountering terrorist content online.

The project has received funding from the European Commission – Internal Security Fund under Grant agreement No 101080100 and will be realised between November 2022 and November 2024. The consortium realising the project is composed of 8 beneficiaries from 6 different European countries, involving [NCSR-D](#) (Greece), the [French Ministry of Interior](#) (France), [Tremau](#) (France), [Civipol](#) (France), [Violence Prevention Network](#) (Germany), [IVSZ](#) (Hungary), [D-Learn](#) (Italy) and [INACH](#) (Netherlands).

2 Purpose and Scope

The toolkit has been designed to serve as a valuable resource for HSPs, offering accessible and user-friendly tools. The primary aim is to empower HSPs in addressing their specific needs related to compliance with the Terrorist Content Online (TCO) Regulation from the EU. Through a thoughtful selection of tools based on the take-aways of the mapping report realised during the first 6 months

of the project, the toolkit strives to simplify the complex landscape of TCO compliance, ensuring that HSPs can navigate these regulatory challenges efficiently and with ease. By providing accessible resources, the toolkit seeks to enhance the capability of HSPs to meet TCO requirements seamlessly, contributing to a more streamlined and effective operational environment in compliance with EU regulations.

The developed tools are summarised in this WP2 Toolkit report in corresponding sections below.

3 Approach for Work Package and Relation to other Work Packages and Deliverables

The project is delivered through five Work Packages (WPs). WP1 will manage and coordinate day-to-day management activities of the project. WP2 is dedicated to the development of relevant tools. WP3 aims to create a training program and to develop e-training modules, while WP4 helps raise awareness of HSPs. WP5 is dedicated to the dissemination and exploitation of results.

The main deliverable of WP2 is a toolkit aimed at providing user-friendly resources for small HSPs to be compliant with TCO regulations. This deliverable follows the mapping report, which delineated both the human and technical requirements of HSP in relation to their adherence to TCO compliance.

Conducted over a six-month period ending in May 2023, the mapping report reflected on the needs of European micro and small HSPs, gathering their insights to identify challenges and gaps in relation to the TCO Regulation.

Regarding the findings, one of the key takeaways of the mapping report was that micro and small HSPs tended to have a very limited awareness and knowledge of the TCO Regulation. As a comparison, we found that HSPs were clearly more aware of the Digital Services Act (DSA) than the TCO Regulation. However, despite the lack of awareness and preparedness, a large majority of HSPs perceived themselves as at low risk for terrorist content.

In general, several stakeholders also pointed out the lack of resources that micro and small HSPs were facing. This affected their willingness to invest in developing the right processes and implementing efficient tools not only to be compliant with new regulations but also to serve their own business needs. This would only be changed in case an imminent manifestation of a (terrorist) threat would push them to do so, otherwise these investments would be likely to be postponed for as long as possible.

Furthermore, micro and small HSPs fundamentally lacked the tools but even more importantly the processes to efficiently implement the provisions of the TCO Regulation. A large majority of the responding HSPs had not set-up the tools and processes to be compliant with the TCO Regulation. This is to some extent reflected by the fact that only about 20% of the HSPs responding to our online survey moderated all content generated by users in their services.

Based on the aforementioned results of Task 2.1 (the mapping report), a self-assessment questionnaire, a process map, and a content moderation tool have been developed to support small

HSPs in the implementation of the TCO Regulation to address the dissemination of terrorist content online on their platforms.

The first tool, an assessment questionnaire, will help provide first-hand information to HSPs about their current compliance level with the TCO, bringing awareness to the different requirements of the TCO and responding to the issues of lack of knowledge that the mapping report raised.

The second and third tools aim to tackle the raised issue regarding the lack of resources and processes to implement provision of the TCO Regulation. These tools provide the grounds for HSP to implement actionable resources to be more compliant with the TCO. The accessibility of these resources aims to cultivate a stronger commitment from HSP to invest in regulatory practices related to TCO, effectively addressing the current resource gaps faced by TCO.

3.1 Methodology and Structure of the Deliverable

This report reflects the progress made on the development of the toolkit during May-November 2023 since the results of the mapping reports have been available. It seeks to offer a comprehensive overview of the tools created. The report will be structured in three main sections, one for each tool, and give an overview of the purpose of the tools, their development processes, and their intended impact.

Through this detailed analysis, we aim to present not only the tangible progress made but also the strategic rationale and anticipated implications of each tool within the overarching developmental framework.

4 Tool n°1: Self-Assessment Questionnaire

4.1 Purpose and objective

The self-assessment questionnaire aims at helping small and medium HSPs understand their level of compliance with the TCO. This questionnaire is meant as a first step for HSPs to understand how their current internal processes for content moderation regarding terrorist content align with the TCO Regulation. The objective of this tool is to provide HSP with a compliance score, which helps them situate themselves in the path to full TCO compliance.

4.2 Design and specifications

Tool summary and requirements

The structure and content of the FRISCO Self-Assessment Questionnaire was collaboratively designed in the context of WP2, and is maintained in an internal shared document, with all partners able to provide their comments, to be assessed by the responsible partners (TREMAU, CIVIPOL). The questionnaire contains 34 questions relevant to the TCO Regulation. It is divided into sections, taking into account all the obligations to which they are subject within this framework, step by step. After completing the questionnaire, users receive a compliance score and are thus able to easily identify their gaps and needs with respect to the TCO Regulation.

In more detail, the questionnaire defines the following organisational elements.

1. **Introduction:** A text snippet introducing the context and scope of the questionnaire.
2. **Applicability:** A set of questions that determine if the TCO Regulation, and thus the questionnaire, applies to the user currently visiting the tool. It includes an introductory text, and three questions. If the answer to any of the questions is *NO*, the process concludes, and the user is informed that they are not affected by the regulation.
3. **Main Body:** The core of the questionnaire, containing all questions organised in **Parts** and **Categories**. There are two parts, *Requirements* and *Additional Requirements*, each comprising questions under specific categories.
4. **Conclusion (optional):** The system displays a text snippet indicating that the process is finalised, and possibly provides the link to the downloadable summary of the questionnaire, including the user's compliance score.

Given the scope of the tool and the general design of the questionnaire, the following requirements were taken into account for developing the online questionnaire.

Introduction and Conclusion: The relevant field **MUST** support HTML content.

Categories: The following categories (Table 1) are defined for the current questionnaire version.

Table 1: Question categories

Question Categories
Legal Representatives
Contact Point

Removal Orders
Preservation of Data
Information to Users
User Complaint Mechanisms
Transparency Obligations
Contact with Authorities
Specific Measures
Report to Authorities

Questions: the following characteristics apply to the included questions:

- Every question **MUST** belong to **EXACTLY** one category.
- A question **MAY** be associated with an explanatory text snippet.
- The text snippet **MUST** support HTML.
- Questions **CAN** be of one of the following types: *plain text, single choice, multiple choice*.
- Every question **MUST** carry a compliance weight.
- A question **MAY** act as a terminator for its category. That is, if a specific choice is selected, the remaining questions within the category are omitted.

Questionnaire Results: Questionnaire results culminate in a *compliance score*, a summary of the provided responses, and the distribution of the overall compliance score to the answered questions.

- 1 Results **MUST** be available only to the user answering the questionnaire.
- 2 Results **MUST** be downloadable upon request at any time in the future, under the condition that the user has provided their email after completing the questionnaire.
- 3 User-targeted export **SHOULD** be provided as a PDF file.
- 4 Exports and aggregate exports targeting FRISCO personnel **MUST** be provided in an editable format (e.g., CSV or MS[®] Excel files) in addition to the PDF export.

Technical Choices

The aforementioned requirements, in conjunction with current best practices for the development of web applications, led to the adoption of the following technical choices:

- Backend framework: Django (python) (<https://www.djangoproject.com>)
- API lib: Django Rest Framework (<https://www.django-rest-framework.org/>)
- Authentication lib: Rest Framework SimpleJWT (https://django-rest-framework-simplejwt.readthedocs.io/en/latest/rest_framework_simplejwt.html)
- Cryptography module: fernet (<https://cryptography.io/en/latest/fernet/>)
- PDF creation: pandas (<https://pandas.pydata.org/docs/>)
- Frontend Framework: React (typescript) (<https://react.dev/>)

- Authentication: jwt-decode (<https://www.npmjs.com/package/jwt-decode>)
- Plotting: reactflow (<https://reactflow.dev/>)
- Styling: tailwindcss (<https://tailwindcss.com/docs/installation>)
- Text Editor: ckeditor5-build-classic (<https://www.npmjs.com/package/@ckeditor/ckeditor5-build-classic>)
- Basic web UI elements: radix-ui (<https://www.radix-ui.com/primitives>)
- Icons: ionicons (<https://ionic.io/ionicons>)
- Date utility: date-fns (<https://date-fns.org/>)

The code for the tool is maintained and publicly available at a dedicated GitHub repository (X), where issue tracking and feature requests are also managed.

4.3 Description: feature and functionalities

Administration panel

The definition and management of the online Self-Assessment Questionnaire is carried out via the tool's administration interface. Administrators are provided with a username and password by the responsible FRISCO partners and are subsequently able to use the administration features through the web interface of the panel.

There are four main administration sections, described below:

The **Questionnaire Overview** section allows administrators to define general characteristics of the questionnaire, namely its title, its introductory informative text snippet, and its outro.

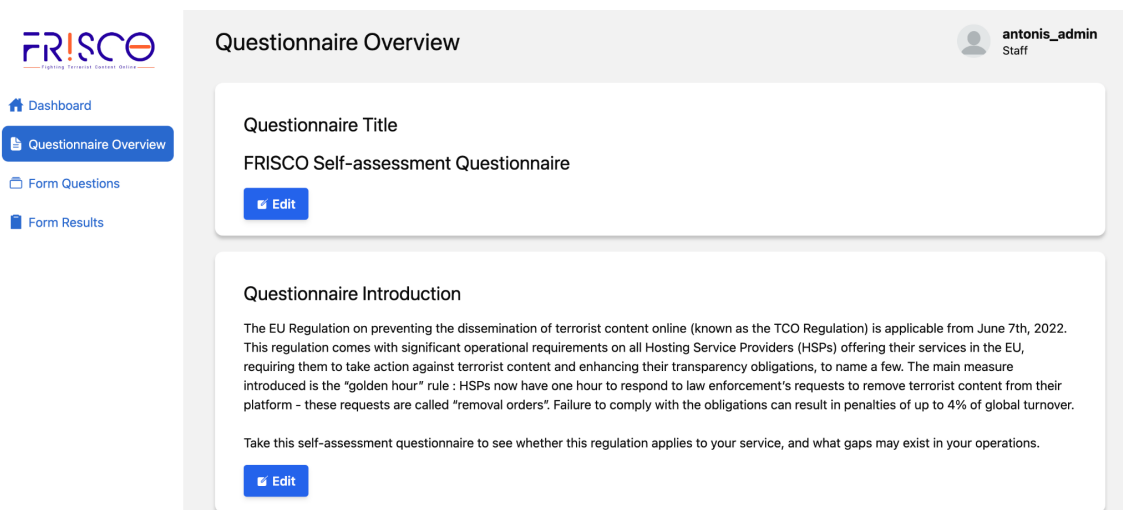


Figure 1: Administration Panel - Questionnaire Overview

The **Form Questions** section is where most of the elements for building the questionnaire are defined. Question categories are presented in the respective panel, from where the administrator can define and edit the categories applicable to the questionnaire.

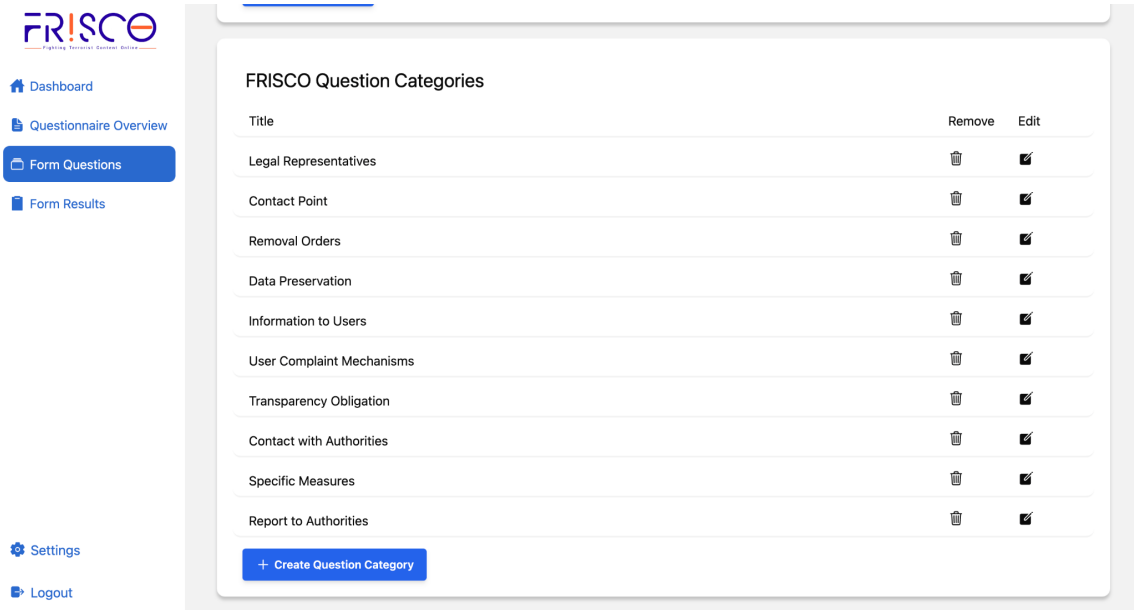


Figure 2: Question Categories List

The Category Creation form prompts the user to enter the title of the category along with a rich text snippet acting as explanatory input for the category, to help questionnaire users understand the scope and coverage of the category.

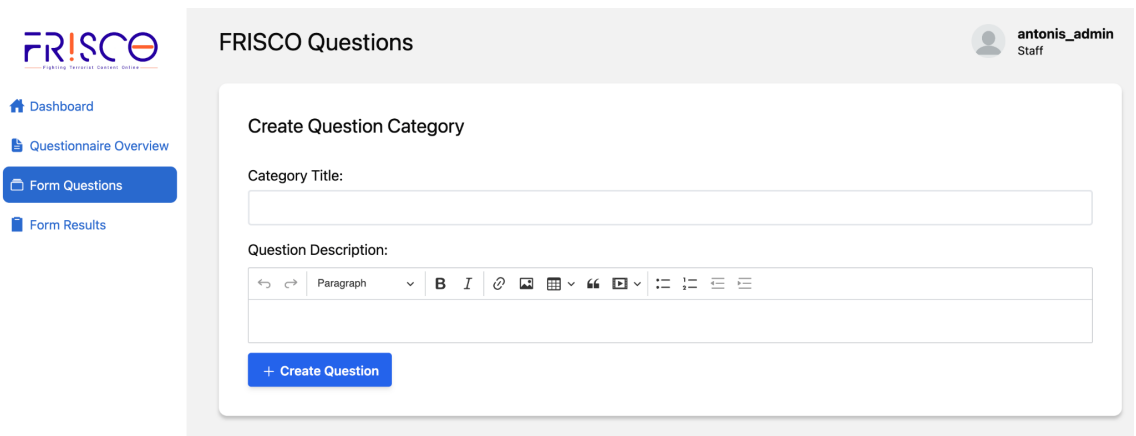
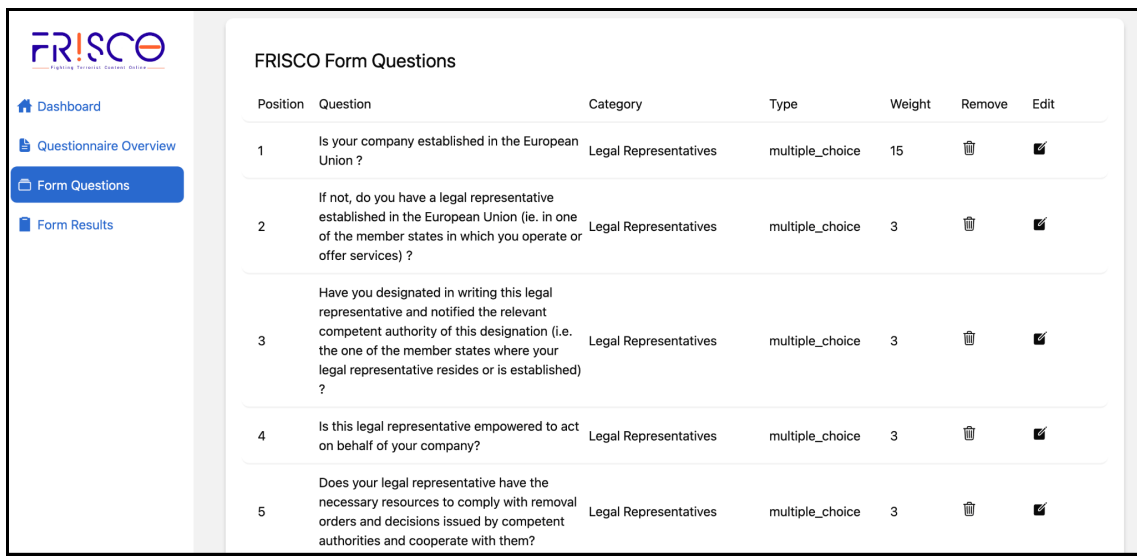


Figure 3: Question Category Description

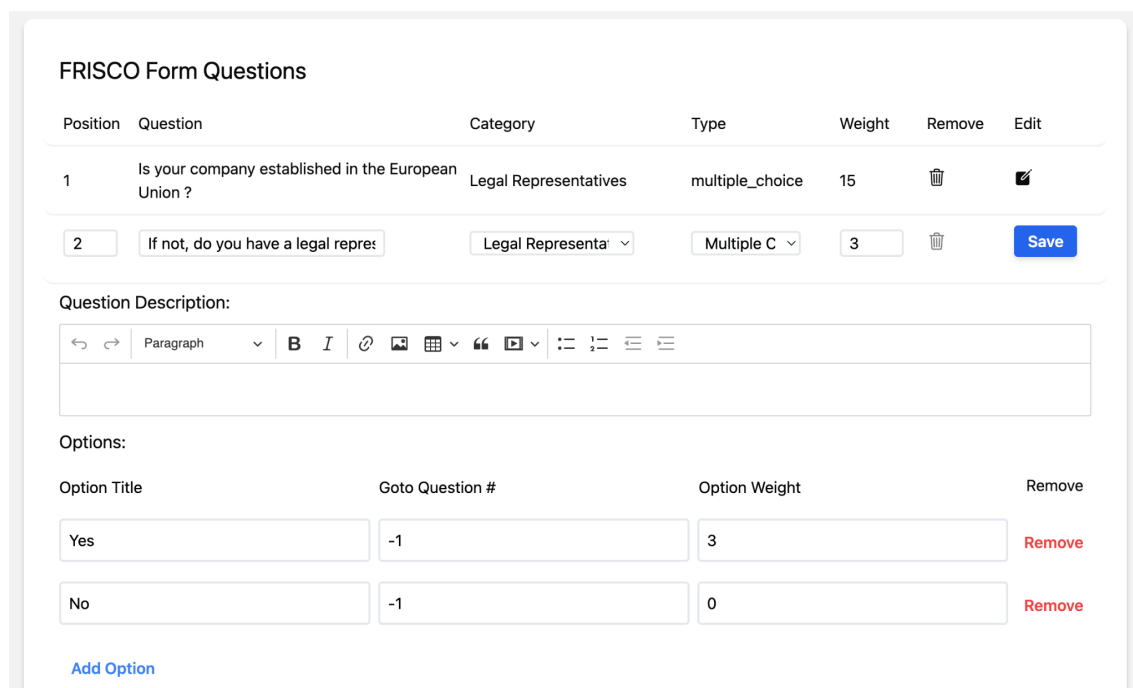
Similarly, Questions are presented in a distinct panel, displaying their basic information (order, title, category, type, weight) and allowing administrators to remove or edit them.



Position	Question	Category	Type	Weight	Remove	Edit
1	Is your company established in the European Union ?	Legal Representatives	multiple_choice	15		
2	If not, do you have a legal representative established in the European Union (ie. in one of the member states in which you operate or offer services) ?	Legal Representatives	multiple_choice	3		
3	Have you designated in writing this legal representative and notified the relevant competent authority of this designation (i.e. the one of the member states where your legal representative resides or is established) ?	Legal Representatives	multiple_choice	3		
4	Is this legal representative empowered to act on behalf of your company?	Legal Representatives	multiple_choice	3		
5	Does your legal representative have the necessary resources to comply with removal orders and decisions issued by competent authorities and cooperate with them?	Legal Representatives	multiple_choice	3		

Figure 4: Question Panel

Already defined questions can be edited inline via the relevant edit button, which opens the question editor that allows administrators to alter the parameters of the question.



FRISCO Form Questions

Position	Question	Category	Type	Weight	Remove	Edit
1	Is your company established in the European Union ?	Legal Representatives	multiple_choice	15		
2	If not, do you have a legal repre:	Legal Representa	Multiple C	3		

Question Description:

Paragraph **B** *I*

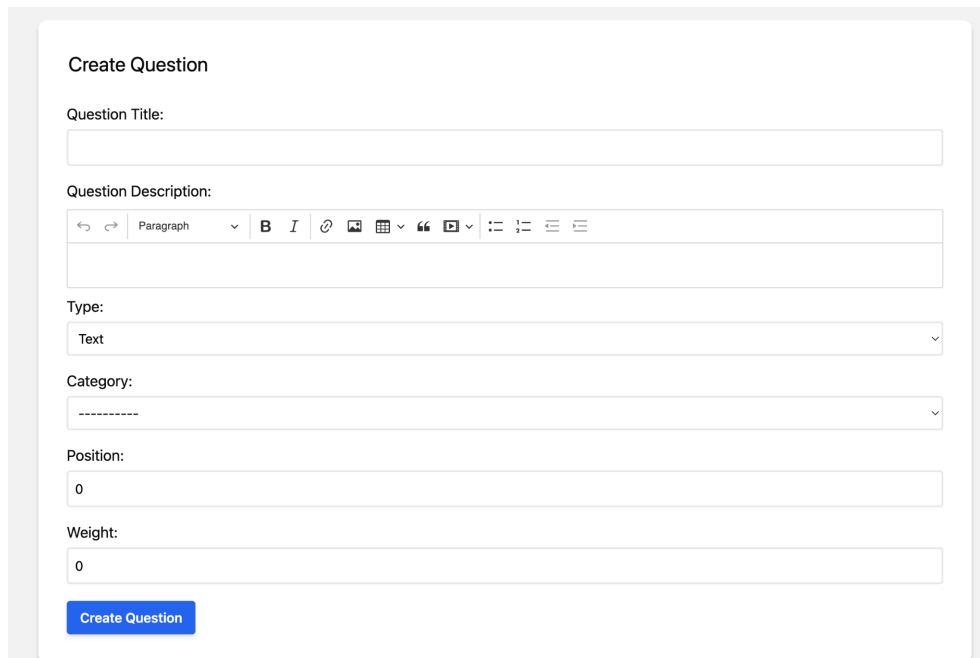
Options:

Option Title	Goto Question #	Option Weight	Remove
Yes	-1	3	Remove
No	-1	0	Remove

[Add Option](#)

Figure 5: Question editor

The same fields are presented when creating a new question, via the add button at the bottom of the question list.



The screenshot shows a 'Create Question' form with the following fields and controls:

- Question Title:** A text input field.
- Question Description:** A rich text editor with a toolbar containing icons for undo, redo, paragraph style, bold, italic, link, unlink, image, table, quote, video, bulleted list, numbered list, indent, and outdent.
- Type:** A dropdown menu with 'Text' selected.
- Category:** A dropdown menu with a dashed line as a placeholder.
- Position:** A text input field with the value '0'.
- Weight:** A text input field with the value '0'.
- Create Question:** A blue button at the bottom.

Figure 6: Question creation form

All questions are assigned to exactly one of the defined categories as per the requirements. Currently supported types for the answers to the questions are text, multiple choice, multiple select, file, email, and phone.

The **Dashboard** provides administrators with an aggregate view of the questionnaires collected so far. The dashboard showcases information on the usage of the tool (answers submitted), and the distribution of answers for multiple choice and multiple select questions using appropriate charts.

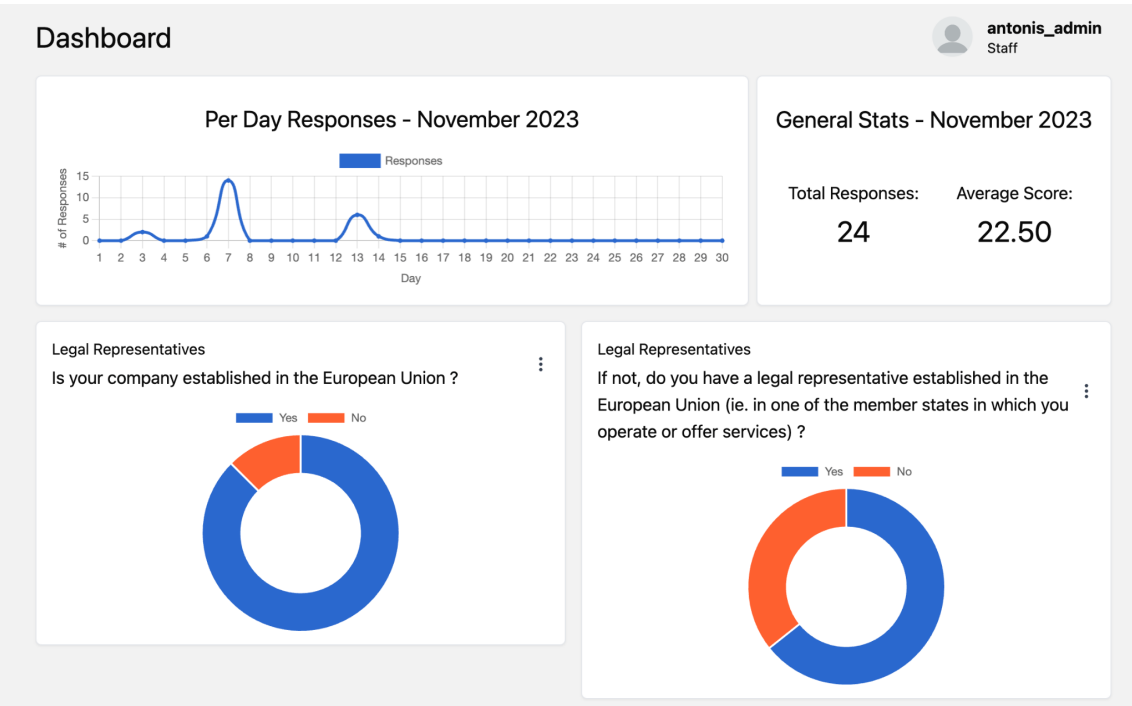


Figure 7: Aggregate Dashboard

Finally, the **Form Results** section of the administration panel provides more fine-grained access to the collected responses. Administrators are able to access and download individual questionnaire responses, filtering them with respect to date or status (completed or pending), and download the filtered subset of responses as a CSV file.

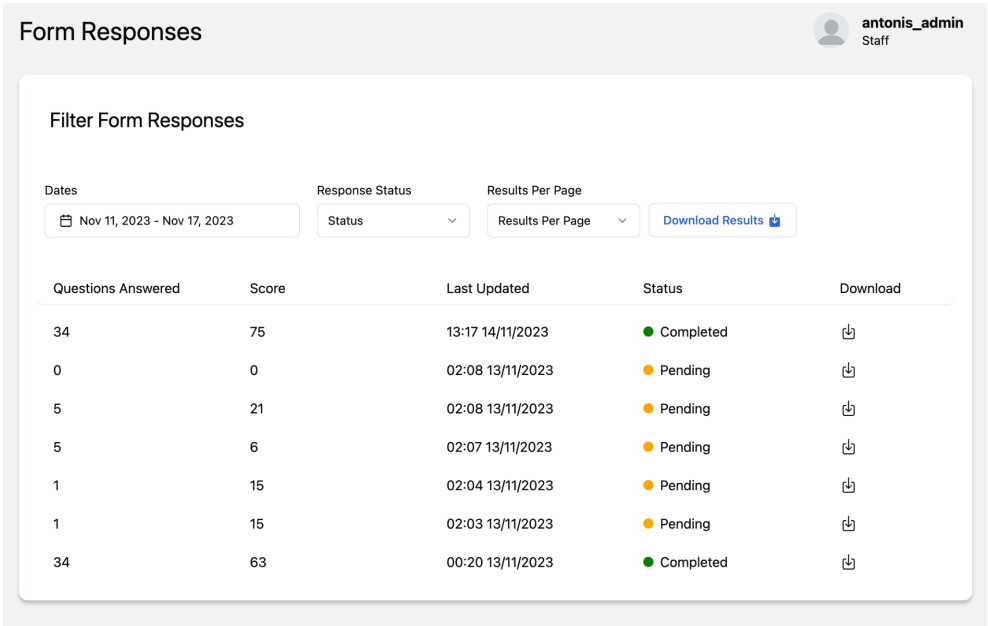


Figure 8: Responses panel

Online questionnaire

Following the introduction page, the online tool essentially comprises a series of screens, each corresponding to the questions defined in the questionnaire.

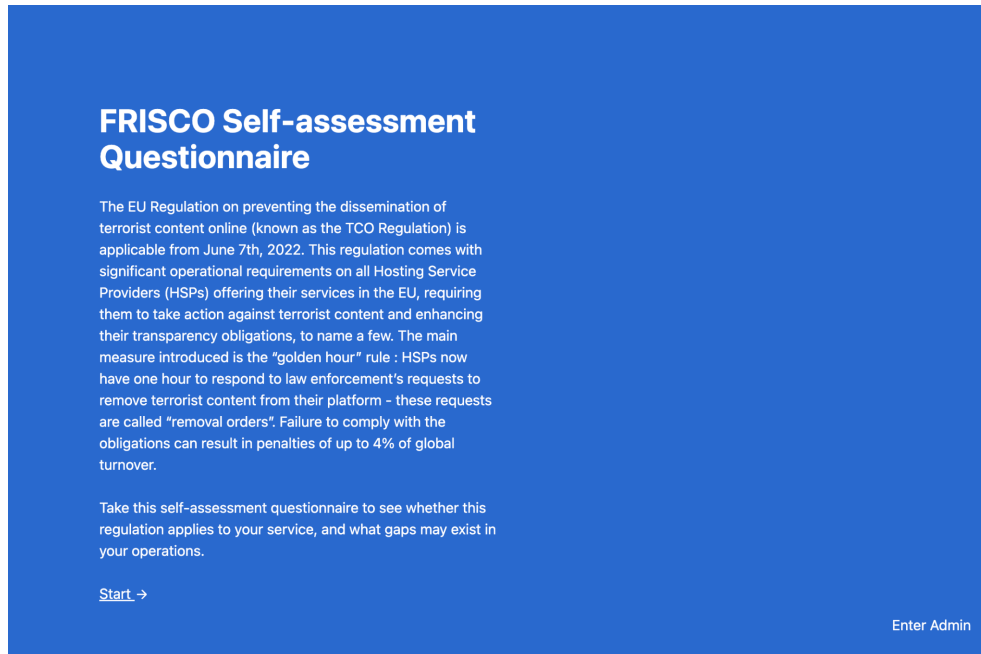


Figure 9: Questionnaire introduction

The user proceeds with providing their answers to the defined questions in the defined order. For questions where additional information is provided, an information symbol appears on the side; the users are presented with the additional information when hovering the information symbol.

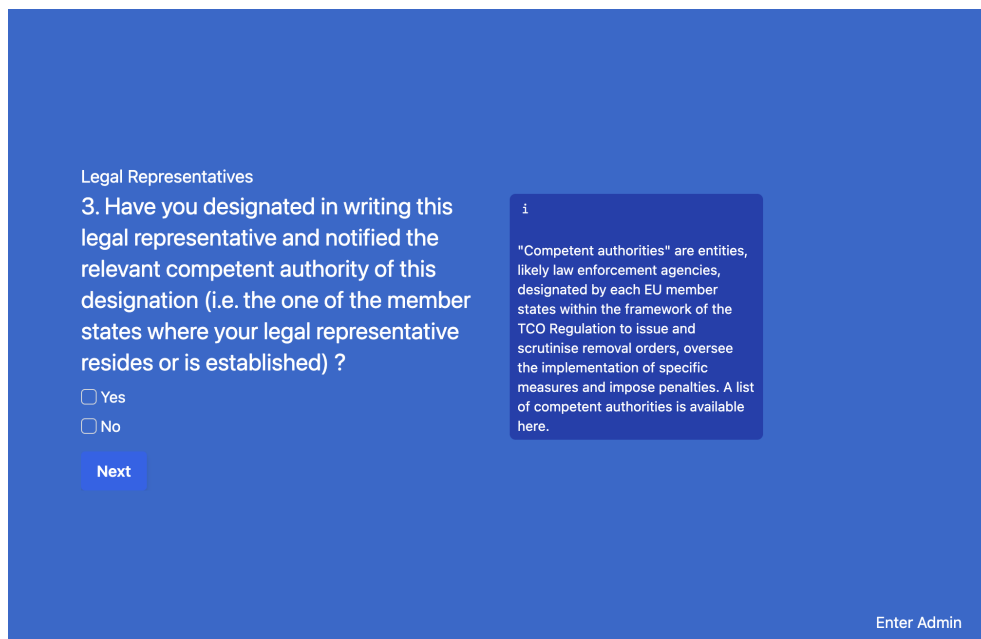


Figure 10: Presentation of a question with additional information attached

When completing the questionnaire, users are informed on their compliance score and gain access to a download of a file containing their detailed responses to each question. Additionally, they can provide their email address to receive their report there.

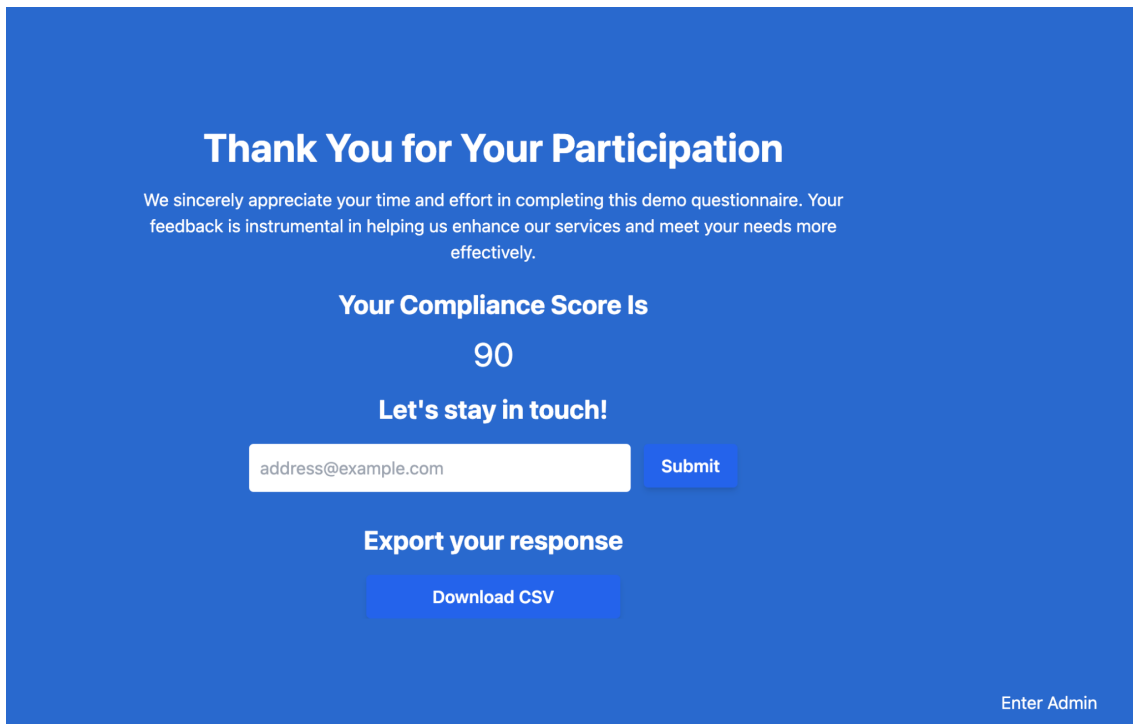


Figure 11: Questionnaire conclusion screen

5 Tool n°2: Process map

5.1 Purpose and objective

FRISCO's process map is an interactive tool that structures and describes the entire compliance process with the TCO Regulation and related duties for HSPs in a holistic way, so to say from the exposure to terrorist content to the transparency reports. This tool, focused on HSPs' operational needs, provides a precise breakdown of the TCO Regulation, step by step, and is based on a holistic and chronological approach to compliance. The process is displayed gradually, thanks to interactive and/or animated transitions, and the whole map can be downloaded entirely in the end. Legend, colour code and steps increase ergonomics and user's experience. Finally, more information, resources and tools are made available via the tool.

- Goal: Providing a detailed overview of the different steps and processes from the moment terrorist content appears to the final transparency reports. The various steps are broken down depending on the type of removal order outlining the different measures that should be taken and considering special measures.
- Interactive and user-friendly: Users will be able to understand the various steps included in the process map, easily develop their personalised workflow by answering clear questions using a simple interface, and be able to view and examined both distinct steps and their overall path.

5.2 Development process

The FRISCO Process Map concept was designed and refined via a static diagrammatic representation of the core steps and processes to be enacted by HSPs.

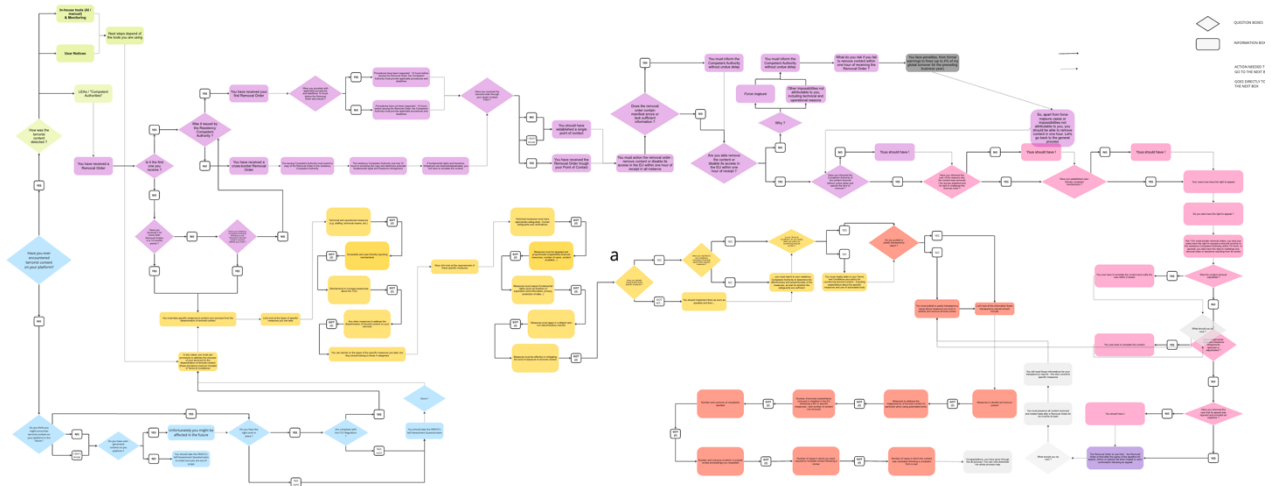


Figure 12: FRISCO Process Map Diagram

The map essentially comprises elements of different types and defines their associations depending on user responses where user input is asked. Hence, the following element types are specified:

1. **Binary (yes/no) question:** Each arrow leads to the next step.
2. **Multiple choice question:** Options are presented automatically after the question. The user is called to select one of the *option* nodes.
3. **Check Nodes:** User confirms that they have completed the process/action described by the node.
4. **Action Nodes:** The node itself provides an introduction on the actions to be performed/checked. Clicking on the node presents a checklist of the different sub-actions entailed in the action described by the node.
5. **Checklist:** A collection of check nodes. **All** of them must be confirmed to move to the next step of the flow.

Based on these elements the process map was transferred to an interactive version, implemented as a web application using the static map as the conceptual baseline and modelling map elements in accordance with the element typology.

5.3 Description: feature and functionalities

The interactive process map presents the chain of steps and processes incrementally to the user, with the steps covered at any given moment presented as the workflow and the expected user action/decision submitted through the input box at the upper right corner of the map's screen.

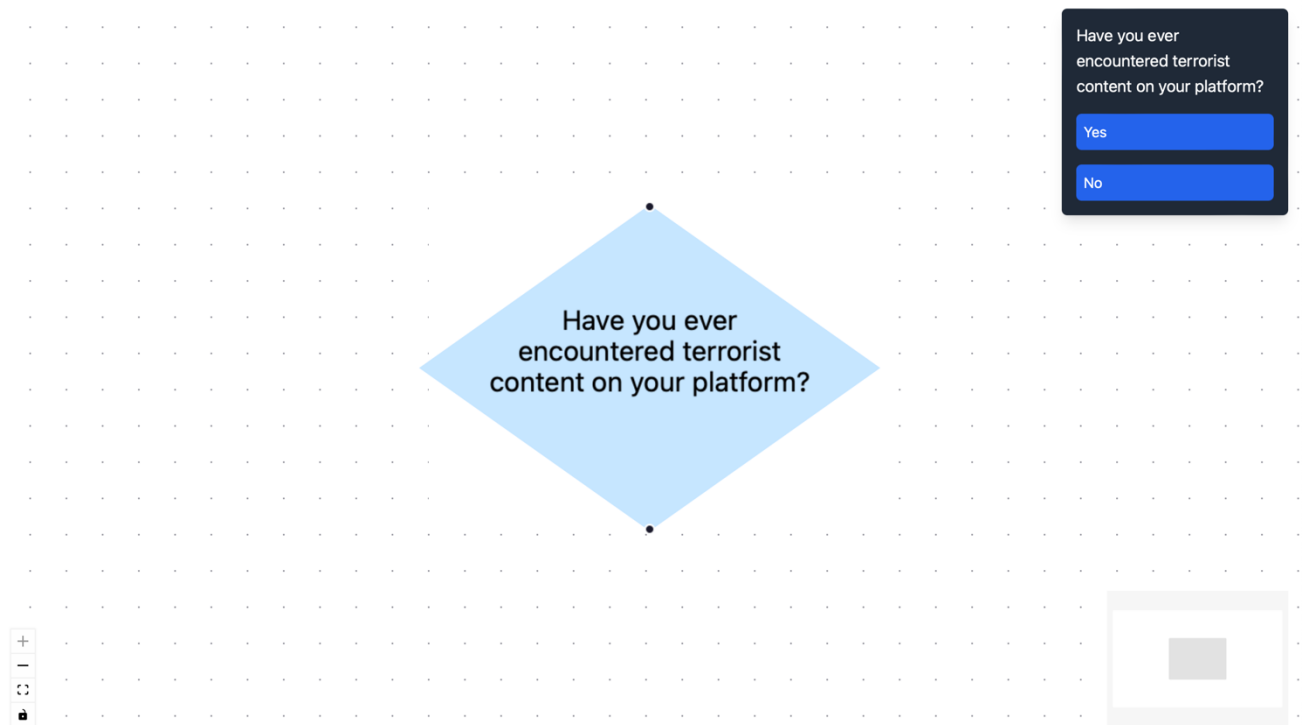


Figure 13: FRISCO Process Map - Initial Step

As users provide their input the map moves forward; ultimately the complete process path for the given user is displayed.

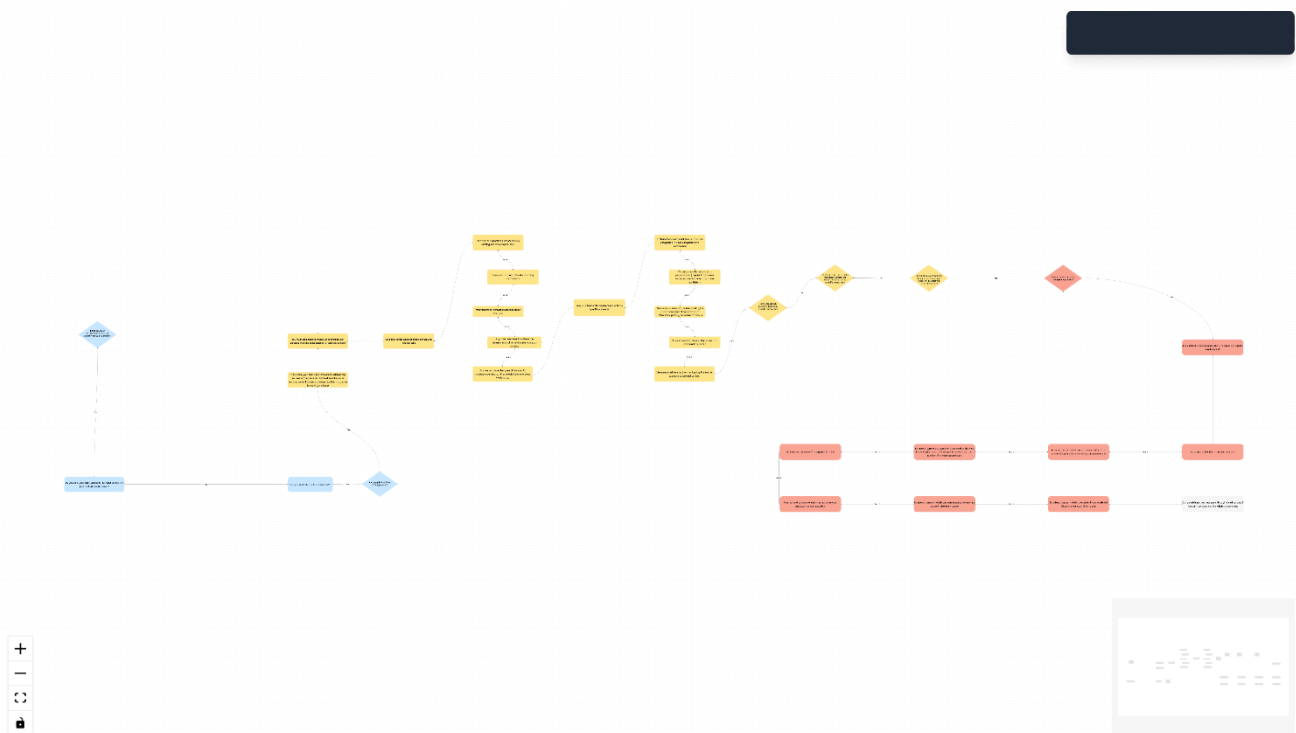


Figure 14: FRISCO Process Map - Full User Path

6 Tool n°3: Content moderation tool

6.1 Purpose and objective

FRISCO's content moderation tool is a user-friendly trust and safety tool that addresses user-generated content related risks. The tool is based on Tremau's in-house solutions and tailored to hosting service providers' needs in relation to the new regulations such as the TCO, thanks to the project's findings and resources. This tool is crafted to cater to the specific needs of HSPs, with the primary goal of optimising and streamlining content moderation workflows and processes. It is provided to them via a SaaS contract by contacting Tremau.

Potential integration with the detection tool developed in the framework of the sister project ALLIES could be made in the future. There are ongoing discussions on the possibility of cooperating and partnering with ALLIES.

6.2 Development process

Based on the results of the mapping report, Tremau's in-house team identified the different needs of HSPs regarding the TCO. These needs mainly revolve around the capacity to manage and moderate potential terrorist content as well as the transparency report requirement.

Following this, Tremau's product and tech team conducted a thorough evaluation of their existing in-house solution to ascertain its adaptability to the specific requirements of HSPs. In response to TCO specifications, essential features, including the TCO transparency report feature, were created to fulfil the designated TCO requirements.

6.3 Description: feature and functionalities

FRISCO's user-friendly content moderation tool serves as a trust and safety solution designed to mitigate risks associated with user-generated content. Utilising Tremau's in-house solutions and customised to meet the specific requirements of hosting service providers, the tool incorporates insights derived from the FRISCO project. At its core, this tool provides a content moderation platform in which incoming flags from different sources of reports are aggregated. Moderators can subsequently view the reported cases and make related decisions (i.e, keep or remove content, signal user, etc.) based on platform policies.

The live functioning of this tool requires an API integration with the system of the HSP that will be using the tool.

The subsequent section describes the various functionalities of the tool.

Dashboard

A graphical user interface providing an overview of the key metrics and KPIs relevant to the content moderation platform. This dashboard shows the number of flagged cases and appeals, as well as data analytics regarding cases per time, labels, etc.

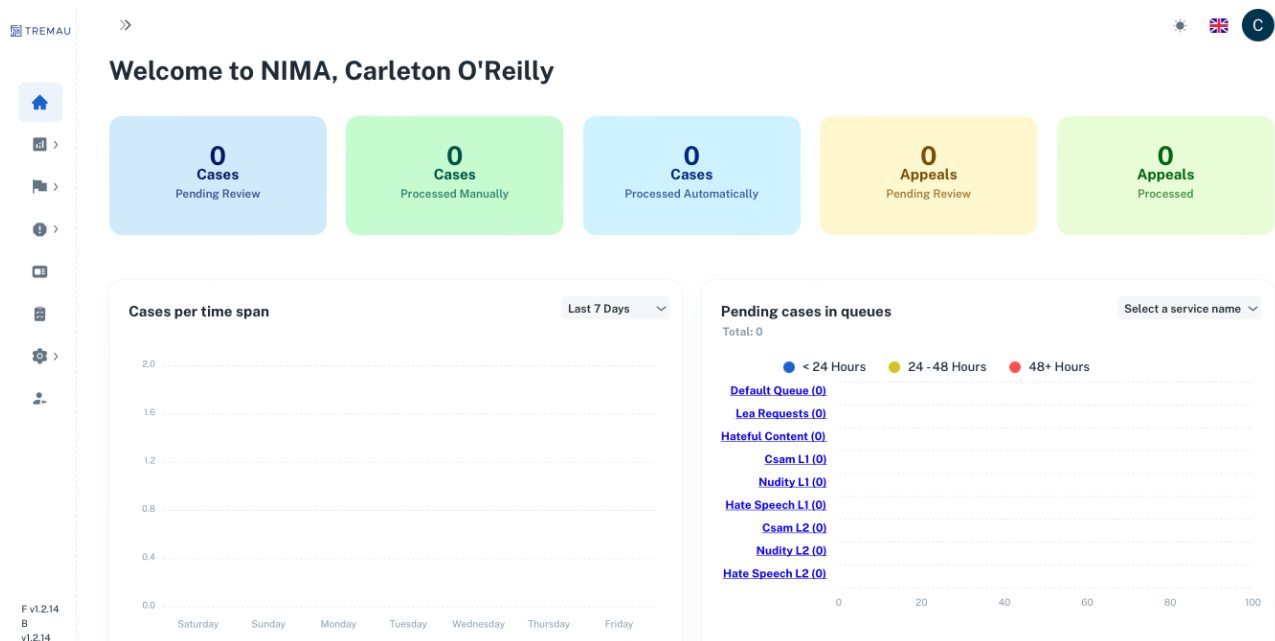


Figure 15: Screenshot of dashboard

Reports List

The tool provides information on the number of cases that are pending. Content moderators need to understand how flags and reports are evolving and being processed on a day-to-day basis.

Based on how flags and cases are ramping up in the user's system, users can consequently increase the number of individuals in specific content moderator user groups.

The “Pending cases in queues” analytics component enables you to view at a glance how your content moderation queues are performing.

The “My Queues” screen lists all queues assigned to the currently signed in user and informs the content moderator on how many cases are pending review as per SLAs, enabling the latter to set his/her right level of goals and aims.

This feature also includes the possibility to search cases by user ids involved in the reports, we provide both reporter ids and reportee ids.

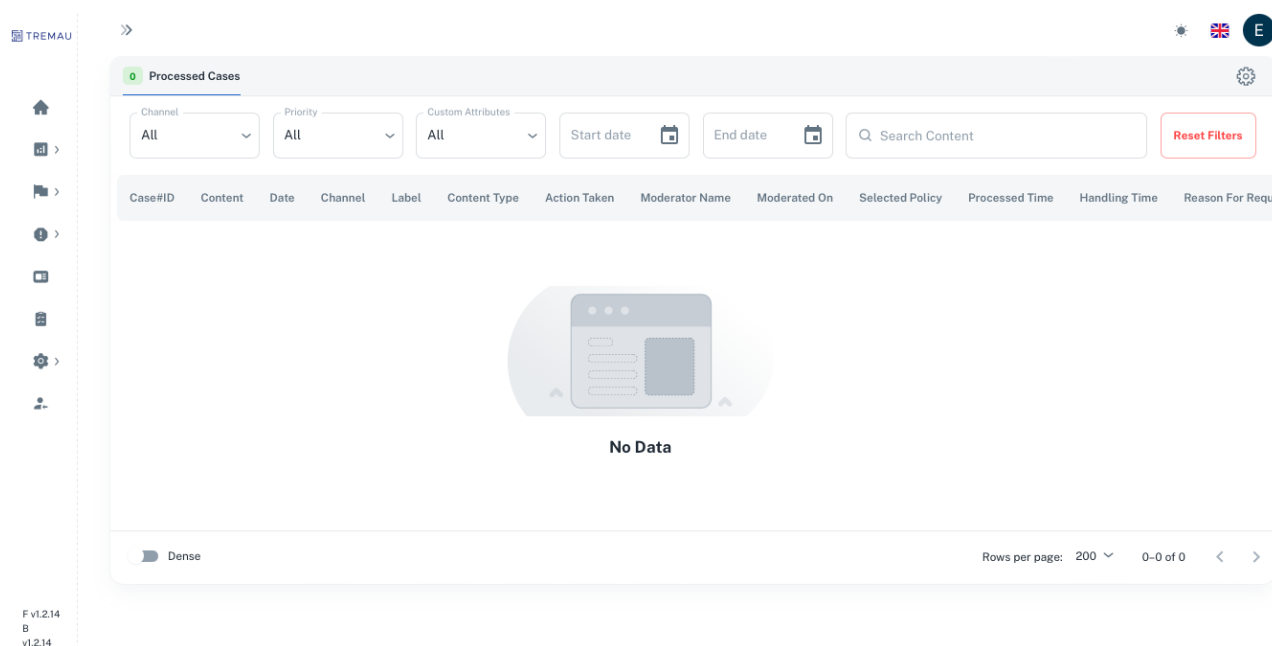


Figure 16: Screenshot of Report list

Appeal List

The tool offers an appeal mechanism for users to request a review of a moderation decision for content which was reported. This appeal mechanism is an internal remedy which is instituted through an appeal form where users can explain why they believe their content was reported incorrectly and add any evidence to substantiate their appeal/counterclaim.

All the appeals can be seen in the appeals list mirroring the appearance of the report list.

User Management

This tool offers the possibility to set up new user accounts for teams and colleagues, as well as create user groups and assign users to those user groups. Creating user groups for Trust and Safety spaces can help streamline content moderation processes.

By creating groups of users, it is possible to assign specific content moderation queues to leverage a team’s expertise to tackle reports in which they might be specialised. For example, create a group of Hateful Speech Experts for collaborators and assign to them only flags that pertain to Hateful Speech.

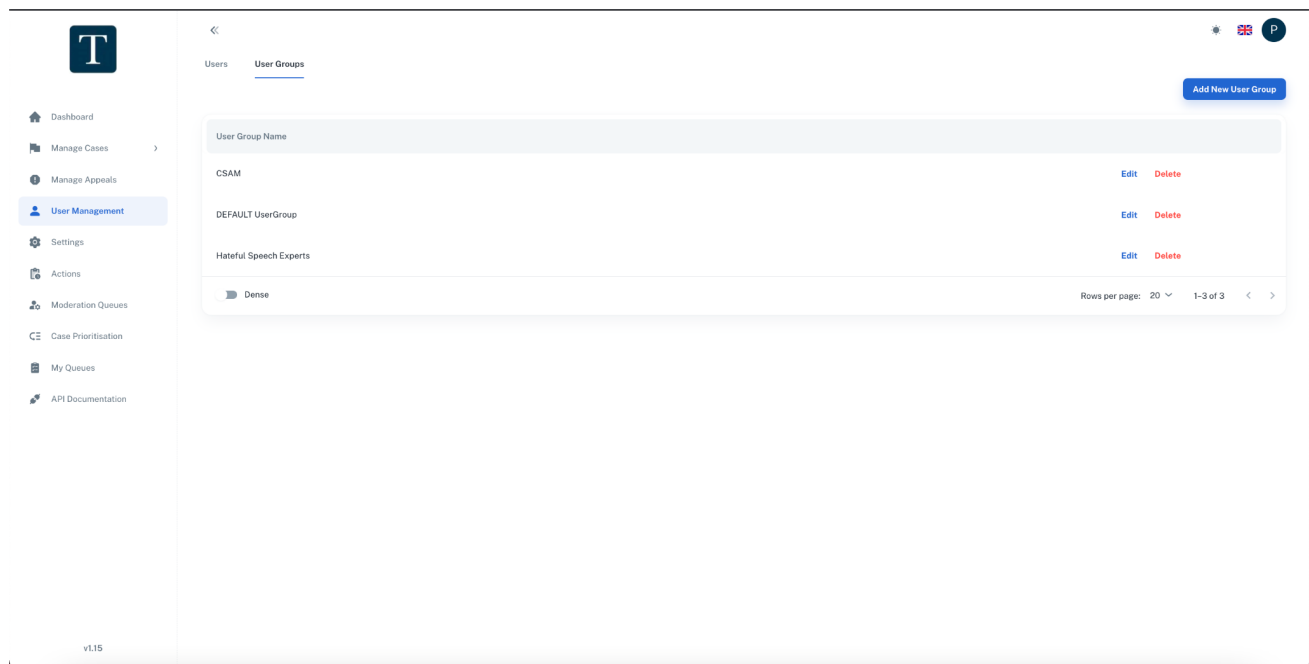


Figure 17: Screenshot of User Management feature

Policy Configuration

The tool offers the possibility to set up a Policy Enforcement Strategy. Policies can be configured in the settings. Users can configure policies in the settings, incorporating sub-policies, tiers (e.g., levels of severity for each policy violation), and labels. These features aid in categorising specific cases by breaking them down into distinct categories. Moderators can subsequently assign cases to each policy and its corresponding sub-policies.

Keyboard shortcuts can be added to each policy configuration.



The tool provides an automated workflow allowing to set up statement of reasons notifications (i.e., a notification to a user justifying why a certain decision regarding a case was made). These notifications are sent to perpetrators of Policy Guidelines, upon content moderation decisions.



Queues Configuration

The Query Based Builder for creating moderation queues exposes a set of criteria which the user can use to redirect flags based on these criteria to specific moderation queues.

These criteria are:

- **Channel** - This is the incoming source of the report or flag, Example: User Report, LEA, Trusted Flagger, a specific AI detection tool, ...
- **Label** - The label(s) is(are) provided by the reporters in the case of User Reports and specifies a potential category of harmful content (Pornography, Hateful Content, ...), the label can also be provided by AI detection tools
- **Content Type** - The NIMA platform is content agnostic which means we support as many content types as your online platform needs (Image, Chat, Livestream Videos, Audio, ...)

This Query Based Builder UI offers the user the ability to add as many conditions as needed to filter incoming flags and redirect them to the specific moderation queue you are creating, or you want to create.

Moderation Actions

Moderation actions represent references to endpoints on a user's server. They allow users to determine actions (e.g., removing content) based on internal policies regarding user-generated content. These API endpoints are called every time a content moderator clicks on a corresponding action button.

The tool enables the user to configure actions and endpoints via two tabs: User and Post.

Configuring Actions are based on two inputs:

- Action name
- A URL that will be used to call an API endpoint on your server when an action is triggered.

TREMAU

Dashboard Analytics Reports List Appeal List My Queues Admin Settings User Management Policy Configuration Statement Of Reasons Queues Configuration Moderation Actions API Documentation NIMA User Guide

F v1.2.14
B v1.2.14

Moderation Actions

User Post

+ Create New

Action Name	Endpoint URL	Test	Edit	Delete	Statements of reason	Flags	Appeals
Restrict visibility of content	https://401ff9e6fc03354fba63255dd3106234.m.pipedream.net	Test	Edit	Delete	✓	✓	✓
Ban User for 10 days	https://401ff9e6fc03354fba63255dd3106234.m.pipedream.net	Test	Edit	Delete	✓	✓	✓
Suspend User account	https://401ff9e6fc03354fba63255dd3106234.m.pipedream.net	Test	Edit	Delete	✓	✓	✓
Permanent Chat Ban	https://401ff9e6fc03354fba63255dd3106234.m.pipedream.net	Test	Edit	Delete	✓	✓	✓
Permanent Game Ban	https://401ff9e6fc03354fba63255dd3106234.m.pipedream.net	Test	Edit	Delete	✓	✓	✓
24 Hour Game Ban	https://401ff9e6fc03354fba63255dd3106234.m.pipedream.net	Test	Edit	Delete	✓	✓	✓
7 day chat ban	https://401ff9e6fc03354fba63255dd3106234.m.pipedream.net	Test	Edit	Delete	✓	✓	✓

Figure 20: Screenshot of moderation action feature

LEA Portal Intake

This tool also provides the opportunity to integrate LEAs requests directly in the system. This allows it to handle legal enforcement requests about flagged content directly through the online platform. These requests are handled confidentiality and releases of non-public information about users to law enforcement officials is only given in response to appropriate legal process, such as a subpoena, court order, or search warrant – or in response to a valid emergency request.

9/8/23, 3:53 PM

Law Enforcement Form

**Law Enforcement Information Request Form**

Welcome to Tremau's legal request submission form. Tremau will only release non-public information about its users to law enforcement officials in response to appropriate legal process, such as a subpoena, court order, or search warrant – or in response to a valid emergency request.

Enter Your Agency or Office

Name of the agency or office requesting information

Enter Your Agency/Office Address**Enter Your Name**

First and last name

Enter Your Title

Your title in the agency or office requesting information

Enter Your Official Email

Your official/work email

Enter Your Official Phone Number<https://panel.tremau.net/law-enforcement/DMBhBdUMtcuKVzrskZccZKcwzylaX655/>

1/3

9/8/23, 3:53 PM

Law Enforcement Form

Enter Your City and State

Upload a Screenshot of Relevant Post

Upload a screenshot of the post you're seeking information regarding.



Drop or Select file

Enter Exact Post Text

If you're unable to provide a screenshot, please provide the exact text of the post you're seeking information regarding. If you included a screenshot, please ignore this field.

Select Information Request Type

Subpoena, search warrant, court order, etc.

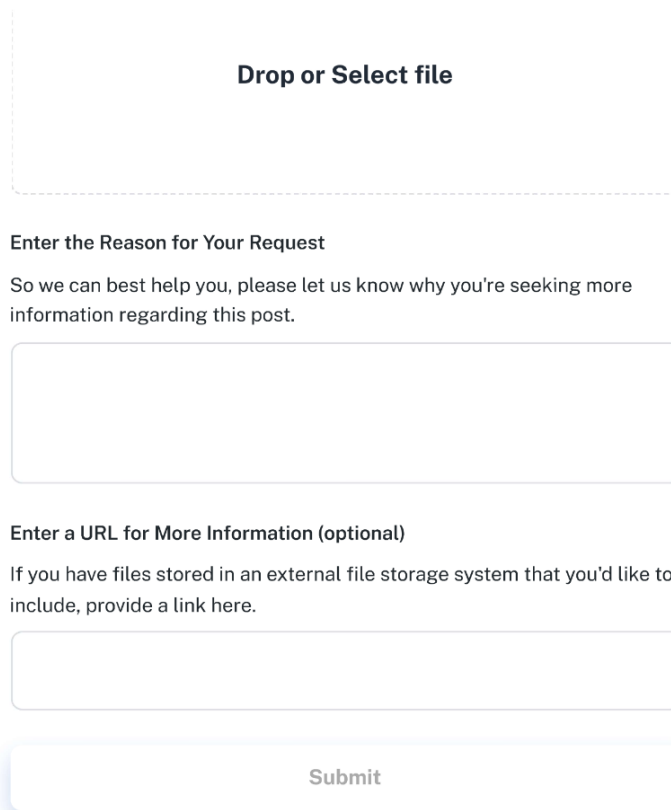
Upload Information Request Type Supporting Documentation

Example: PDF copy of search warrant



9/8/23, 3:53 PM

Law Enforcement Form



The form is titled "Law Enforcement Form". It features a dashed rectangular box with the text "Drop or Select file" centered inside. Below this is a section titled "Enter the Reason for Your Request" with a subtext: "So we can best help you, please let us know why you're seeking more information regarding this post." This is followed by a large, empty rectangular text input field. Below the input field is another section titled "Enter a URL for More Information (optional)" with a subtext: "If you have files stored in an external file storage system that you'd like to include, provide a link here." This is followed by another empty rectangular text input field. At the bottom of the form is a prominent, rounded rectangular button with the text "Submit" centered on it.

Figure 21: LEA request submission form

TCO transparency report

The tool enables the user to reduce its reporting burden by effortlessly producing compliant-by-design transparency reports in four simple steps.

Based on the type of report and time period specified, the tool will compile an easily comprehensible Transparency Report regarding the TCO Regulation. These reports are in a machine-readable format, which will encompass information pertaining to the management of user data and the removal of content, along with government inquiries for user records, various pertinent metrics and valuable insights. These reports can be made for specific time periods.

7 Access to FRISCO tools

Tools No 1 (self-assessment questionnaire) and No 2 (process map), are and will always be available as open-source software, with the specific licence accompanying the tools to be discussed amongst FRISCO partners. The source code for the development version of the tools is maintained and accessible at a dedicated [GitHub repository](#).

The current version of the two tools themselves are available to test and use at the following addresses:

- a. Self-assessment questionnaire: <http://78.46.226.222>
- b. Process map: <http://78.46.226.222/flowchart>

Only a description of the content moderation tool will be available on the FRISCO website. Interested parties will be encouraged to initiate contact with Tremau and undergo the requisite procedural step of signing a Software as a Service (SaaS) contract. This deliberate approach ensures that access is extended responsibly and securely, aligning with the commitment to providing a secure and tailored solution to HSPs within the FRISCO framework.

8 Conclusions and next steps

A toolkit composed of three user-friendly tools was developed to respond to the needs outlined in the mapping report of WP2. These tools aim to help HSPs in developing the right processes and implementing efficient methods not only to be compliant with new regulations but also to serve their own business needs.

The self-assessment questionnaire provides the first means for HSPs to understand how their current internal practices meet the requirements of the TCO. The tool gives them an overview of the main TCO Regulation requirements as well as a compliance score to help situate themselves within the compliance process.

The process map is an interactive tool that structures and describes the entire compliance process with the TCO Regulation and related duties for HSPs in a holistic way, so to say from the exposure to terrorist content to the transparency reports. The process map can be seen as a compliance workflow builder. It aims at helping HSP develop the processes to respond to removal orders, implement the policies and workflows to be compliant with the TCO, and gain better understanding about the TCO requirements in relation to their current practices.

Finally, the content moderation tool addresses user-generated content related risks and TCO compliance. Essentially, it provides a content moderation platform in which incoming flags from different sources of reports are aggregated. Moderators can subsequently view the reported cases and make related decisions (i.e, keep or remove content, signal user, ...) based on platform policies. The built-in TCO feature enables HSPs to subsequently create TCO transparency reports, as well as align with TCO requirements through increased internal process compliance. This tool is provided through cloud access to interested HSPs which can be integrated with their internal system.

Ultimately, these tools respond to the needs of HSPs regarding TCO compliance in two major ways: informational and practical. That is, they offer comprehensive and tailored information to TCO compliance as well as provide guidelines and resources to enhance internal processes for better content moderation practices. Following the initial testing phase where the tools will be tested by the different stakeholders of the FRISCO consortium, feedback will be given to WP2 participants. Subsequently, refinement of these tools is scheduled over the coming months to elevate the overall user experience and make potential adjustments in light of the testing results.

-----End of Document-----