



International Panel on the Information Environment

IPIE Submission to the European Commission

Implementation of Regulations for
Transparency Reporting Obligations Under
the Digital Services Act

Technical Paper 2024.1

IPIE SUBMISSION TO THE EUROPEAN COMMISSION ON THE IMPLEMENTATION OF REGULATIONS FOR TRANSPARENCY REPORTING OBLIGATIONS UNDER THE DIGITAL SERVICES ACT

Technical Paper TP2024.1

SYNOPSIS

This submission by the [International Panel on the Information Environment](#) (IPIE) sets out recommendations for the European Commission's consultation regarding the Implementation of Regulations for Transparency Reporting Obligations under the Digital Services Act (Regulation (EU) 2022/2085). Specifically, the submission recommends the following amendments to the proposed mandatory qualitative template for Articles 15(1)(c), 15(1)(e), 42(2)(a), and 42(2)(b) in Annex II to the Draft implementing Regulations:

- Summary of the content moderation engaged in at the providers' own initiative: This field should specifically contain information regarding content labelling, and any corrective information provided, or content or account moderation that has been and is being applied.
- Meaningful and comprehensible information regarding the applied detection method: This field should include information about the methods used for actions that flow from detection, such as taking content moderation measures, content labelling, and providing corrective information. The information should include the parameters used to detect systemic risks.
- Summary of the use made of automated means for the purpose of content moderation: This field should contain summary information regarding the parameters set for automated means, especially concerning how content is labelled and actions that flow from such labelling, such as flagging, fact-checking, de-amplification, and removal.
- Qualitative description of the automated means: This field should contain specific information regarding the parameters set for automated means, especially concerning how content is labelled and actions that flow from such labelling, such as flagging, fact-checking, de-amplification, and removal.
- Specification of the precise purposes to apply automated means: This field should contain information indicating which risks have been identified, that is, what specific problems the automated means is intended to address, not just the intended outcome of the use of automated means.

Introduction

This submission provides requested feedback on the online consultation for the European Commission's consultation regarding the Implementation of Regulations for Transparency Reporting Obligations under the Digital Services Act, specifically the proposed mandatory qualitative template for Article 15(1)(c), 15(1)(e), 42(2)(a) and 42(2)(b) in Annex II to the Draft implementing Regulations.

The IPIE is an international organization based in Switzerland that strives to empower policymakers, industry leaders, and civil society with invaluable insights for a better global information environment. With contributions from hundreds of leading scientists worldwide, the IPIE is uniquely placed to provide robust empirical evidence on information integrity in the global online information environment.

The submission draws on the IPIE's report *Strategies for Improving the Global Information Environment Results from a Systematic Review and Meta-Analysis* (IPIE, SFP2023.1). Based on the scientific findings of the IPIE, this submission recommends that the proposed qualitative template be amended in four areas: (1) summary of the content moderation engaged in at the providers' own initiative; (2) meaningful and comprehensible information regarding the applied detection method; (3) summary of the use made of automated means for the purpose of content moderation; (4) qualitative description of the automated means; and (5) specification of purposes to apply automated means

In conclusion, the IPIE welcomes the Draft implementing Regulations as an important step towards a better information environment and the recommendations herein are designed to strengthen the effectiveness of the Regulations by including information regarding measures that have been endorsed by the scientific community.

1. Legal context

This submission considers the proposed mandatory qualitative template for Article 15(1)(c), 15(1)(e), 42(2)(a) and 42(2)(b) in Annex II to the Draft implementing Regulations. According to the Annex:

“Article 15(1)(c) prescribes that ... the transparency report [of intermediary services] shall include meaningful and comprehensible information about the content moderation engaged in at the provider’s own initiative.

Article 15(1)(e) prescribes that the transparency report shall include information on any use made of automated means for the purposes of content moderation, including a qualitative description, a specification of the precise purposes, and any safeguards applied.”

Article 15 Transparency reporting obligations for providers of intermediary services

1. Providers of intermediary services shall make publicly available, in a machine-readable format and in an easily accessible manner, at least once a year, clear, easily comprehensible reports on any content moderation that they engaged in during the relevant period. Those reports shall include, in particular, information on the following, as applicable:

...

(c) for providers of intermediary services, meaningful and comprehensible information about the content moderation engaged in at the providers’ own initiative, including the use of automated tools, the measures taken to provide training and assistance to persons in charge of content moderation, the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients’ ability to provide information through the service, and other related restrictions of the service; the information reported shall be categorised by the type of illegal content or violation of the terms and conditions of the service provider, by the detection method and by the type of restriction applied;

...

(e) any use made of automated means for the purpose of content moderation, including a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the automated means used in fulfilling those purposes, and any safeguards applied.

“Article 42(2)(a) prescribes that providers of very large online platforms shall include in their transparency report information on the human resources that the provider dedicates to content moderation in respect of the service offered in the Union. Article 42(2)(b) prescribes that providers of very large online platforms shall include in their transparency report information on the qualifications and linguistic expertise of the person carrying out the activities referred to in Article 42(2)(a), as well as training and support given to staff.”

Article 42 Transparency reporting obligations

1. Providers of very large online platforms or of very large online search engines shall publish the reports referred to in Article 15 at the latest by two months from the date of application referred to in Article 33(6), second subparagraph, and thereafter at least every six months.
2. The reports referred to in paragraph 1 of this Article published by providers of very large online platforms shall, in addition to the information referred to in Article 15 and Article 24(1), specify:
 - (a) the human resources that the provider of very large online platforms dedicates to content moderation in respect of the service offered in the Union, broken down by each applicable official language of the Member States, including for compliance with the obligations set out in Articles 16 and 22, as well as for compliance with the obligations set out in Article 20;
 - ...
 - (c) the indicators of accuracy and related information referred to in Article 15(1), point (e), broken down by each official language of the Member States. The reports shall be published in at least one of the official languages of the Member States...

2. The Draft Mandatory Qualitative Template

The Draft implementing Regulations will give effect to the objectives set out in Recital 2 of the DSA to ensure “a transparent and safe online environment.” This includes requirements for the publication of information by Internet intermediaries that is machine-readable and broken down per calendar month, and for very large online platforms and very large search engines, set out in biannual reports.

Article 1 of the Draft implementing Regulations specifies that pursuant to Articles 15, 14, and 42 of the DSA, a qualitative and a quantitative template should be used. Both templates must be completed to fulfil the transparency reporting requirements.

The draft mandatory qualitative template set out in Annex II asks for the following information:

Information	Instructions
Summary of the content moderation engaged in at the providers' own initiative	This field must contain a summary of the content moderation the provider has engaged in during the reporting period on its own initiative. This summary must, as a minimum, include a description of the types of restrictions applied by the provider. If the provider uses soft moderation restrictions, such as the attachment of labels to or warnings about certain information, an outline of these policies must also be included.
Meaningful and comprehensible information regarding the applied detection method	This field must provide meaningful and comprehensible information regarding the detection methods the provider has used to detect information that is allegedly illegal or incompatible with the platform's terms and conditions.
Updates to the terms and conditions	This field must outline the updates the provider has made to the terms and conditions during the reporting period that affect or have affected the provider's own initiative content moderation.
Measures taken to provide training and assistance to persons in charge of content moderation	This field must provide meaningful and comprehensible information on the measures that have been taken during the reporting period to provide training and assistance to persons in charge of content moderation, including both internally and externally employed content moderation.
Summary of the use made of automated means for the purpose of content moderation	This field must contain a summary of the use made of automated means for the purpose of content moderation during the reporting period.
Qualitative description of the automated means	This field must provide a qualitative description of the automated means used for the purpose of content moderation as applied by the provider.
Specification of the precise purposes to apply automated means	This field must provide the specification of the precise purposes outlined by the provider to apply automated means during any parts of the content moderation process.
Safeguards applied to the use of automated means	This field must indicate the safeguards that the provider applied alongside the use of automated means in the content moderation process.
Summary of the content moderation governance structure	This field must contain a summary of the governance structure of the content moderation teams both internally employed or externally contracted, both internally and externally, by the service provider. The summary must, as a minimum, include the hierarchy of the content moderation teams, their locations, and the teams' capacities.
Qualifications of the human resources dedicated to content moderation	This field must provide information on the qualifications of the staff dedicated to content moderation, including the level as defined by the European Qualifications Framework.
Training given to human resources dedicated to content moderation	This field must provide information on the type, frequency, and content of the training given to the staff dedicated to content moderation.
Support given to human resources dedicated to content moderation	This field must provide information on the support, including mental and physical, at the disposal of the staff dedicated to content moderation.

3. Risk Mitigation

The IPIE supports the aim of the draft qualitative template in Annex II, but is concerned that the information asked for will not adequately capture information regarding risks and risk

assessments. Although Articles 15 and 42 do not directly mention risks, the reference to “content moderation,” which is often used as a risk mitigation measure, suggests that risk should be part of the appraisal.

Articles 34 and 35 of the DSA set out the risk assessment and the mitigation requirements for very large online platforms and very large online search engines. Article 34 specifies that risk assessments should identify systemic risks such as dissemination of illegal content; foreseeable negative impacts on fundamental rights, public discourse, elections, and security; foreseeable risks of gender-based violence; and foreseeable risks to public health, to minors, and to the physical and mental well-being of individuals (see Recital 79). Article 35 specifies that these risks should be mitigated, and it sets out a list of possible mitigation measures, including content moderation. Thus, it is our understanding that content moderation, as set out in Article 3, is intended to be used to eliminate or mitigate these risks according to the principle of proportionality.

There is therefore an overlap between the information required for the risk assessments pursuant to Articles 34 and 35 and the transparency obligations pursuant to Articles 15(1)(c), 15(1)(e), and 42(2)(a). It is the IPIE’s submission that this link should be made more explicit in the draft mandatory qualitative template.

This recommendation for a more explicit link between the transparency requirements in Articles 15 and 42 and the risk assessment requirements draws on the approach taken by the European Commission in its proposed Baseline Framework for Digital Services Risk Management (European Commission (August 2023)). This Risk Management Framework resulted from the Commission’s investigation into Russian disinformation campaigns targeted at Ukraine in 2022. The purpose of the investigation was “to develop methodologies for risk assessments and metrics for risk mitigation under the DSA framework (Articles 34 and 35)” (European Commission (August 2023), p. 12).

The Commission’s investigation examined online platforms’ practices before the DSA took effect, and it found that the platforms did not implement measures on a systemic level and did not have policies that effectively reduced risks overall (European Commission (August

2023), p. 20). In response to these findings, the Commission proposed a Risk Management Framework intended to offer the following:

1. “A qualitative system for evaluating the severity of a particular risk factor in context;
2. A quantitative method for determining the scale and intensity of the risk factor;
3. A qualitative method for identifying whether or not a platform has designated a mitigation measure for a particular risk; and
4. A set of quantitative techniques for assessing the effectiveness of mitigation measures deployed to address these risks” (European Commission (August 2023), p. 13).

Specifically, the Risk Management Framework contains the following assessment and mitigation metrics:

Assessment metrics	Mitigation metrics
Audience size	Speed and consistency of removal
Exposure	De-amplification
Engagement	Non-follower engagement
Prevalence	Consistency of labelling
Amplification factor	Responsiveness to user notification
Volume	Redress of denial of service
Rechanneling	Restrictions on inauthentic behavior
Toxicity	Restriction on algorithmic exploitation
	Denylisting URLs

The benefit of the framework is that it “has the potential to reveal gaps in existing platform policies, enforcement processes, transparency reports, and the kind of data that is available to researchers and regulators” (European Commission (August 2023), p. 23). As will be seen below, the IPIE has examined the effectiveness of similar and overlapping mitigation metrics. Thus, the IPIE’s identification of countermeasures is already present in the EU context in relation to the DSA.

The IPIE has identified an emerging scientific consensus that endorses content labelling, the provision of corrective information, content or account moderation, and information and media literacy as effective strategies for improving the information environment (IPIE, SFP2023.1, p. 5).

The IPIE defines these countermeasures as follows:

Countermeasure	Description	Example
Content labelling	Labelling posts, accounts, and stories with tags about fact-checking, funding, or advertising, or any other forms of tagging or flagging, including providing further content without the user having to click through to receive the additional information.	A platform adds a “disputed” label to a user post, or a platform labels posts by state media with a “warning” sign.
Corrective information	An organization, platform, or individual provides accurate information without regard to whether users have preconceptions about it.	Governments, private enterprises, or users publicly debunk a rumor or conspiracy on social media in a separate, unlinked piece of content.
Content or account moderation	Taking down or marking content; using human or algorithmic moderation to suspend and block accounts.	YouTube downranks content, or X (formerly Twitter) reduces interactions with accounts that users don’t follow.
Information and media literacy	Educating users to identify misinformation by giving them tips or suggestions or by training them.	Facebook offers Tips to Spot False News, including “be skeptical of headlines,” “look closely at the URL,” and “investigate the source.”

Considerable overlap can therefore be found between the European Commission’s Risk Assessment Framework, discussed above, and the findings of the IPIE.

The IPIE also contends that the elements of the Commission’s Risk Management Framework should be included in the draft mandatory qualitative template in Annex II in the manner set out below.

4. Bridge between Articles 15 and 42 and Articles 34 and 35

The IPIE considers that the information about the use and effectiveness of the mitigation strategies set out above should not be provided in relation to the publication of risk assessments. The information that will be provided in the risk assessments that relate to risk should also be part of the transparency obligations under Articles 15(1)(c), 15(1)(e), 42(2)(a), and 42(2)(b). Such information should therefore be included in the draft mandatory qualitative template in the Draft implementing Regulations.

Further, scientific experts affiliated with the IPIE have considered numerous countermeasures, but due to a lack of information and research they have not been able to reach a consensus regarding their effectiveness. These countermeasures include advertisement policies, content reporting, content user sharing, disclosure, redirection, security or verification, and self-fact checking. It is therefore vital for the advancement of scientific understanding of the information environment that very large online platforms and

very large search engines disclose information in a timely fashion. Waiting for 15 months for the publication of risk assessments erects an unreasonable barrier to researcher access to data.

Excluding information pertaining to risk and risk mitigation from the draft mandatory qualitative template for Articles 15(1)(c), 15(1)(e), 42(2)(a), and 42(2)(b) would significantly reduce the utility of the information.

The IPIE therefore proposes the following amendments to the draft mandatory qualitative template:

Information	Instructions	IPIE recommendations
Summary of the content moderation engaged in at the providers' own initiative	This field must contain a summary of the content moderation the provider has engaged in during the reporting period on its own initiative. This summary must, as a minimum, include a description of the types of restrictions applied by the provider. In case the provider makes use of soft moderation restrictions, such as the attachment of labels or warnings to certain information, an outline of these policies must also be included.	This field should specifically contain information regarding content labelling, any corrective information provided, or any content or account moderation that has been and is being applied.
Meaningful and comprehensible information regarding the applied detection method	This field must provide meaningful and comprehensible information regarding the detection methods the provider has used to detect information that is allegedly illegal or incompatible with the platform's terms and conditions.	This field should include information on the methods used for actions that flow from detection, such as content moderation measures, content labelling, and providing corrective information. The information should include the parameters used to detect systemic risks.
Updates to the terms and conditions	This field must outline the updates the provider has made to the terms and conditions during the reporting period that affect or have affected the provider's own initiative content moderation.	No comment.
Measures taken to provide training and assistance to persons in charge of content moderation	This field must provide meaningful and comprehensible information on the measures that have been taken during the reporting period to provide training and assistance to persons in charge of content moderation, including both internally and externally employed content moderation.	No comment.
Summary of the use made of automated means for the purpose of content moderation	This field must contain a summary of the use made of automated means for the purpose of content moderation during the reporting period.	This field should contain summary information regarding the parameters set for automated means, especially concerning how content is labelled and actions that flow from such labelling, such as flagging, fact-checking, de-amplification, and removal.

Qualitative description of the automated means	This field must provide a qualitative description of the automated means used for the purpose of content moderation as applied by the provider.	This field should contain specific information regarding the parameters set for automated means, especially concerning how content is labelled and actions that flow from such labelling, such as flagging, fact-checking, de-amplification, and removal.
Specification of the precise purposes to apply automated means	This field must provide the specification of the precise purposes outlined by the provider to apply automated means during any parts of the content moderation process.	This field should contain information indicating which risks have been identified, that is, what specific problems the automated means is intended to address, not just the intended outcome of the use of automated means.
Safeguards applied to the use of automated means	This field must indicate the safeguards that the provider applied alongside the use of automated means in the content moderation process.	No comment.
Summary of the content moderation governance structure	This field must contain a summary of the governance structure of the content moderation teams both internally employed or externally contracted, both internally and externally, by the service provider. The summary must, as a minimum, include the hierarchy of the content moderation teams, their locations, and the teams' capacities.	No comment.
Qualifications of the human resources dedicated to content moderation	This field must provide information on the qualifications of the staff dedicated to content moderation, including the level as defined by the European Qualifications Framework.	No comment.
Training given to human resources dedicated to content moderation	This field must provide information on the type, frequency, and content of the training given to the staff dedicated to content moderation.	No comment.
Support given to human resources dedicated to content moderation	This field must provide information on the support, including mental and physical, at the disposal of the staff dedicated to content moderation.	No comment.

References

European Commission, August 2023. *Digital Services Act: Application of the Risk Management Framework to Russian Disinformation Campaigns*.

European Commission, 2023. Draft Implementing Regulation – Ares(2023)8428591: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-_en.

International Panel on the Information Environment, 2023. *Strategies for Improving the Global Information Environment: Results from a Systematic Review and Meta-Analysis*. SFP2023.1. Zurich, Switzerland: IPIE. <https://www.ipie.info/research/sfp2023-1>.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance), PE/30/2022/REV/1, OJ L277. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065>.

Contributors

Drafting Author: Ann Kristin Glenster (Consulting Scientist, Norway/UK). Reviews: Philip Howard (IPIE President, Canada/UK), James Ward (Researcher, USA/UK). Copyediting: Beverley Sykes.

Copyright Information



This work is licensed under an Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Technical Papers

Technical Papers are prepared on specialized topics for which an independent scientific review or technical perspective is deemed essential. Most Technical Papers are used to inform scientific consensus building. Some Technical Papers are prepared as submissions to, or at the request of, international organizations or governments. Technical Papers are produced by individual authors at the request of the IPIE, but any opinions, findings, conclusions, or recommendations expressed in Technical Papers are those of individual authors and may not reflect the views of the international scientific community.

About the IPIE

The International Panel on the Information Environment (IPIE) is an independent and global science organization committed to providing the most actionable scientific knowledge about threats to the world's information environment. Based in Switzerland, the mission of the IPIE is to provide policymakers, industry, and civil society with independent scientific

assessments on the global information environment by organizing, evaluating, and elevating research, with the broad aim of improving the global information environment. Hundreds of researchers from around the world contribute to the IPIE's reports. www.IPIE.info.

For more information, please contact the International Panel on the Information Environment (IPIE), secretariat@IPIE.info. Seefeldstrasse 123, P.O. Box, 8034 Zurich, Switzerland.



International Panel on
the Information
Environment

Seefeldstrasse 123
P.O. Box 8034 Zurich
Switzerland

