CONTENT & JURISDICTION PROGRAM

TOOLKIT

CROSS-BORDER CONTENT MODERATION

MARCH 2021
I&JPN REF: 21-104
www.internetjurisdiction.net/content/toolkit

INTERNET & JURISDICTION POLICY NETWORK

The Internet & Jurisdiction Policy Network is the multistakeholder organization fostering legal interoperability in cyberspace. Its stakeholders work together to preserve the cross-border nature of the internet, protect human rights, fight abuses, and enable the global digital economy. Since 2012, the Internet & Jurisdiction Policy Network has engaged more than 400 key entities from six stakeholder groups around the world including: governments, the world's largest internet companies, the technical community, civil society groups, leading universities and international organizations.

DESIGN & LAYOUT
João Pascoal Studio

www.joaopascoal.com

CITATION
Internet & Jurisdiction Policy Network Toolkit
Cross-border Content Moderation (2021)

TABLE OF CONTENTS

1. ISSUE FRAMING

Every day, hundreds of millions of posts and hundreds of thousands of hours of videos are uploaded on major internet platforms and made globally accessible, greatly facilitating freedom of expression. At the same time, legitimate concerns are raised regarding increasing harmful behaviors, including hate speech, harassment, security threats, incitement to violence, or discrimination.

Protecting human rights and freedom of expression when dealing with such abuses on the internet is a major transnational challenge. In the absence of clearly agreed substantive and procedural frameworks to handle the disparity of national laws content legal in one country can be illegal in another one.

Moreover, although potentially abusive content remains an extremely small proportion of the massive amount posted, the absolute amount of individual restriction decisions to be made is nonetheless unprecedented. Case-by-case determinations need to carefully account for context and intent, but within very limited resources and response times given viral propagation.

Protecting human rights and freedom of expression when dealing with such abuses on the internet is a major transnational challenge. In the absence of clearly agreed substantive and procedural frameworks to handle the disparity of national laws content legal in one country can be illegal in another one.

In this context, opposing demands are made regarding the expectations of service providers: one asking them to thoroughly police content posted on their platforms to guarantee the respect of national laws and protect their users; and the other objecting to them making determinations on their own and exercising proactive content monitoring, for fear of detrimental human rights implications.

Clear common guidelines and due process mechanisms are needed to address this common challenge for all actors, maximizing the necessary remediation of harm and minimizing restrictions to freedom of expression.

States around the world are increasingly adopting legislations regulating online content and service providers. Some of these laws are putting on service providers the responsibility of determining illegal content, others aim to regulate the whole sector. The resulting fragmented regulatory environment imposes potentially conflicting responsibilities for service providers providing their services across multiple jurisdictions.

At the same time, service providers are developing more detailed terms of service to deal with content or behaviour that they do not want on their platform or that may be illegal. In order to deal with the immense quantity of user-generated content, service providers have also developed flagging tools, hash databases and algorithms for the identification and removal of illegal or problematic content.

Enabling interoperability and coexistence between such heterogeneous governance frameworks can reconcile the need for collective solutions with the recognition of the autonomy of actors, as well as the diversity of their cultural references and normative authority. It can provide solutions that are as distributed and scalable as the internet itself.

This requires: communication between all stakeholders to help them understand each other's situation, concerns and intentions; agreed norms of behavior to foster informal or structured coordination; and processes to develop practical cooperation mechanisms.

The Content & Jurisdiction Program Contact Group, consisting of experts from governments, internet companies, technical operators, civil society, leading universities and international organizations has, over the years, identified the key issues that could structure new models of transnational cross-border content moderation.

A common objective of the different actors should be the definition of high substantive and procedural standards regarding:

- Applicable substantive norms, including the interplay between agreed international and regional human rights, national laws, and companies' community guidelines,
- The respective obligations of states and the respective responsibilities and protections of other actors, including the identification of allegedly illegal content,
- Decision-making, standards and procedures, including the escalation path for individual decisions and appeal mechanisms,
- Legitimate purposes, necessity and proportionality regarding the geographic scope of restrictions,
- The necessary due process and transparency standards that should be applied across borders.

2. I&JPN METHODOLOGY

The Internet & Jurisdiction Policy Network fosters a new approach to transnational policy-making. Its innovative methodology identifies relevant stakeholders to define common problems and produce solutions to pressing and complex policy challenges. The neutral and replicable approach, structures interactions among diverse policy actors who would normally not have the opportunity to work together on practical and concrete outcomes.

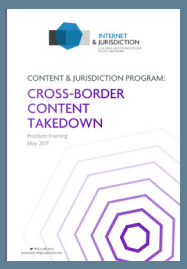
Since 2016 in regular iterations, the Content & Jurisdiction Program Contact Group engages a select set of these global policy actors while trying to ensure balanced geographical representation from governments, internet companies, technical operators, civil society, leading universities and international organizations. Using the I&JPN Methodology, Contact Groups have iteratively developed concrete outcomes pertaining to specific facets of cross-border content moderation and restriction challenges. Based on this methodology, future Contact Groups will continue to develop specific policy outcomes on focused issues while also addressing emerging challenges.

Meet the Contact Group members from 2018 - 2020 here

The Internet & Jurisdiction Policy Network fosters a new approach to transnational policy-making. Its innovative methodology identifies relevant stakeholders to define common problems and produce solutions to pressing and complex policy challenges.

FRAMING COMMON PROBLEMS

Issues can best be addressed when formulated as problems that stakeholders have in common rather than with one another. As a first step stakeholders are consulted to develop a shared framing of the issue at hand and build a shared vernacular. This helps develop a common understanding of the policy problem and helps identify key areas for cooperation where stakeholders can work collaboratively to develop practical and operational solutions.



How can we manage globally available content in light of the diversity of local laws and norms applicable on the internet?

I&JPN Content & Jurisdiction Framing Paper (2017)[i]

SETTING COMMON OBJECTIVES

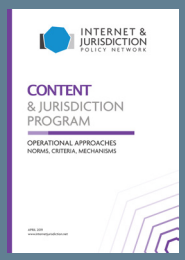
Based on these areas of cooperation, a dedicated Contact Group, guided by a neutral and independent coordinator, identifies key structuring questions that guide discussions amongst stakeholders and provide a framework within which concrete policy solutions can be developed. These discussions documented as Policy Options define common objectives to ensure better policy coherence and structure further work.

This document aims at providing, through a forward-looking approach, guiding elements to structure further discussion on possible frameworks regarding transnational content restrictions. It documents the key substantive and procedural dimensions that can help overcome divergences regarding the responsibilities of intermediaries.

I&JPN Content & Jurisdiction Policy Options (2018)[ii]

DEVELOPING COMMON APPROACHES

Based on the objectives identified, intense work in the Contact Group aims to develop scalable, interoperable policy solutions. These can take the form of Operational Norms – to help actors organize their own behavior and mutual interactions; Operational Criteria – to guide actors who develop, evaluate & implement solutions; and Operational Mechanisms – that offer concrete avenues for cooperation.



The work of the dedicated Contact Group of the Internet & Jurisdiction Policy Network, aims to contribute to policy discussion by addressing the key elements of a general framework regarding responsible content moderation and restrictions.

I&JPN Content & Jurisdiction Operational Approaches (2019)[iii]

FOSTERING LEGAL INTEROPERABILITY

Further work is conducted to evangelize, communicate and aid the implementation of these policy solutions. This may take the form of Toolkits compiling thematic Outcomes developed by the Contact Group. This helps further legal interoperability in two dimensions:

- **Interoperability between actors**: to enable automation of the technical workflow among public authorities and private actors across borders to ensure due process at scale.
- **Interoperability between norms**: to reduce the potential of conflicts in rule-setting, implementation and enforcement among different regimes.

i. https://www.internetjurisdiction.net/uploads/pdfs/Papers/Content-Jurisdiction-Program-Paper.pdf
ii. https://www.internetjurisdiction.net/uploads/pdfs/Papers/Content-Jurisdiction-Policy-Options-Document.pdf
iii. https://www.internetjurisdiction.net/uploads/pdfs/Papers/Content-Jurisdiction-Program-Operational-Approaches.pdf

The Cross-border Content Moderation Toolkit frames approaches towards identification and reporting of problematic online content and builds a common understanding of the requisite processes that can ensure due process. This resource can be useful to service providers, in the design of their content moderation activities and notifiers in the detection and reporting of problematic or abusive content. It can also help legislators and policy-makers determine procedures for dealing with different types of content and abusive behaviour. This Toolkit provides tools that seek to help improve the interactions between the different actors to act on abusive content while also strengthening corresponding procedures and mechanisms to promote freedom of expression online. The Content & Jurisdiction Program Contact Group will continue to engage on the topics addressed in the Toolkit with the objective of refining them and developing new tools.

This Toolkit provides resources that seek to help improve the interactions between the different actors to act on abusive content while also strengthening corresponding procedures and mechanisms to promote freedom of expression online.

The subsequent components of this Toolkit are a joint contribution by some of the most engaged experts in this field to advance the ongoing debate on the complex issues of cross-border content moderation. They should however not be understood as the result of a formal negotiation validated by these Members' organizations. They are a best effort by the Members of the Program's Contact Group to address the important cross-border issues pertaining to content moderation that have been curated by the I&JPN Secretariat into the framework of this Toolkit.

3.

TOOLKIT CROSS-BORDER CONTENT MODERATION

STRUCTURE

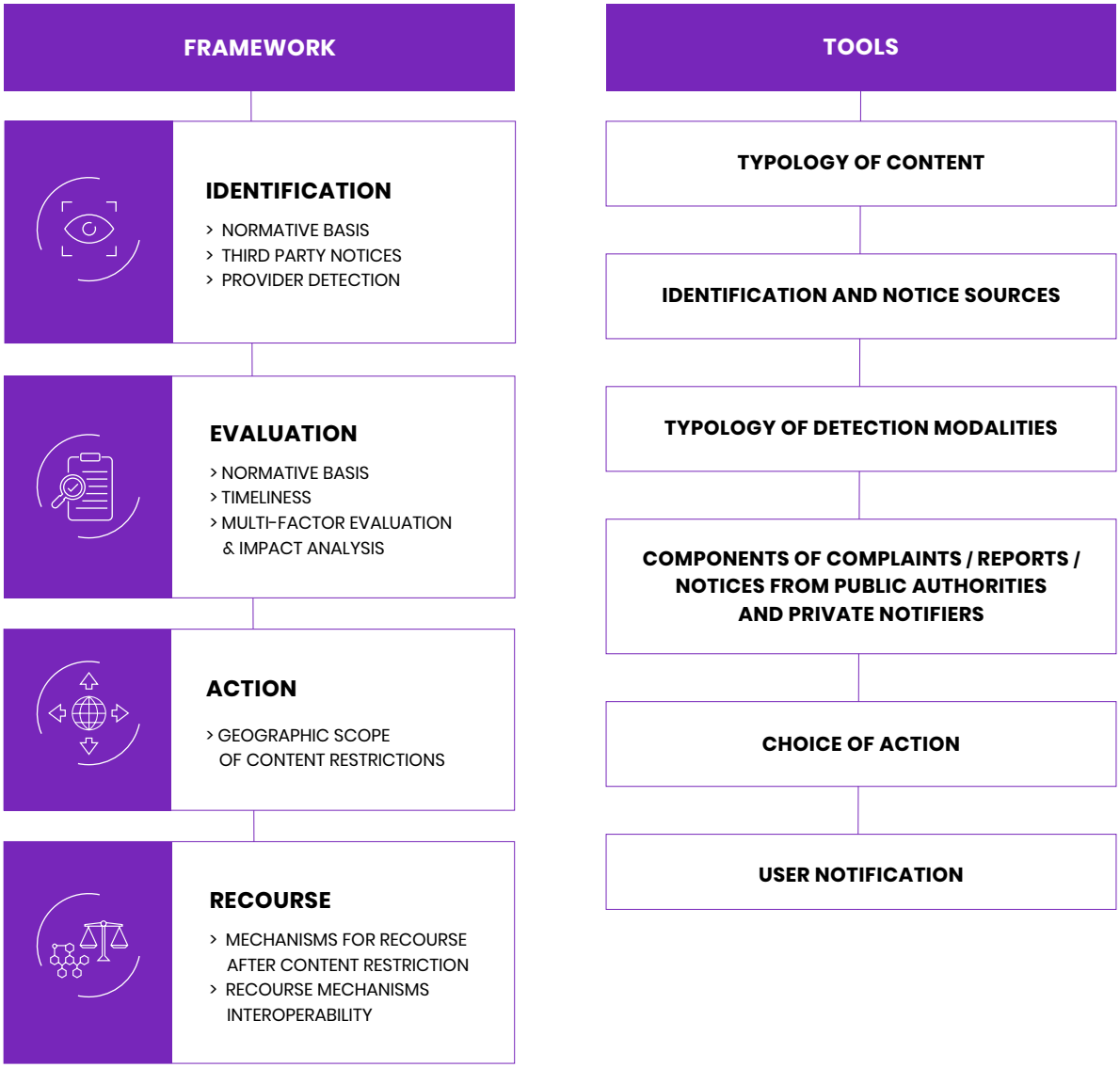
FRAMEWORK

TOOLS

STRUCTURE

The following Toolkit curates resources that practitioners can use in their everyday work to guide the process of content moderation while strengthening due process. These tools have been developed by the multistakeholder Content & Jurisdiction Program Contact Group throughout 2019-20 and also draw on the Operational Approaches document* published by the Contact Group in April 2019.

This Toolkit has a twofold structure. The first section provides a framework for understanding and approaching the four different stages in the process of content moderation (Identification, Evaluation, Choice of Action and Recourse). The second section contains practical tools that can help resolve some of the challenges of cross-border content moderation.



*Content & Jurisdiction Operational Approaches (2019)

3.1

FRAMEWORK

IDENTIFICATION & NOTIFICATION

EVALUATION

ACTION

RECOURSE

IDENTIFICATION

The content moderation process begins with the identification of unacceptable, abusive, harmful or even illegal content on the digital premises of an online intermediary. While globally there are differences between what is illegal, considered harmful or unacceptable, and each platform has different rules, there are many types of prohibited content or behaviour on which there is a degree of coherence. The resources in this section help to develop a common vocabulary for unacceptable, abusive harmful or illegal content. These resources also document the means and sources for identification of such content as well as the normative bases for such identification.

Initial identification can come from a wide variety of sources that can be grouped into either the intermediaries' own detection sources or coming from third parties. The following resources provide an overview of the sources of such identification. A comprehensive typology of such sources/actors can be found in the tools section.

I&JPN REF. 19-116

NORMATIVE BASIS

A major challenge in the drafting and implementation of domestic laws, as well as companies' community guidelines[1], is the need to reconcile competing rights, namely freedom of expression and the prevention of harm.

1. Reconciling diverse normative bases

A plurality of sources form an increasingly elaborate normative landscape, combining:

a. Overarching international or regional human rights principles, in particular:
   i. The Universal Declaration of Human Rights (UDHR);
   ii. The International Covenant for Civil and Political Rights (ICCPR), in particular articles 6, 17, 18, 19, 20, 24, 26;
   iii. The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD);
   iv. The Declaration of the Rights of the Child.

b. The diversity of applicable national and regional laws, either existing[2] or newly drafted for the digital context because of growing concerns regarding abusive online content.

c. The increasing importance of companies' terms of service and community guidelines.

2. International normative coherence

The following categories can be used to ascertain the degree of normative coherence on illegal content across jurisdictions. The frontiers between these different categories are however not rigid. Debates exist regarding where some topics fall.

a. There IS universal agreement that the content/behavior is illegal AND there is strong substantive convergence around the world on the corresponding threshold criteria (example: child sexual abuse material);

b. There IS universal agreement that the content/behavior is illegal, BUT significant national variations exist in the criteria determining illegality (example: defamation);

c. The content/behavior IS NOT universally considered as illegal, BUT the application of specific domestic laws on the local territory is considered acceptable by other countries, in particular for historic reasons (example: criminalization of Holocaust denial);

d. The content/behavior IS NOT universally considered as objectionable AND some countries even consider that it should not be allowed to make it illegal (example: laws discriminating against or criminalizing certain sexual orientation).

1. A comprehensive typology of such sources can be found in the tools section (of this Toolkit).

2. This may include relevant laws related to media as applicable.

3. Types of regulation

Some providers aim for community guidelines as uniform as possible for all their users. This produces a de facto global harmonization of applicable rules on their respective spaces. Others however rely on community-driven moderation, for instance organized by topic (e.g. Reddit) or by language (e.g. Wikipedia).

Private providers have sometimes initiated rules on new issues, e.g. on non-consensual posting of adult content (aka "revenge porn"), with some legislative initiatives emerging as a result.
A conceptual framework could be envisaged to distinguish more clearly between:

a. Regulation/moderation **ON** a platform: under the responsibility of the administrators of sub-forums and groups, such group rules can be more restrictive than the rules of the overall platform.

b. Regulation **BY** a platform: community guidelines and decision-making rules establish the general framework for all content on the corresponding space, including potentially the latitude given to group administrators.

c. Regulation **OF** platforms: domestic laws, regional legislation or international agreements defining the general responsibilities of providers in terms of content moderation, including how to reconcile their business capacity to determine their terms of service and the duties that could result from an extensive market position.

THIRD-PARTY NOTICES

Service providers receive notices of potentially unacceptable, abusive, harmful or even illegal content from external actors. These notifications can originate from either public authorities or private individuals and entities.

1. Public authorities issue:

a. Formal orders on the basis of domestic laws. However, due to constraints of volume and timely action, this is sometimes done without validation by a local court that could clearly establish the illegality of the content with all appropriate procedural guarantees. This puts on private entities the responsibility to make this determination, with incentives to over-restrict content in situations of uncertainty. Additional procedures (with appropriate protections) for expedited domestic evaluation could be developed.

b. More informal requests on the basis of terms of service or community guidelines, for instance through so-called internet referral units. Clearer procedures are necessary to ensure transparency and accountability regarding the use of such a channel by public authorities.

2. Private notices can come from:

a. Specialized notifiers, for instance for copyright or child abuse material. More clarity is however needed regarding, inter alia, their procedures, decision-making criteria, due diligence requirements and avenues for recourse if their notices are to be taken prima facie.

b. Media, inter alia for fact-checking (e.g. during electoral periods), or reporting cyber harassment against journalists targeted because of their professional activities.

c. Individual flaggers (including "trusted flaggers") via platform tools. The role of user reports should be seen alongside the increasing role of automated detection means for means for identifying and removing violating content. Easy-to-use flagging tools remain nonetheless necessary.

I&JPN REF. 19-118

PROVIDER DETECTION

Responding to pressure, major providers increasingly implement proactive detection and this can only be done through intensive use of algorithmic tools, including artificial intelligence. The use of hash databases to prevent re-upload of content previously detected as justifying restriction is also spreading.

The performance of such tools however strongly varies according to the different types of problematic content: performing well for images with easily recognizable elements, it remains much less accurate for anything requiring a strong evaluation of context.

In spite of significant progress, major challenges must be addressed, as automated tools:

1. Still lack the required accuracy to detect all infringing content or correctly identify objectionable content, risking under-or over-restriction.
2. Largely ignore contextual considerations, including external context, culture, and intention.
3. Risk making decisions without a proper balance between competing interests, by ignoring legal rules, legal interpretations, nuances in platform terms of service, et cetera.
4. Raise serious transparency issues: automated removal or restriction may provide insufficient information about its rationale, making it difficult to either understand the restriction decision or challenge it.
5. May still be circumvented through technical means, for example, by changing metadata or encrypting content, thus allowing certain harmful content to stay online or be viralized further through encrypted channels.
6. Can exhibit undetected biases due to the datasets they were trained on.

Human review remains a necessity in the decision-making process for individual restrictions and a significant collaborative effort is needed more generally to allow proper evaluation and oversight of algorithmic tools.

EVALUATION

Once potentially abusive or illegal content has been identified (through internal or external sources), the next step is an in-depth evaluation of said content towards making a decision on whether the content merits action. This part of the framework sets out the challenges that need to be considered, such as:

- Timeliness, i.e. the tension between the need for quick action versus ensuring the respect of due process;
- The different factors that need to be considered to evaluate whether an action is justified, including its potential impact(s).

TIMELINESS

Recent legislative efforts have put an increasing emphasis on short response times[3] for removing specific types of content, in particular regarding terrorism and violent extremism.

1. Tensions

a. Response times can be measured by reference to different operational events, including: time of upload, notification to the platform or notification to the user. Clarity regarding this factor is necessary in any rule establishing compulsory response times.

b. The tension between short response times, accuracy of the measure and the need to ensure the protection of rights can be summarized in the following statements:
- i. Some decisions on content restrictions have to be made quickly to prevent harm,
- ii. The faster the decision, the greater the risk of errors or inaccuracies in restriction decisions, or its impact on users' rights,
- iii. Ensuring accuracy of the decision and the full respect of users' rights and interests requires careful evaluation and thus time.

2. Rationale supporting quick decisions

There are a number of reasons incentivizing quick decisions:

a. Normative incentives, including the respect of national laws, the limitation of service provider liability, and the prospect of fines.

b. Economic incentives, including the volume of content and requests to be handled (as a matter of resources), the satisfaction of users and other interested parties (such as advertisers).

c. Operational considerations, including, the nature and volume of content and restriction requests, the distinction between clear cut and harder cases, and the consideration of harm from restriction versus harm from keeping the content accessible.

d. Interest considerations, including for whom the rationale for restrictions are more pressing: respect of national laws, or national security, may be more pressing as an interest for governments than for users.

3. Potential risks of quick decisions

a. Incorrect decision-making:
- i. False positives, leading to over-restriction. These include wrongly identified infringing content based on competing values (e.g. nudity as art vs. norms against nudity), contextual analysis (i.e. external situations where there is no offense from certain content, e.g. breastfeeding), analytical errors (i.e. content was misidentified).
- ii. False negatives, leading to content unduly remaining accessible, thus materializing harm, as the counterpart to false positives.

3. The case of live-streaming has not been specifically addressed, and would be worthy of a dedicated discussion.

b. Procedural fairness:
   i. Limited capacity of users to challenge a decision before the restriction takes place. Ex ante capacity to contest is time intensive. These risks differ with regard to the type of content and related harms.
   ii. Limited transparency to challenge restriction decisions before or after they take place, when automated detection systems are used, and there is a lack of information on the process for detection or the cause of restriction.

c. Substantive harm: risks according to the harm that too rapid decisions could produce, for example:
   i. On fundamental rights, such as freedom of expression or privacy.
   ii. On users' interests in the correct functioning of the platform, such as collaboration and discussion.
   iii. On larger public interests, such as democracy and public debate.

d. Different types of content (e.g. child abuse material vs. political content) have different risks of harm from false positives or false negatives in restriction decisions.

e. Downstream consequences of wrong decisions must also be taken into account as possible harms. A wrong decision of restriction can impact the user that generated the content but can also impact the audiences, the information environments and political decisions. The pressure to act quickly can disproportionately impact particular user groups based on language or type of content.

f. The need for quick decisions, even when justified, may have other negative effects with regards to people reviewing content, their preparation for decisions on restrictions, and their protection from harm produced by exposure to harmful content.

4. Criteria affecting how quickly a decision can be made include:

a. Whether it is a clear-cut case or a hard/disputed issue.

b. Whether restriction is pursuant to international human rights law, domestic local law, or terms of service and community guidelines, and whether there are some potential conflicts between these norms.

c. What the type of content is, both in terms of format (text, picture, video) and subject matter.

d. What the type and amount of harm would be in case of restriction or not, related to the impact the delay may have.

e. What the different effects may be in relation to users at large and those allegedly affected by certain content.

f. What action will be implemented, among the diversity of measures that could be used to restrict access to material. For instance, content removal restricts access to all users, while placing content behind a login system restricts it from users who have not registered for a particular website. These measures and others have varying impacts on accuracy and effects on user rights and defining the least restrictive one in difficult cases takes additional time.

I&JPN REF. 19-120

MULTI-FACTOR EVALUATION & IMPACT ANALYSIS

The information available in notices needs to be sufficient for decision-makers to understand inter alia what prohibition is being referred to, what specific content is allegedly violating it and whether the content does violate the prohibition. When the assessment is made that the content violates the prohibition, the action implemented needs to respect the standard of proportionality. Ensuring proportionate action on individual items of content requires evaluation of a diversity of factors and a broader appreciation of the potential impact of the measure.

1. Multi-factor evaluation

a. What is the context of the content at issue?

Content posted online is, by default, globally available. Nevertheless, the user making the content available and those accessing it perceive it within specific contexts (history, references, orientation, linguistic community, etc.).

In order to take this fundamental tension into account, decision-makers can identify where and from whom the specific piece of content originates. A larger discussion on the methods for and difficulties in identifying origination would be useful to fully understand the challenges behind the identification of context. This issue is compounded when taking into account situations where the content itself is "mirrored" across multiple different websites/platforms.

In addition, to fully understand context, decision-makers can first try to determine where the content is hosted and displayed. In other words, it is crucial to unpack the potential differences between where the website's domain is registered, the website/platform owner's country of incorporation, where the content is hosted/hashed, and where it is available.

Finally, evaluation of the context needs to be conducted by people with the capacity to understand the language and corresponding cultural environment.

b. What are the motives of those who have posted/re-posted this content?

The motive of the users who have posted or re-posted the content is important to consider. Decision-makers should keep in mind that there can be more than one motive, including the following: economic, political, humor, satire, social commentary. Those motives need to be evaluated within their linguistic and cultural environment. Where motive can be (or has been) ascertained, this may help decision-makers as they think through the various options available for restriction.

c. What motives might other actors have in "receiving"/having access to this content?

Decision-makers can consider what motives other actors can have in "receiving" or having access to this content. These can align, or be independent from, the intent of the user(s) posting it. It is also important to address the risk associated with the content, including the imminence of danger associated with it.

d. Are there particular jurisdictions/actors that may have an interest in and/or be impacted by this decision, and if so what do their laws/rules say about this kind of content?

Other jurisdictions/actors may have an interest in the decision, or be impacted by it. If the decision-makers identify that this is a possibility, it is important to consider what these interests may be, and in particular which of their specific laws or rules could apply. The above points pertaining to context and motives of users posting and receiving content may inform the identification of other relevant jurisdictions whose interests can be considered in a potential comity or conflict-of-law analysis.

e. **Are prohibitions of this kind of content universal/widely shared/inconsistent across jurisdictions?**
The decision-maker can determine the level of international normative coherence, understood as a basic assessment of the degree of global consensus on the unacceptability/illegality of such content. As a general matter, it may be useful to consider whether the content fits into one of the four categories described above in Normative Basis.

f. **What is the format of the content at issue? How does the format of the content at issue impact its potential virality?**
The format of the content at hand (e.g. text, image, video, hyperlink, etc.) can often determine the virality of the content and is an important criterion to analyze with respect to the ability of the content to be shared across pages, platforms and devices. In addition, the format of the content is an important (but not unique) characteristic in determining the file size and its implications on accessibility and storage.

2. Impact analysis

Evaluation includes the consideration of a range of potential impact, including (but not limited to):

a. **Impacts on freedom of expression**
Laws that restrict freedom of expression must meet the legality, legitimacy, and necessity tests drawn from Articles 19 of the Universal Declaration of Human Rights (UDHR) and the International Covenant on Civil and Political Rights (ICCPR). Laws on expression nevertheless vary across jurisdictions, and as a result decisions about how to restrict digital content must pay particular attention to potential conflicts of laws.

In particular, it is important for decision-makers to discern, where possible, what other jurisdictions may have connections to the content at issue. This can include for instance the identification of:
i. where the user(s) responsible for the content is situated,
ii. where the platform hosting the content is headquartered, and/or
iii. where significant audiences for the content are located.
The stronger the identifiable connections to countries whose laws could be read to protect the content at issue, the more cautious decision-makers should be before ordering restrictions that could have impacts in those jurisdictions.

b. **Impacts on privacy**
Enforcement of laws restricting expression online is imperfect and never complete. The extent to which authorities seek to limit this imperfection tends to correspond to the degree to which it frustrates an authority's legitimate interests and/or results in harm to other individuals in its jurisdiction.

Allowing content that has been determined to violate one country's laws to remain available elsewhere online opens up the possibility that individuals in the censoring country may continue to access it by circumventing technical restrictions. It is however important to recognize that most internet users do not use circumvention tools, and those that do tend to use them episodically. As a result, the "harm" that may flow from the possibility of circumvention should be scrutinized carefully and on a case-by-case basis.

Efforts to eliminate digital content, including efforts to prevent "re-posting", tend to have broad extraterritorial impacts that extend beyond freedom of expression. In particular, efforts to proactively identify possibly infringing content often create conditions that can lead to privacy infringements. Decision makers ordering content restrictions should be aware that their orders could impact the privacy and data protection rights of individuals both inside and outside their jurisdiction, and they should take steps to ensure such orders avoid or minimize infringing these rights.

c. Economic impacts

The hosting, display, and transmission of digital content can implicate a wide range of private enterprises, including web hosts, internet registries, internet service providers, mobile network operators, content-delivery networks, social media platforms, and financial intermediaries. Orders to restrict content, depending on their formulation, often impact multiple entities directly or indirectly.

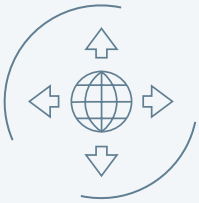
Decision makers ordering content restrictions should consider the extent to which private actors on the receiving end of such orders have the technological and economic means to implement them. Given the importance of fostering competition and innovation in the ICT sector, particular attention should be paid to the impacts such restrictions may have on smaller or start-up actors.

d. Setting precedent

Decisions to restrict a particular piece of content are rarely made in isolation. Such decisions tend to build on previous actions taken with respect to similar content, and also to impact future decisions. Decision makers should be cognizant of the possibility that any given restriction could be cited as precedent for future decisions by other decision makers in other contexts.

To the extent restriction decisions cumulatively reveal patterns, they can also impact the decisions of individuals whether or not to post content. While this can constitute effective "deterrence" against future violations, where restriction patterns are vague and protections for expression are unclear it can also lead to the "chilling" of legitimate expression.

The extent to which decisions can be contextualized and narrowly applied will help mitigate against misreading or misapplication by individuals or other decision makers.



ACTION

Once a piece of content has been determined to be abusive, harmful or illegal, a determination must be made on what specific action is the most proportionate in restricting or limiting access to it. A non-exhaustive list of the different types of actions at the disposal of service providers can be found in the Tools section. A key consideration in the choice of action must be a determination of the appropriate geographic scope of content restrictions in order to preserve the broadest availability of legitimate content.

I&JPN REF. 20-102

GEOGRAPHIC SCOPE OF CONTENT RESTRICTIONS

The principle of **proportionality,** along with other principles including necessity and legality, is well established in jurisprudence, especially in cases related to restrictions on speech and expression. However, the **global accessibility** of content posted online by users makes it subject to a plurality of national laws with potentially different or even conflicting regulatory obligations.

This calls for an additional criteria regarding the **geographic scope or reach** of content restrictions to become an integral part of the proportionality test in determining an appropriate course of action. As public authorities and private actors increasingly have to define the territorial scope of restrictions, the following **operational norm**[4] of 'geographically proportionate and relevant action' and two corresponding criteria can inform their decision making.

Geographically Proportionate and Relevant Action[5]

Decisions by public authorities and private actors preserve the broadest availability of legitimate content.

The **two following criteria** of 'International normative coherence' and 'Default action by service providers', further detailed on the next page, can provide a conceptual framework to operationalize this norm on a case-by-case basis and help determine the necessary and proportionate geographic scope of restrictions:

International normative coherence which categorizes the degree of convergence among legislations across jurisdictions.

Default action by service providers expected when a piece of content is deemed illegal or contrary to community guidelines.

This approach is intended to help guide a **diversity of stakeholders**, be it for judges in the treatment of cases submitted to them, policymakers in the development of corresponding legislation, content moderators in implementing either legislations or practices of particular platforms, and platforms themselves when developing their terms of service and community guidelines.

4. The notions of "operational norm" and "criteria" and these formulations come from the work of the multistakeholder Contact Group on Content & Jurisdiction that worked in 2018-19 in the perspective of the 3rd Global Conference of the Internet & Jurisdiction Policy Network in Berlin on 3-5 June 2019.

5. The original formulation of "Geographically proportionate action and international normative consistency" was rephrased by Stakeholders as part of the Berlin Roadmap. See Content & Jurisdiction Operational Approaches, p. 17. (https://www.internetjurisdiction.net/uploads/pdfs/Papers/Content-Jurisdiction-Program-Operational-Approaches.pdf)

International Normative Coherence

The following categories can be used to ascertain the degree of normative convergence on illegal content across jurisdictions. The frontiers between these different categories are however not rigid. Debates exist regarding where some topics fall.

1. There **IS** universal agreement that the content/behavior is illegal **AND** there is strong substantive convergence around the world on the corresponding threshold criteria. Example: child sexual abuse material.

2. There **IS** universal agreement that the content/behavior is illegal, **BUT** significant national variations exist in the criteria determining illegality. Example: defamation.

3. The content/behavior **IS NOT** universally considered as illegal, **BUT** the application of specific domestic laws on the local territory is considered acceptable by other countries, in particular for historic reasons. Example: criminalization of Holocaust denial.

4. The content/behavior **IS NOT** universally considered as objectionable **AND** some countries even consider that it should not be allowed to make it illegal. Example: laws discriminating against or criminalizing certain sexual orientation.

Note: The maps above illustrating normative coherence (particularly in 2 above) are not intended to be geographically accurate and is a simplified representation.

Default Action by Service Providers

The normative basis invoked for a content restriction has a direct relation with its geographic extent, as illustrated by the table below, which can help identify the default action associated with each case:

	GEOGRAPHICALLY LIMITED RESTRICTION	GLOBAL RESTRICTION
Illegal according to local laws	Unless the rationale for the request is clearly contrary to international human rights standards, by default, the content item is restricted locally by the provider (for instance through geo-IP filtering).	A global restriction can exceptionally be implemented by the provider in response to a request/order if a multi-factor evaluation meets a sufficient threshold including high international normative coherence.
Contrary to ToS/ Community Guidelines	Taking into account the context, including local circumstances, content is restricted in the most geographically proportionate manner.	The content is generally globally restricted when clearly in violation of the ToS/ Community Guidelines, except if a court issued a local stay-up order.

RECOURSE

Recourse is an essential part of due process in content moderation. It is independent from the evaluation conducted ahead of the action being taken on a piece of content and must provide avenues for users to challenge such action and obtain redress. This framework section outlines some possible recourse mechanisms that are structurally separated from the internal decision-making and even reconsideration processes of a service provider, and identifies some structuring questions regarding their interoperability.

MECHANISMS FOR RECOURSE

Every day, hundreds of millions of posts and hundreds of thousands of hours of videos are uploaded on the major internet platforms and made globally accessible, greatly facilitating freedom of expression. At the same time, legitimate concerns are raised regarding increasing harmful behaviors. Addressing abuses while protecting human rights has become a central issue of the global digital society.

Service providers have an important role to play in the identification and moderation of content that is illegal or not in compliance with their terms of service (ToS) and community guidelines. This has been translated into various normative frameworks, including self-regulation, codes of conduct or hard regulation. In addition, the numerous decisions on content restriction taken by providers are expected to be made in short timeframes to limit potential harm.

The use of automated tools increasingly allows detection at scale of potentially infringing content, but entails risks of bias and false positives or negatives. The increased reliance on ToS / community guidelines as the basis for content restriction decisions has in parallel magnified the norm-setting and decision-making roles of providers.

In order to ensure that content moderation and restrictions are proportionate and conducted responsibly, renewed attention is being paid to recourse mechanisms allowing users to contest a decision to restrict their content. New approaches at various degrees of development have emerged in recent years, including:

- **Company-established independent review**[iv]
  Some companies explore mechanisms to provide an independent appeal of their content restriction decisions made on the basis of their community guidelines. It is understood as a company-specific instrument with binding authority at the third level of a decision-making escalation path following initial first instance decisions and reconsideration.

- **Country-based self-regulation councils**[v]
  The establishment of independent self-regulatory bodies (Social Media Councils) at the national level is proposed to provide inter alia review mechanisms against content moderation decisions by providers.

- **Review by national authorities**
  Some actors have proposed that specific public authorities at the national level may have a formal role in reviewing content restriction decisions made by providers. The opinion of such bodies would be binding on the company, and geographically limited to the country.

- **Global advisory council**
  Finally, proposals for a global council with advisory power on companies' terms of service (ToS) and community guidelines have emerged, to increase transparency and accountability regarding this important normative basis.

---

iv. www.internetjurisdiction.net/content/companyindependentreview
v. www.internetjurisdiction.net/content/selfregulationcouncils
Note: The list above is non-exhaustive, and does not address or prejudge the degree of support for any of those proposals.

I&JPN REF: 19-136

RECOURSE MECHANISMS INTEROPERABILITY

The recent multiplication of initiatives and approaches to recourse mechanisms illustrates that actors have identified this issue as important and express the desire to address it. On the other hand, this proliferation raises major questions of interoperability, including:

1. **Jurisprudence coherence:** How can instances in which a decision made within one recourse mechanism contradicts the conclusions of another one be addressed? Should cases decided in such a manner have an impact on providers' ToS / community guidelines?
2. **Overlap:** How can duplication of efforts be avoided? In particular, if various separate mechanisms consider the same content restriction decision, how can coordination be best fostered?
3. **Liability:** How would potentially competing or complementing decisions by various recourse mechanisms impact providers' liability? What consequences for providers' liability do conflicting decisions infer?
4. **Relation with national courts:** How can multiple recourse mechanisms interact with national courts? In particular, could decisions by independent review mechanisms be appealed before national courts?
5. **Respective responsibilities of actors:** What roles can each type of actor play in the various recourse mechanisms, to ensure that users' rights are respected, that processes remain efficient, and that excessive burdens are not created?

Every recourse mechanism that is implemented will partly address these issues. Yet, unless frameworks for coordination and cooperation between actors are established, there are significant risks that uncoordinated actions lead to unintended consequences, including a lesser protection of users' rights, duplication of efforts and high costs. Jointly developed norms and criteria can help structure the interactions between various mechanisms and ensure that interoperability is included by default in the implemented approaches.

3.2

TOOLS

This section of the Toolkit provides practioners with a non-exhaustive list of tools such as typologies, criteria and request formats that can aid interactions between different stakeholders, including states and service providers within the context of identification, evaluation and choice of action on abuse.

TYPOLOGY OF CONTENT

IDENTIFICATION AND NOTICE SOURCES

TYPOLOGY OF DETECTION MODALITIES

COMPONENTS OF COMPLAINTS / REPORTS / NOTICES FROM PUBLIC AUTHORITIES AND PRIVATE NOTIFIERS

CHOICE OF ACTION

USER NOTIFICATION

I&JPN REF: 19-115

TYPOLOGY OF CONTENT

A broad diversity of types of content can potentially be illegal, harmful or against service providers' terms of service. In order to ensure coherence and interoperability between the different actors involved in content restrictions or moderation, sufficiently common terminology and interpretations are highly desirable.

The typology in the link below is **not a normative index of content that should be restricted**. It is a non-exhaustive effort by the Content & Jurisdiction Contact Group to describe the main issues at stake. It is structured in categories and includes a description of content that may be problematic or offensive in different situations and contexts.

A classification shared by the diverse actors is key for ensuring that appropriate steps are taken in the content moderation process. This document should help all actors to develop coherent, diversified and nuanced approaches for each type of content.

Some of the categories such as child abuse material, content to organize violence or support violent organizations, medical misinformation or abetting self-harm or suicide may fall within legitimate restrictions to freedom of expression as set out in the Article 19 of the International Covenant on Civil and Political Rights (ICCPR). Other categories, relating to privacy, racial and other forms of discrimination may require reconciling or balancing of different human rights. The majority of the content categories relate to forms of expression that in a certain context may be harmful or problematic but not necessarily illegal.

Coherent labelling of content is important not only for consistency during the four stages of the moderation process, as they should not be conducted by the same individual or entity, but also to help all actors involved have a common understanding of what is at stake. The most obvious examples of incoherence are the changing labels for certain categories. For example child sexual exploitation images or (CSAM) used to be called child porn, so called revenge porn is now known as non-consensual sexual images or (NCSI). Although some terms are commonly used, each term has its own context and interpretation, thus complicating the moderation process.

The complete table can be accessed online[6].

6. The full outcome can be accessed here: www.internetjurisdiction.net/content/outcomes/typology-of-content

I&JPN REF: 20-117

IDENTIFICATION AND NOTICE SOURCES

Identification of content potentially illegal or violating company's terms of service/community guidelines comes from a diversity of sources. This document intends to identify the different categories of such sources without assigning any value or hierarchy. The terms used are as neutral as possible and aim to include the terminology currently used in practice.

The notifiers covered in this typology could be represented in multiple categories and the typology does not establish the relevance of the notification provided, which should be evaluated on its own merits.

INDIVIDUALS	
Individual Notifiers	Any individual either directly targeted or flagging 3rd party content in their personal capacity.
Coordinated Individual Notifiers	Coordinated reporting/flagging by groups of individuals.

ORGANIZATIONS OTHER THAN PUBLIC AUTHORITIES	
Civil Society Organizations and Academia	Various definitions may apply, for example: According to the UN, "Civil society is the "third sector" of society, along with government and business. It comprises civil society organizations and non-governmental organizations". The World Bank refers to the "...wide array of non-governmental and not for profit organizations that have a presence in public life, express the interests and values of their members and others, based on ethical, cultural, political, scientific, religious or philanthropic considerations."
Press and Media	Private, public or community-based organizations having the scope to provide information to citizens with editorial responsibilities and subject to specific regulations.
Private Sector and Commercial Interest Groups	Industry associations, lobbying groups, organizations founded and/or funded by businesses that operate in a specific industry or hired/outsourced PR, marketing companies.
Political Parties	Political parties[7] are associations that participate in the management of public affairs, including the presentation of candidates for elections.

7. Similar definition by the Venice Commission and OSCE ODIHR accessible at: https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL-AD(2010)024-e

TRUSTED NOTIFIERS / FLAGGERS / PRIORITY FLAGGERS / REPORTERS / REPORTING AGENCIES	
Trusted notifiers / flaggers / priority flaggers / reporters / reporting agencies[8]	Individuals or organisations that are given a special status or a special reporting channel by platforms, which might not be available to regular users. This does not confer on them any specific legal status and these entities usually notify a platform based on terms of service or community standards infringements. This term has widely variable usage, most often referring only to partnerships between companies and their users, and sometimes being invoked in notification processes that involve governments.

PLATFORMS	
Platform Staff	Staff members whose job description includes content moderation/ management.
Sub-contractors	Companies or individuals hired to moderate content on behalf of internet platforms.
Non-compensated / Volunteer moderator(s)	These actors may be granted administrator-type privileges over certain fora and can form part of the escalation chain in reporting content that violates ToS or the law. These moderators occupy a space in between individuals, trusted flaggers, and sub-contractors.
Proprietary AI Tools	Machine learning software for content moderation.
Mutualized Hash Database	A shared database of "hashes", i.e unique digital "fingerprints" of content, that is set up between concerned actors to prevent further uploads of previously removed content.

PUBLIC AUTHORITIES[9]	
Government institutions (Ministries of Communication, Digital, Information...)	Executive organs of the State.
Regulators and other administrative Bodies	Regulators or agencies with a specific mandate to regulate online content.
Internet Referral Units	Specialist units most commonly established by police forces to liaise directly with internet platforms and service providers to alert them to potentially illegal content that contravenes the companies' ToS.
Law enforcement	Agencies mandated to enforce the law. Most commonly different police services.
National Courts	Legally binding court decisions issued by national courts.

INTERNATIONAL	
Regional & International Courts	Legally binding court decisions.
International Governmental Arrangements	International and Intergovernmental Organisations and other networks of governments (such as Christchurch call or Global Media Freedom Initiative).

8. Trusted Notifiers are transversal across this typology and can fall in multiple categories.
9. These categories often exist at the national/federal levels, as well as state/province and city/municipal levels.

I&JPN REF: 19-118

TYPOLOGY OF DETECTION MODALITIES

The following Typology of Detection Modalities is a list of recognized actions available to service providers and network / hosting intermediaries to identify allegedly harmful or illegal content. This list was developed to illustrate and map a spectrum of possible responses and is not meant to endorse any specific actions.

CONTENT PROVIDER / PLATFORM AND SEARCH ENGINES	
Actions	Description / Technical Tools
Account Authentication / Verification	Authentication is to ensure that the person is who s/he claims to be and to verify the identity data. It may include confirming email address, date account was established, whether the profile is complete, etc. Authentication relates more to an internal process where the verification is about external data.
Content Monitoring	Content monitoring involves the process of implementing procedures and filters to identify content or online behaviors which may be violative of ToS, community guidelines or local laws. Some monitoring is conducted by humans but much of it is done automatically through algorithms or AI. This may include evaluating behavior of users (e.g. who they follow or what they share), how other accounts interact with them (e.g. who mutes, follows, shares, or blocks the user), or if there are coordinated actions taken by groups or across platforms with the intent to harm. Once potentially harmful content/behavior is identified, it may be flagged for review.
Hashing (and hash databases)	This technology creates a unique digital signature (or "hash") of an image or video, which can then be compared against hashes of other photos or videos. This can help detect and remove or prevent the upload of a new image or video if its hash matches the hash stored in a database of items previously identified as justifying restrictions. Hash databases have been used for instance regarding child sexual abuse material, violent extremism, unauthorized dissemination of intimate images ("revenge porn") or copyright.
Notice and Take Down: Temporary or Permanent Removal	Mechanism where an individual can issue a legal request to a content host that requires the host to take down, delete or restrict access to allegedly harmful content. Examples include "Right to be Forgotten" policies in the EU and the copyright-oriented notice and takedown regime of the United States Digital Millennium Copyright Act.

NETWORK / HOSTING INTERMEDIARIES	
Actions	Description / Technical Tools
Notice and Take Down: Temporary or Permanent Removal	Hosting intermediaries may take content offline, based on company policies and procedures, court orders, or state regulations

I&JPN REF: 20-106

COMPONENTS OF COMPLAINTS / REPORTS / NOTICES FROM PUBLIC AUTHORITIES AND PRIVATE NOTIFIERS

A variety of notifiers, representing different stakeholder groups, identify potentially problematic content and notify platforms. This document intends to identify the minimum basic components that are needed for such notices.

REFERENCING				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Request Number	Request ID number generated by the notifier that identifies the specific demand; used for reference tracking and potential audits.	R	M	M
Time and date	Time and date when the notice was issued or generated.	M	M	M
Country	Indicates the country of origin of the request /demand.	R	R	M
Case Number	Identifies the corresponding legal case in the requesting country, if applicable.	M	M	M
Type of Notifier	Reference to Typology of Notifiers	R	M	M

TARGET				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Account information	Identifies the specific target of the request: user identifiers or accounts (criteria of specificity).	M	M	M
File Type	The type of allegedly infringing content (text, picture, video).	R	R	R
Content Language[10]	The language of expression of the content.	R	R	R
URL	URL to the piece of content (a timestamp of the alleged infringement in the case of multimedia content is recommended).	M	M	M

10. This type of information might be useful for platforms to select the appropriate AI tools or moderators to optimize time for human review.

TIMING				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Deadline	Identifies specific deadlines attached to the demand, if any.	N/A	N/A	M
Emergency	Identifies whether the circumstances correspond to a demonstrable situation of emergency.	Yes/No	Yes/No	Yes/No
Rationale for Emergency	Justification and demonstration of the emergency (e.g. its nature, link of the request to the emergency, how the action can avert the emergency).	M* if emergency is indicated.	M* if emergency is indicated.	M* if emergency is indicated.

CONFIDENTIALITY				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Confidentiality	Specifies whether specific circumstances justify that some parts or all of the demand not be communicated to the concerned user.	Yes/No	Yes/No	Yes/No
Rationale for Confidentiality	Justification of non-notification.	M* if confidentiality is indicated.	M* if confidentiality is indicated.	M* if confidentiality is indicated.
Confidentiality timeline	Duration of the confidentiality exception.	R	M* if confidentiality is indicated.	M* if confidentiality is indicated.

ANONIMITY				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Confidentiality	Specifies whether specific circumstances justify that some parts or all of the demand not be communicated to the concerned user.	Yes/No	Yes/No	Yes/No

SIGNATURE				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Signature	Identifies the signature and/or stamp of the notifier.	N/A	M	M

CONTACTS				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Issuing Authority	Contact details to which response notifications should be directed to.	R	M	M

CASE				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Category of violation	What is the category of alleged abuse[11]	M	M	M
Problem reported	Description of the issue and an explanation of the motivation to report.	R	M	M
Supporting Elements	Elaboration on the context, facts and potential harm.	R	R	M
Normative Basis	Reference to national legal framework or terms of service clause upon which this demand is based, ideally with an explicit link to an online version in English of the corresponding law/ jurisprudence if available.	R	R	M
Evaluation by notifier	An explanation of the prior evaluation conducted by the notifier	R	M	M

REQUESTED ACTION				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Action sought	Indication of the specific action requested.	N/A	N/A	M* if based on legality

ISSUING AUTHORITY				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Issuing Authority	The authority and/or POC that has issued the demand and its details.	N/A	N/A	M

SELF-CERTIFICATION / DECLARATION / ATTESTATION				
	Brief Explanation	**Mandatory/ Recommended for Individual Flaggers**	**Mandatory/ Recommended for Identified Flaggers**	**Mandatory/ Recommended for Public Authorities**
Self-Certification	Self-certification by the notifier (e.g.: I declare the information to be true and that there is no improper motivation or illegitimate purpose for this request.)	R	M	M

11. Refer to Operational Criteria A - Content Typology in the Operational Approaches document for examples

I&JPN REF: 19-122

CHOICE OF ACTION

The actions[12] that can be implemented to deal with content that is illegal, harmful or contrary to terms of service/community guidelines are increasingly diversified. The choice of the appropriate measure in each case is an important component to achieve the least restrictive effect.

The following Typology of Actions is a list of recognized actions available to platforms, intermediaries or states to block allegedly harmful or illegal content. This list was developed to illustrate and map a spectrum of possible responses and should not be understood as a normative index of actions that should be considered as equally valid.

CONTENT PROVIDER / PLATFORM AND SEARCH ENGINES	
ACTIONS	**DESCRIPTION/TECHNICAL TOOLS**
Additional context	Additional context may be required for posting certain types of content, and may include explanatory information or URLs to additional sources of information and alternative perspectives. For instance, graphic images of historical, artistic or scientific significance, might require context for the users to understand or appreciate the image. It may also be used in situations where content is deemed to be extremist or a form of disinformation.
Labelling	Label content with a warning for a specific type of content (i.e. violent content).
Age Verification / Age-gating	Age verification is undertaken by platforms to ensure that content is accessed only by users of the appropriate age. Age-gating prevents access to content and services by underage users according to national, regional or international laws.
Right of reply	Response for alleged defamatory content where the publisher/poster has the opportunity to post a reply, counter-speech, or disclaimer.
Account suspension	Accounts may be de-activated or suspended for a temporary period of time due to policy violations or invalid traffic. During this time, the account may not be accessible (i.e. error message will show), or visible to the public, or key functionality may be de-activated (ability to post, comment, read data). Alerts may be sent to give account holder time to address the issue. If the issues are not resolved, the account may be disabled.
Account disabled	Users or content providers who are not in compliance with the relevant governing policy, may have their account permanently disabled so that it is no longer visible or active. Some platforms may not allow the user to create a new account on the same platform.

12. "Among primary sources used to compile this list were:
Internet Society - Perspectives on Internet Content Blocking: An Overview
Daphne Keller - A Glossary of Internet Content Blocking Tools
Internet Society - Summary of Content Blocking Techniques
IETF - A Survey of Worldwide Censorship Techniques

Anonymizing source documents	In cases involving alleged defamatory information (i.e. "Right to be Forgotten"), names may be removed from source documents such as newspaper articles or public documents, and replaced with initials or a random letter (i.e. X or Y).
Block search indexing	In cases involving allegedly defamatory information, content from individual pages can be de-indexed/de-referenced so it cannot be found through internal (i.e. news archive) or external search engines. This is done either by including a noindex meta tag in the page's HTML code, or by returning a 'noindex' header in the HTTP request.
Block keywords	Content providers and search engines can block specific keyword search terms to prevent associated content from being found via search results. For example, on Tumblr, searches for keywords associated with adult content will come back with no results, even if there are matches.
Take down: temporary or permanent removal	Mechanism where an individual can issue a legal request to a content host that requires the host to take down, delete or restrict access to allegedly harmful content. Examples include "Right to be Forgotten" policies in the EU and the copyright-oriented notice and takedown regime of the United States Digital Millennium Copyright Act.
Down-ranking / voting (modifying the visibility of content)	Down ranking is used to demote content visibility (as Google web search has done on DMCA grounds) for content posted by confirmed bad-faith actors who intend to manipulate or divide the conversation.
Quarantining	Potentially harmful content may be quarantined to prevent it from being viewed by users. Quarantined content usually will display a warning for users who may not wish to view it, or require users' opt-in to view it.
Geo-blocking / Geo-IP-filtering / Withholding Content	Platforms can "withhold content" or block target content, or users and content at once. This can be done by blocking all users from a geographic region, from specific IP addresses, or other applications. An example of geographic blocking is "country withheld content" (CWC) which could happen, for instance, if a tweet violates local laws or if it is blocked due to a court order.
Shadow banning	Shadow banning restricts the visibility and reach of a user's content without their knowledge. This discreet ban allows the user to perform all the normal activities on a site but may prevent his/her profile or posted content from being visible to others or restrict the reach of the content by preventing it from appearing in feeds or showing in search results. This might be done to allow problematic or possibly harmful content to remain up while preventing those not seeking it from finding it. It may prevent bad actors from simply starting a new account if they knew of the ban, or alternately, it may encourage bad actors to leave a platform due to lack of engagement. It is also a common technique for combating bots and trolls. Other terms for this include, stealth banning, ghost banning or comment ghosting.
Platform-based blocking	In cooperation with platform, content or specified search results are blocked from coming back from the search engine. This is often initiated by national authorities to block "illegal" content within a geographic region and thereby avoid blocking an entire platform. In some cases, it may be done by platforms to block content that violates its terms of service or points to malware.

EXTRALEGAL BLOCKING	
ACTIONS	**DESCRIPTION/TECHNICAL TOOLS**
Blocking / Interferencen / RST Packet Injection	A specific type of packet injection attack that is used to interrupt an established stream by sending RST packets to both sides of a TCP connection; as each receiver thinks the other has dropped the connection, the session is terminated. This is also known as a "man in the middle" attack.

NETWORK / HOSTING INTERMEDIARIES	
ACTIONS	**DESCRIPTION/TECHNICAL TOOLS**
Deep packet inspection-based blocking	A device is inserted in the network that blocks based on keywords and/or other content (e.g. file name). This technique is often used for data protection, anti-spam and anti-malware (anti-virus), and traffic prioritization.
Keyword block lists	A keyword block list is a tool used by hosting intermediaries to filter keywords, and other forms of ID for video or audio. The filtering can be automated or done in combination with human monitoring. States also employ keyword blocking to censor content.
URL or HTTP Header Based Blocking	A device is inserted in the network that intercepts web requests and looks up URLs against a block list.
IP and protocol-based blocking	A device is inserted in the network that blocks traffic based on IP address and/or application (e.g. VPN) between the end user and the content.
Internet Service Providers (ISPs) [point of control]	ISPs are very effective points of control as they are easily identifiable and can readily identify the regional and international traffic of all users. Filtration mechanisms can be placed on an ISP via governmental mandates, ownership, or voluntary/coercive influence. ISPs can stop all its users from going to a website or using an app. Blocking can be done based on a URL, IP address (all content associated with IP or partial); technical specifications (such as blocking a port to prevent use of VOIP).
Geographic IP-filtering	A website can partially or fully block users with IP addresses from a certain country or based on GPS, Wi-Fi network identification, or other technical information.
Performance degradation	Performance degradation involves the intentional decrease in connectivity and response speed throughout a given network. "Bandwidth throttling," for instance, may be done to manage network congestion or to partially block a percentage of traffic from specified IP addresses or other applications.
Packet dropping	Packet dropping interrupts traffic flow by not properly forwarding packets associated with the harmful content. This technique is most effective when the packet contains transparent identifiers linked to the specified content, such as the destination IP. It often results in over blocking.
DNS-based blocking / Geographic TLD blocking	At the network or ISP level, Domain Name System (DNS) traffic is funneled to a modified DNS server that can block lookups of certain domain names.
DNS Interference	DNS interference results in an incorrect IP address being returned in response to a DNS query to a censored destination. Users may receive an error message.
Domain name reallocation / seizure	Domain names may be reallocated or seized legally (i.e. criminal copyright violations) or extrajudicially when a top-level domain (TLD) deregisters a domain name to prevent DNS servers from forwarding and caching the site.
Network disconnection or adversarial route announcement	This is a form of technical interference where a whole network can be cut off in a specified region when a censoring body withdraws all of the Border Gateway Protocol (BGP) prefixes routing through the censor's country. This is an extreme and extensive form of blocking usually only undertaken for short periods under dire circumstances.
Server Takedown	If undesirable content is hosted in the censoring country the servers can be physically seized or the hosting provider can be required to prevent access.

I&JPN REF: 20-116

USER NOTIFICATION

User notification is a critical part of moderation of any type of online content and an essential element of due process. It may allow users to provide additional information or modify their upload or posting before any restrictive measure[13] is decided or implemented. It is in any case a precondition to reconsideration or recourse processes.

Building on the work of the Content & Jurisdiction Contact Group in 2019, the present document intends to reconcile the practical constraints regarding the timing of user notification with strengthening due process.

1. I&J Operational Norm and Criteria

The Operational Approaches developed by the Content & Jurisdiction Contact Group in 2019 identified the importance of early user notification and laid out the following Operational Norm:

> ***Users are notified ahead of the enforcement of restriction decisions regarding their content. If justifiably demonstrable according to clear pre-agreed criteria that advance notification is not practical, advisable, or permissible, users are notified expeditiously after the enforcement of a restriction decision. Some situations may justify an exception to the general principle of user notification.***

Regarding the content of notification, Criteria I of the Operational Approaches further clarified that:

> ***The notification should contain information pertaining to the normative basis and rationale for restriction along with the specific/respective channels, information and applicable timelines for recourse. For content restricted on the basis of the providers' ToS/Community Guidelines, notification also contains information pertaining to the specific clause/guideline that was violated.***

However, the implementation of the above Operational Norm requires taking into account important constraints.

2. Implementation constraints

A considerable volume of potentially objectionable[14] posts must be detected, reviewed and decided upon expeditiously to ensure the timely prevention and remediation of online harm. However, public authorities have a limited human and technical capacity to ensure detection of illegality at scale.

---

13. See Criteria H: Choice of Action in Content & Jurisdiction Operational Approaches

14. Although contentious posts represent only a very small proportion (less than 1%) of overall activity, hundreds of millions per year (for the largest operators) have to be reviewed for illegality or violation of Terms of Service.

Moreover, applying to each case the elaborate court procedures developed for traditional publishing would create a considerable burden on the judicial systems and introduce delays incompatible with the rapid and potential global propagation of illegal content. In this context, national (or regional) regulations increasingly impose upon operators the responsibility to address illegal content, with short action timelines under penalties.

The same volume and time constraints apply to service providers, compounded by the need to also address violations of their own increasingly detailed rules. While artificial intelligence tools increasingly supplement flagging systems for the detection of content to review, human intervention is necessary for decision-making.

Notifying users before a content restriction is decided and implemented naturally implies delays. It is thus frequently not practical when time is of the essence to address harm or the illegality (resp. violation of companies rules) is sufficiently evident upon rapid review. Prior notification may not be permissible by law, and is inadvisable in case of ongoing or imminent real world harm.

Likewise, notification in the course of review is only meaningful if it allows the user to either modify the posting so that it does not infringe any more on the relevant normative basis, or provide relevant context or information useful for the evaluation.

An important distinction must therefore be made between:

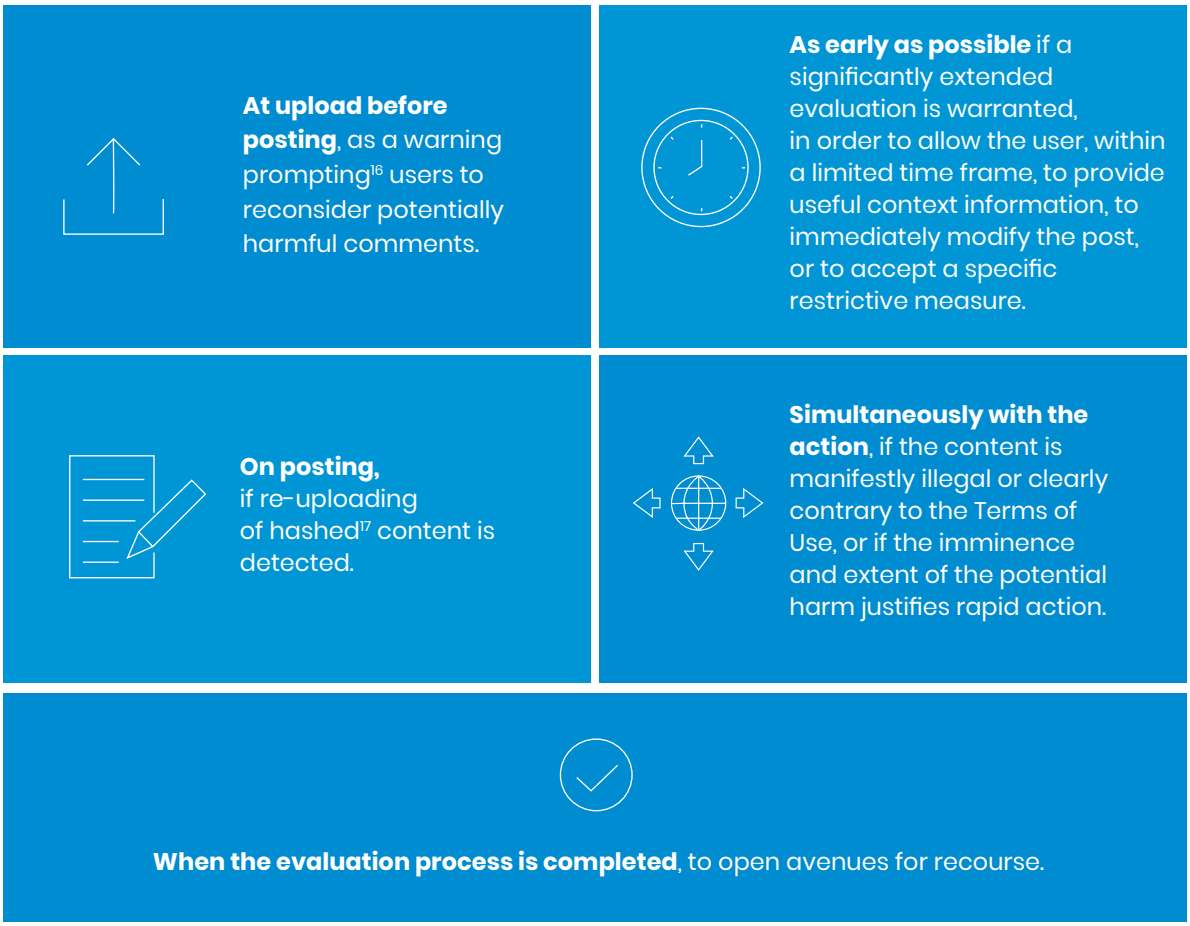
- Content that is **manifestly illegal or clearly contrary to the terms of use** of the intermediary, and
- Content that **requires a more extensive evaluation** in light of the context, to establish the right balance between prevention of potential harm and protection of freedom of expression.

Indeed, some recent legislations (e.g. the German NetzDG) do recognize different response times on operators according to how manifest the illegality is.[15] Likewise, internet companies' internal escalation paths allow more time to handle non-manifest violations of their rules. This allows for a time-bound extensive evaluation by the company.

15. Under Netz DG, companies must take down or block access to manifestly unlawful content within 24 hours of receiving a complaint. Other illegal content must be taken down or blocked within 7 days of receiving a complaint. Alternatively, social networks may refer the content concerned to a "recognised institution of regulated self-governance" (the FSM) with the understanding that they will accept the decision of that institution. The institution must then decide on whether the content is unlawful within 7 days. Social networks can exceed the 7 day deadline when determining the illegality of the content depends on "the falsity of a factual allegation" or "other factual circumstances".

3. Timing of user notification

In light of the above, user notification can be meaningfully implemented by the company at different stages of the evaluation, under the following conditions:



In the second situation, the notification should also indicate if the service provider is applying temporary measures to limit the distribution or virality of the content during the extended evaluation. The company can also separately solicit advice from a competent third body.

As notification is a precondition to reconsideration or recourse processes, in all cases it needs to contain, as indicated in Criteria I above, precise information regarding the available reconsideration and appeal mechanisms, and, in some jurisdictions, the possibility of recourse to a higher authority.

16. See for instance Twitter's "Want to revise this?" and Instagram's "do you really want to post this?"

17. Hash databases contain references to already evaluated content that has been previously determined as very dangerous or harmful (e.g. CSAM or terrorist content).

4. INTERNET & JURISDICTION POLICY NETWORK

> **Managing the way that a large number of separate legal frameworks apply to the internet is one of the biggest policy challenges of our time – more complex than building the internet itself.**
>
> **Vint Cerf Co-inventor of the internet, writing in the Financial Times ahead of the 2nd Global Conference of the Internet & Jurisdiction Policy Network in 2018**

The Internet & Jurisdiction Policy Network is the multistakeholder organization fostering legal interoperability in cyberspace. Its stakeholders work together to preserve the cross-border nature of the internet, protect human rights, fight abuses, and enable the global digital economy. Since 2012, the Internet & Jurisdiction Policy Network has engaged more than 400 key entities from six stakeholder groups around the world including: governments, the world's largest internet companies, the technical community, civil society groups, leading universities and international organizations.

The regular Global Conferences of the Internet & Jurisdiction Policy Network are institutionally supported by six international organizations: Council of Europe, European Commission, ICANN, OECD, United Nations ECLAC, and UNESCO. Host partner countries include France (2016), Canada (2018) and Germany (2019).

The Community

6
STAKEHOLDER GROUPS

70+
COUNTRIES

400+
ENTITIES

STATES

INTERNET COMPANIES

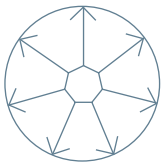
TECHNICAL OPERATORS

CIVIL SOCIETY

INTERNATIONAL ORGANIZATIONS

ACADEMIA

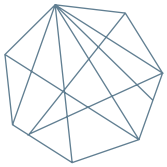
Mission



INFORM

The debates to enable evidence-based policy innovation

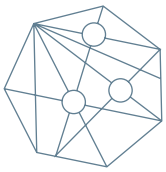
Informational asymmetry and mistrust between actors often result in uncoordinated policy action.
The I&JPN facilitates **pragmatic** and well-informed policy-making by framing issues and taking into account the **diversity of perspectives** while documenting tensions and efforts to address problems.



CONNECT

Stakeholders to build trust and coordination

Cooperation is important in a digital environment that is increasingly polarized, and where actors function in policy silos, with insufficient factual information.
The I&JPN serves as the **connective tissue** between stakeholder groups, regions, and policy sectors, as well as by **bridging gaps** within governments or organizations.



ADVANCE

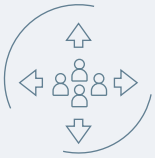
Solutions to move towards legal interoperability

The Policy Network strives to develop shared **cooperation frameworks** and **policy standards** that are as transnational as the internet itself.
The Network promotes a **balanced and scalable approach** to policymaking, aiming for legal interoperability, taking inspiration from the fundamental principle that enabled the success of the internet and the World Wide Web.

Core activities

POLICY PROGRAMS

EVENTS



KNOWLEDGE MUTUALIZATION

5. ACKNOWLEDGEMENTS

This Toolkit is based on the work of the Members of the Content & Jurisdiction Program Contact Group of the Internet & Jurisdiction Policy Network 2017-2020, and its Outcome Documents, as well as the Roadmaps that resulted from the Global Conferences of the Internet & Jurisdiction Policy Network in 2016 (France), 2018 (Canada) and 2019 (Germany).

The Secretariat is grateful for the hundreds of hours of intense work of the Members of the Content & Jurisdiction Program Contact Groups, and their alternates, composed of senior level representatives from governments, internet companies, technical operators, civil society, leading universities, and international organizations from around the world since 2017.[20] The Secretariat also expresses thanks to Wolfgang Schulz, Director of the Alexander von Humboldt Institute for Internet and Society who serves as Contact Group Coordinator (2017-present).

The following list of Members and their appointed alternates indicates the affiliation of stakeholders at the time they served in the Contact Group. Members served in their personal capacity.

Rasha Abdul Rahim, Researcher Technology and Human Rights, Amnesty International • **Guilherme Fitzgibbon Alves Pereira** Head of Science, Technology, Innovation, and Cooperation, Embassy of Brazil in Berlin, Brazil - Ministry of Foreign Affairs • **Ana Andrijevic**, Research and Teaching Assistant, University of Geneva • **Chinmayi Arun**, Assistant Professor of Law and Executive Director, Centre for Communication Governance, National Law University, Delhi • **Luca Belli**, Senior Researcher, Fundação Getulio Vargas Law School • **Susan Benesch**, Project Director, Dangerous Speech Project • **Guy Berger**, Director, Freedom of Expression and Media Development, UNESCO • **Beatrice Berton**, Policy Officer, Europol • **Theo Bertram**, Senior Manager, Public Policy EMEA, Google • **Oli Bird**, Head of International Policy, United Kingdom - Office of Communications (Ofcom) • **Ellen Blackler**, Vice President, Policy Strategy, The Walt Disney Company • **Nikki Bourassa**, Program & Policy Officer, Global Network Initiative • **Jillian C. York**, Director for International Freedom of Expression, Electronic Frontier Foundation • **Agnes Callamard**, Director, Columbia University, Global Freedom of Expression Project • **Greg Callus**, Ombudsman, Financial Times • **Maria Paz Canales**, Executive Director, Derechos Digitales • **Jorge Cancio**, Deputy Head of International Affairs, Switzerland - Federal Office of Communications • **Juan Carlos Lara**, Research & Public Policy Director Derechos Digitales • **Jordan Carter**, Chief Executive, InternetNZ • **Mark Carvell**, Head of International Online Policy, United Kingdom - Department for Culture Media and Sport • **Adeline Champagnat**, Advisor to the Prefect in Charge of the Fight Against Cyberthreats, France - Ministry of Interior • **Alexander Corbeil**, Research Advisor, Canada - Department of Public Safety • **Nicolas D'Arcy**, Head of Government Affairs, Dropbox • **Jennifer Daskal**, Associate Professor, American University Washington College of Law • **Giovanni De Gregorio**, PhD. Candidate, University of Milano-Bicocca • **Iris de Villars**, Head of Tech Desk, Reporters Without Borders • **Jacques de Werra**, Professor and Vice Rector, University of Geneva • **Agustina Del Campo**, Director, University of Palermo (CELE) • **Harlem Desir**, OSCE Representative on Freedom of the Media • **Elena Dodonova**, Administrator, Media and Internet Division, Council of Europe • **Maria Donde**, Head of International Content Policy, United Kingdom - Office of Communications (Ofcom) • **Kristine Dorrain**, Senior Corporate Counsel, Amazon Web Services • **Louis-Victor Douville De Franssu**, Advisor to the Ambassador for Digital Affairs, France - Ministry of Foreign Affairs • **Nilay Erdem**, Head of Content Policy Stakeholder Engagement, Facebook • **Anriette Esterhuysen**, Director of Policy and Strategy, Association for Progressive Communications • **Miriam Estrin**, Public Policy Manager, Google • **Raquel Gatto,** Regional Policy Advisor, Internet Society • **Jan Gerlach**, Lead Public Policy Manager, Wikimedia Foundation • **Alison Gillwald**, Executive Director, Research ICT Africa • **Tonei Glavinic**, Director of Operations, Dangerous Speech Project • **Caroline Greer**, Head of European Public Policy, Cloudflare • **Hiroki Habuka**, Deputy Director, Information Economy Policy Division, Japan - Ministry of Economy, Trade and Industry • **Gazala Haq**, Head of Public Policy and Government Affairs, EMEA, Dropbox • **Pablo Hinojosa**, Strategic Engagement Director, APNIC • **Daniel Holznagel**, Legal Officer, Germany - Federal Ministry of Justice and Consumer Protection • Xianhong Hu, Program Specialist, UNESCO • **Raman Jit Singh Chima**, Global Policy Director, Access Now • **David Kaye**, Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, United Nations •

20. An overview of the Members of the Content & Jurisdiction Program Group by year can be found here : https://www.internetjurisdiction.net/news/content-jurisdiction-program-contact-group-members

Daphne Keller, Director of Intermediary Liability, Stanford Law School, Center for Internet and Society • **Gail Kent**, Global Public Policy Manager, Facebook • **Gayatri Khandhadai**, Project Coordinator, Association for Progressive Communications • **Edison Lanza**, Special Rapporteur for Freedom of Expression, Organization of American States • **Judith Lichtenberg**, Executive Director, Global Network Initiative • **Emma Llanso**, Director, Free Expression Project, Center for Democracy & Technology • **Hawley M. Johnson**, Associate Director, Global Freedom of Expression, Columbia University • **Rebecca MacKinnon**, Director, Ranking Digital Rights, New America Foundation • **Katherine Maher**, Executive Director, Wikimedia Foundation • **Jeremy Malcolm**, Senior Global Policy Analyst, Electronic Frontier Foundation • **Nathalie Marechal**, Senior Policy Analyst, New America Foundation • **Giacomo Mazzone**, Head of Institutional Relations and Member Relations, European Broadcasting Union • **Corynne McSherry**, Legal Director, Electronic Frontier Foundation • **Ayman Mhanna**, Executive Director, Samir Kassir Foundation • **Drew Mitnick**, Policy Counsel, Access Now • **Gregory Mounier**, Head of Outreach at European Cybercrime Centre (EC3), Europol • **Paul Nemitz**, Principal Advisor, European Commission, DG JUST • **Thiago Tavares Nunes de Oliveira**, Board Member, CGI.br • **Juan Ortiz-Freuler**, Policy Fellow, World Wide Web Foundation • **Javier Pallero**, Policy Lead, Latin America, Access Now • **Elena Perotti**, Executive Director, Public Affairs and Media Policy, WAN-IFRA • **Nick Pickles**, Director, Public Policy Strategy, Twitter • **Jason Pielemeier**, Policy Director, Global Network Initiative • **Eliska Pirkova**, Europe Policy Analyst, Access Now • **Frederic Potier**, National Delegate, France - Inter-ministerial Delegation for the Fight against Racism, Antisemitism and Anti-LGBT Hatred • **Alice Rutherford**, Head of International Online Policy, Security and Online Harms Directorate, United Kingdom - Department for Digital, Culture, Media and Sport • **Alexander Schafter Lex**, Deputy Head of Division, Consumer Policy in the Information Society, Germany - Federal Ministry of Justice and Consumer Protection • **Thomas Schneider**, Vice-Director, Switzerland - Federal Office of Communications • **Nicolás Schubert** Gallardo, Head, Digital Economy Department, Undersecretariat of International Economic Affairs, Chile - Ministry of Foreign Affairs • **Bernard Shen**, Assistant General Counsel, Microsoft • **Sherwin Siy**, Senior Public Policy Manager, Wikimedia • **Carlos Affonso Souza**, Director, Institute for Technology and Society (ITS Rio) • **Alissa Starzak**, Head of Public Policy, Cloudflare • **Peter Stern**, Director, Content Policy Stakeholder Engagement, Facebook • **Ellen Strickland**, Chief Advisor, International, InternetNZ • **Everton Teles Rodrigues**, Expert Advisor, CGI.br • **Lee Tuthill**, Counsellor, Trade in Services, World Trade Organization • **Elfa Ýrgylfadottir**, Director, Media Commission, Iceland - Ministry of Communications • **Achilles Emilio Zaluar Neto**, Ambassador, Brazil -Ministry of Foreign Affairs

I&JPN SECRETARIAT

LEAD CONTENT & JURISDICTION PROGRAM:

Bertrand de la Chapelle, Executive Director
Frane Maroevic, Director, Content & Jurisdiction Program
Ajith Francis, Policy Programs Manager
Sophie Tomlinson, Communications and Outreach Manager
Juri Wiedemann, Young Professional

Paul Fehlinger, Deputy Executive Director
Martin Hullin, Director of Operations and Knowledge Partnerships
Hedvig Nahon, Events and Office Manager

FINANCIAL AND INSTITUTIONAL SUPPORTERS

This Toolkit would not exist without the support of the unique coalition of governments, international organizations, businesses, technical operators and foundations, which enable the work of the Internet & Jurisdiction Policy Network.

Please consult the overview of these key actors and their logos at
https://www.internetjurisdiction.net/about/funding

The Internet & Jurisdiction Policy Network is the multistakeholder organization fostering legal interoperability in cyberspace. Its stakeholders work together to preserve the cross-border nature of the internet, protect human rights, fight abuses, and enable the global digital economy. Since 2012, the Internet & Jurisdiction Policy Network has engaged more than 400 key entities from six stakeholder groups around the world including: governments, the world's largest internet companies, the technical community, civil society groups, leading universities and international organizations.

www.internetjurisdiction.net