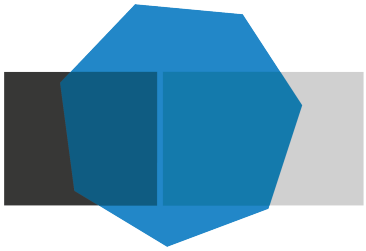


INTERNET
& JURISDICTION
POLICY NETWORK

CONTENT & JURISDICTION
POLICY OPTIONS

CROSS-BORDER CONTENT RESTRICTIONS

November 2017

Input Document for Workstream II of the second Global Internet and Jurisdiction Conference

GLOBAL INTERNET
AND JURISDICTION
CONFERENCE 2018
FEBRUARY 26-28 • OTTAWA, CANADA
conference.internetjurisdiction.net

@IJurisdiction
www.internetjurisdiction.net

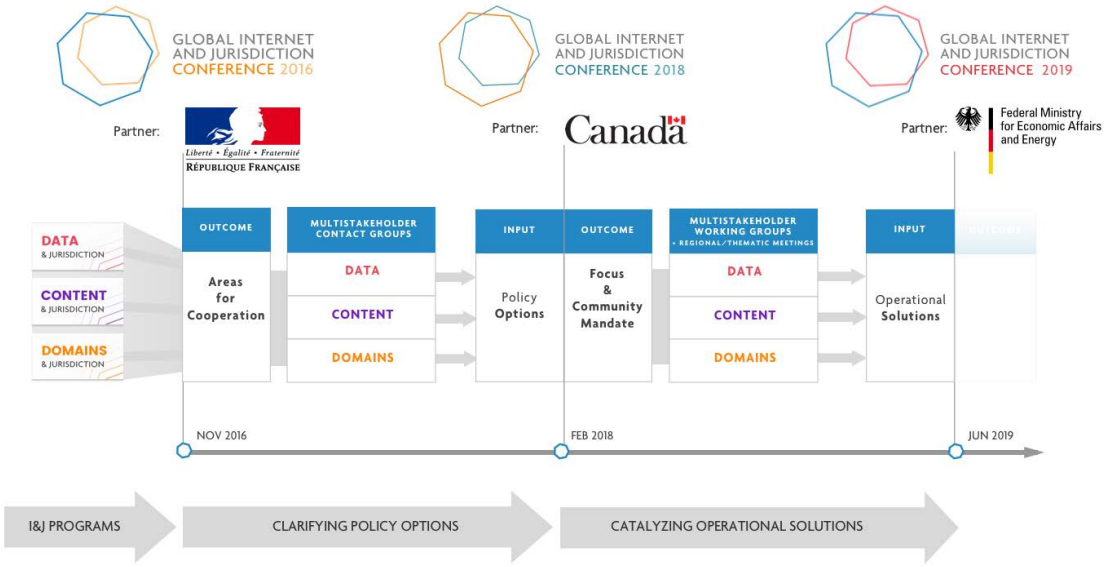
This Policy Options document prepared by the Secretariat of the Internet & Jurisdiction Policy Network presents the results of the work of the multistakeholder Content & Jurisdiction Contact Group. This Group was set up as a result of the first Global Internet and Jurisdiction Conference of the Policy Network, held in Paris on November 14-16, 2016, and which gathered 200 senior-level participants from 40 countries. The Group held seven virtual meetings in 2017 to explore the Areas of Cooperation identified in Paris (see Content & Jurisdiction Framing Paper[1]).

Reflecting preparatory work, this document will serve as input to structure discussions in Workstream II on Content & Jurisdiction on Day 2 of the second Global Internet and Jurisdiction Conference[2] in Ottawa on February 26-28, 2018. On this basis, stakeholders are expected to agree there on focus and community mandates to structure further work in the Internet & Jurisdiction Policy Network.

Feedback on this document can be submitted to the Secretariat until January 26, 2018 via gijc2018-content@internetjurisdiction.net. It will be shared with the Members of the Contact Group.

AN ONGOING PROCESS TOWARDS OPERATIONAL SOLUTIONS

The second Global Internet and Jurisdiction Conference of the Internet & Jurisdiction Policy Network is organized in partnership with the Government of Canada, and institutionally supported by OECD, UNESCO, Council of Europe, European Commission, and ICANN. It will be a milestone moment to identify concrete focus and priorities to develop policy standards and operational solutions to major jurisdictional challenges. This will define the methodology and roadmap in the lead-up to the third Global Internet and Jurisdiction Conference in June 2019, which will be organized in partnership with the Government of Germany.



ABOUT THE CONTENT & JURISDICTION CONTACT GROUP

The Contact Group was set up under the Content & Jurisdiction Program of the Internet & Jurisdiction Policy Network. Parallel Contact Groups have been established in the Data & Jurisdiction and Domains & Jurisdiction Programs, as well. The Group is composed of Members of different stakeholder constituencies, actively involved in addressing the jurisdictional challenges related to online content restrictions. This neutral space allowed participants to map their respective perspectives, compare approaches, foster policy coherence, and identify possible steps for coordinated actions.

---

[1] Secretariat of the Internet & Jurisdiction Policy Network (2017). Framing Paper of the Content & Jurisdiction Program https://www.internetjurisdiction.net/publications/paper/content-jurisdiction-program-paper

[2] https://conference.internetjurisdiction.net

Members of the Contact Group are:

- **Chinmayi Arun**
  Executive Director, Centre For Communication Governance
  National Law University, Delhi
- **Luca Belli**
  Senior Researcher
  Fundação Getulio Vargas (FGV) Law School
- **Theo Bertram**
  Google Policy Strategy, EMEA
  Google
- **Ellen Blackler**
  Vice President, Global Public Policy
  The Walt Disney Company
- **Adeline Champagnat**
  Advisor To The Prefect In Charge of The Fight Against Cyberthreats,
  Ministry of Interior, France
- **Raquel Gatto**
  Regional Policy Advisor
  ISOC
- **Elfa Ýr Gylfadóttir**
  Director, Media Commission
  Ministry of Communications, Iceland
- **Xianhong Hu**
  Assistant Programme Specialist, Communications and Information,
  UNESCO
- **Daphne Keller**
  Director of Intermediary Liability
  Stanford Law School, Center for Internet and Society
- **Gail Kent**
  Global Public Policy Manager
  Facebook
- **Edison Lanza**
  Special Rapporteur for Freedom Of Expression
  Organization of American States
- **Judith Lichtenberg**
  Executive Director
  Global Network Initiative
- **Rebecca Mackinnon**
  Director, Ranking Digital Right
  New America Foundation
- **Jeremy Malcolm**
  Senior Global Policy Analyst
  Electronic Frontier Foundation
- **Frane Maroevic**
  Director, Office of the OSCE Representative on Freedom of the Media,
  OSCE
- **Gregory Mounier**
  Head of Outreach at European Cybercrime Centre (EC3)
  Europol
- **Paul Nemitz**
  Principal Advisor
  DG JUST, European Commission
- **Nick Pickles**
  Head, UK Public Policy
  Twitter
- **Christian Meyer Seitz**
  Head of Division, Consumer Policy in the Information Society, Federal
  Ministry of Justice and Consumer Protection, Germany
- **Wolfgang Schultz**
  Professor
  Humboldt Institute for Internet And Society (HIIG)
- **Thiago Tavares**
  President
  Safernet, Brazil
- **Elena Lopatina**
  Programme Manager, Media and Internet Division
  Council of Europe

ABOUT THE INTERNET & JURISDICTION POLICY NETWORK

The Internet & Jurisdiction Policy Network addresses the tension between the cross-border nature of the internet and national jurisdictions. Its Paris-based Secretariat facilitates a global multistakeholder process to enable transnational cooperation. Participants in the Policy Network 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 200 key entities from governments, internet companies, technical operators, civil society, academia and international organizations around the world. Its Secretariat has convened, organized or contributed to more than 120 policy events in over 30 countries.

Content & Jurisdiction Policy Options

This document aims at providing, in a forward-looking approach, guiding elements to structure further discussion on possible frameworks regarding transnational content restrictions at the second Global Internet and Jurisdiction Conference in Ottawa on February 26-28, 2018, and beyond. It documents the key substantive and procedural dimensions that can help overcome the current divergences regarding the responsibilities of intermediaries.

Online services accepting user-generated content have become major economic actors and key instruments for the exercise of freedom of expression by billions of users around the world, thanks to frictionless posting and the global availability the internet provides. At the same time, this ease of use is abused, leading to a proliferation of harmful behaviors, including hate speech, incitement to violence, or harassment that need to be dealt with. The status quo appears unsustainable and questions are being raised regarding the degree of engagement of intermediaries in response.

Volume represents a major issue. Several hundreds of millions of posts and hundreds of thousands of hours of videos are uploaded every day on the major platforms. At the same time, numerous individual pieces of content are either reported by public authorities as illegal in their country or flagged by users as objectionable. These requests for content restrictions cover a large spectrum of grounds, reflect the disparity of national laws and illustrate the sometimes opposing values of a growing and increasingly diverse user base. In addition, public authorities sometimes flag content using company reporting mechanisms based on Terms of Service, blurring the line between legal requests and company policies.

On a principle level, any state restriction to freedom of expression needs, inter alia, to be authorized by clear and predictable law, to meet the criteria of necessity and proportionality, and to be decided with sufficient due process guarantees. This is however a complex challenge on a transnational basis in the absence of clearly agreed substantive and procedural frameworks.

More importantly, although still small in proportion to the overall volume of posts, the absolute number of individual restrictions decisions to be made is unprecedented[3]. Case-by-case determinations need to account for context and intent in a way analogous to national courts, but within very limited response times given viral propagation. This is a challenge for all actors, including intermediaries directly receiving those requests.

Online platforms are also subject to opposing demands: one asking them to thoroughly police the content posted on their services to guarantee the respect of national laws, and the other objecting to them making determinations on their own and exercising proactive content monitoring, for fear of detrimental human rights implications. Moreover, given that the current non-liability regimes were initially established for "passive" intermediaries, the fear of a potential loss of protection may disincentivize companies from assuming more responsibilities.

[3] For instance, one company recently reported deleting close to 10.000 posts flagged as hate speech per day in 2017 - see: Richard Allan, 'Hard Questions: Hate Speech' (27 June 2017, Facebook Newsroom) https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/ accessed 30 August 2017.

The above elements underline the acute complexity of these challenges. Clear common guidelines and mechanisms become increasingly needed to properly deal with abusive content, i.e.: to maximize the necessary remediation of harm and minimize restrictions to freedom of expression. Existing instruments based on strict territorial jurisdictions are challenged and some institutional innovation might be needed.

A common goal for the different actors could be to develop a framework that clarifies:

- Applicable substantive norms and the interplay between agreed international human rights, national laws, and companies' community guidelines,
- The respective responsibilities and protections of actors in the detection of abuses,
- Decision-making, including the escalation path for individual decisions and appeal mechanisms,
- Necessity and proportionality regarding the geographic scope of restrictions, and the handling of merely controversial content,
- The necessary due process standards that should be applied across borders.

These five interconnected dimensions are detailed below to focus and structure further discussions.

1. Substantive norms

Various sources contribute to a complex patchwork of relevant substantive norms in cyberspace, with different degrees of authority:

- The internationally agreed human rights principles and standards,
- The broad set of national laws and jurisprudences, very diverse in terms of the abusive content covered and the corresponding illegality criteria,
- The increasingly detailed public community guidelines and private implementation procedures discretionarily developed by platforms to organize their respective digital spaces.

Determining when and how these norms coexist, complement each other, overlap or potentially conflict is a first fundamental and delicate challenge.

International human rights principles and standards provide an overarching reference and the UN Guiding Principles on Business and Human Rights[4] (know as the "Protect, Respect, Remedy" Framework) clarified the respective duties of states and responsibilities of companies.

National laws have, a priori, a specific geographic scope. They legitimately govern the conditions of access to content in the corresponding territory when in conformity with human rights standards and rule of law in terms of, inter alia, purpose, adoption procedures, clarity, precision and mechanisms for redress.

---

[4] See: http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf, as well as a recent study by the NYU Stern Centre for Business and Human Rights at: https://issuu.com/nyusterncenterforbusinessandhumanri/docs/final.harmful_content._the_role_of_?e=31640827/54951655

Yet, the typology of abusive content remains imprecise (general lack of common definitions) and inconsistent across legislations given the level of agreement spectrum among states, with:

- Content prohibited in only certain countries, but not in most others, where it can even be protected,
- Content globally considered objectionable, but with very different limits, boundaries and criteria for illegality between countries;
- Content universally objected to, usually in violation of established international standards, with broad agreement on the related criteria.

Companies' community guidelines, by contrast, are often intentionally conceived for global application to simplify the management of large user communities, resulting in a form of de facto harmonization and global restrictions in case of violation. However, these rules need to respect national laws regarding contracts. They can legitimately vary significantly from one service to the other, according to their technical nature, size, business scope, or the specific user communities served.

Many stakeholders express concern about the definition by private companies alone of broad guidelines directly impacting fundamental rights, as well as regarding the confidentiality of the corresponding internal implementation instructions. Even large companies recognize the difficulty of properly developing and applying such substantive criteria, and smaller entities may not be able to do the same.

Irrespective of this complex landscape, recognition of the need to deal with abusive content has grown in the last few years, under the overarching principles of necessity and proportionality. All actors agree on restricting access to the most egregious types of content, such as child sexual abuse imagery. Increased attention is given to terrorist content in declarations and codes of conduct. Terms of Service reference a growing number of forbidden types of behaviors, and national laws increasingly target specific types of abuse online. Some convergence is also observed on the need to tackle certain topics that did not previously garner as much consensus or attention, such as various forms of online bullying or non-consensual pornography ("revenge porn").

Global substantive harmonization however is not achievable and the desirable degree of overall convergence remains an open question. The challenge is thus to develop a better understanding of the types of abuses the global community wants to see addressed, how different norms can coexist in shared online spaces, and how to handle their possible conflicts. Structured interactions among all relevant stakeholders on a topic-by-topic basis represent a possible approach, in order to take into account the specificities of the different types of objectionable content. Initiatives such as the recently established Global Internet Forum to Combat Terrorism[5] and previous cooperation initiatives[6] regarding child sexual abuse imagery also illustrate the possible use of shared repositories of infringing content.

Question: *Should distinct multistakeholder cooperation mechanisms be established for different vertical issues on substantive and procedural dimensions? Could this help platforms define their community guidelines and implementation criteria?*

---

[5] https://newsroom.fb.com/news/2017/06/global-internet-forum-to-counter-terrorism/

[6] https://www.iwf.org.uk/ - https://www.wired.co.uk/article/iwf-hash-lists-child-abuse-images

2. Detection of abuses

A second set of challenges is related to the mechanisms and tools for identifying potentially illegal or inappropriate content, and the degree of proactive efforts expected from companies.

Most current non-liability regimes protect intermediaries until they receive notice of allegedly infringing material, provided they act expeditiously afterwards. Court decisions may or may not have been obtained beforehand. In that context, platforms have established various channels for the submission of requests by public authorities and users, including via flagging tools on their sites, and had to significantly ramp up the human resources dedicated to the screening of such notices. Efforts are also under way to reach agreements on expected reaction times, taking into account the difficult constraint of reconciling efficiency and respect for due process.

Beyond notification, the most difficult issue is defining the degree of responsibility that intermediaries should have to more proactively "monitor" user-generated content. This raises two important questions.

The first one is the very appropriateness of moving beyond relying only on notifications. Moving to a full regime of responsibility and pre-screening akin to the one of publishers remains highly controversial and would require significant legislative changes. Yet, intermediaries do have responsibilities in terms of respect of both national laws and human rights. Any voluntary regime should likely combine three elements: a clearly defined scope in terms of the type(s) of expression covered and expected actions, corresponding liability protections for companies implementing it, and adequate protections for users' human rights. Established practices of impact and risk assessments could be transposed in this context.

Question: *Does the notion of "duty of care" sufficiently encompass these different dimensions to serve as an overarching approach?*

The second question is related to the growing use of algorithms and artificial intelligence for detecting inappropriate content. Significant developments are under way in this field, including the use of hash databases, but the effective performance of these tools should not be overestimated. Analogies with approaches used for targeted advertising are not relevant: what is at stake here implies nuanced decisions with important human rights dimensions that can require determination according to detailed national laws. The extreme importance of context, possible algorithmic bias, and the impact of false positives and negatives push, inter alia, for avoiding automatic takedowns and always maintaining a human decision step.

Question: *Would coordinating efforts to clarify the challenges of algorithmic detection of abusive content and making algorithmic tools more broadly available and transparent be possible options*?

3. Decision-making

The capacity to make decisions with some enforceability regarding online content restrictions currently rests with national courts, the intermediaries themselves or a combination of both. However, the high volume of decisions to be made and uncertainties regarding the conditions for national courts to exercise personal jurisdiction over foreign intermediaries lead to content restrictions being frequently made in the first instance by intermediaries. In this context, developing an appropriate escalation path is essential, both internally and externally.

In particular, a difficult issue is the question of proper appeal mechanisms. Requests for reconsideration addressed to the company itself represent a natural first mechanism that already exists, but this cannot be sufficient to ensure full availability of independent redress. Recourse to national courts against a company decision does not appear an established practice and implies in any case renewed jurisdictional challenges that can only increase the already high costs and duration of such procedures.

Question: *Would the creation of shared guidelines for the design and application of appeals process-es and/or specific and independent appeal bodies (as ICANN did when instituting a permanent Inde-pendent Review Panel (IRP) in the context of the IANA transition) be options to fill this gap?*

4. Necessity and proportionality

Necessity and proportionality are key agreed principles applying to content restriction decisions. In that regard, the question of the geographic scope of such restrictions has become a contentious issue and another challenge is the grey area of merely controversial content.

Geographic scope

Complex and still evolving jurisdictional criteria determine the competence of national courts upon individuals and entities on the Internet. Recent years saw a clear evolution towards courts exercising personal jurisdiction over internet companies located abroad when they provide services in the coun-try. Companies have gradually accepted to implement local content filtering (through geo-IP tools, among others) against manifestly locally illegal content, in particular as a result of properly justified and processed court decisions, but defend their right to ignore or challenge those deemed below a sufficient standard or not corresponding to a clear violation of their Terms of Service.

Question: *Is local filtering (for instance via geo-IP blocking) a proper default option, and how to de-termine the corresponding standard?*

More recently, a few national courts are reviewing cases where a party solicits a global removal of content. There are strong objections to extraterritorial application of national laws in speech-related issues, and courts in different countries can issue contradictory decisions. Some nonetheless argue that it should be possible to envisage global takedowns if it is necessary to remedy a particularly strong harm or violation of human rights.

Meanwhile, intermediaries voluntarily decide, on the basis of their own guidelines, upon direct requests made by governmental entities or flags submitted by users. Such decisions can indeed result in a global takedown, even if more limited and proportionate restrictions could potentially be available. This has incentivized some public actors to invoke community guidelines instead of their national laws to obtain such global restrictions.

In this context, strictly excluding any extraterritorial impact of national laws while at the same time implementing global takedowns on the basis of Terms of Service, would be somewhat paradoxical.

Question: Can *some strict conditions exist under which global restrictions could, as a matter of ex-ception, be ordered and implemented?*

Even if such conditions could be defined, a corollary question would remain: how to resolve potential conflicts regarding the geographic scope of content restrictions?

Question: *Would specific dispute resolution mechanisms be a possible option to explore?*

Controversial content

Even if each individual determination remains complex, some mechanisms and practices have been progressively developed (including in cooperation with governments) to handle content that is clearly illegal or in violation of community guidelines. However, an increasing number of situations concern content not neatly falling in these clear-cut categories and that can be labelled as "controversial". Such content is often posted by users cleverly navigating the frontiers of both laws and community guidelines.

Question: *It has been suggested that independent third-party advisory panels could help assess such grey zone cases before a decision is made. Is this a valid option to explore?*

5. Due process

Whatever the applicable norms, the modalities for abusive content detection and the distribution of decision-making responsibilities, any cooperation framework would require the development of a set of mechanisms ensuring due process across borders, including: transparency, formats of requests, notification of users, and remediation.

Transparency - Deepening, broadening and harmonizing transparency reporting on content restrictions (including by governments), as well as encouraging adoption of this practice by a wider range of operators would be clearly beneficial. Yet, caution should be exercised regarding "transparency for transparency's sake" and the purpose of such exercises should be clearly identified. In addition, without aiming at a set of fixed categories of abusive content, efforts towards some convergence of terminology[7] (including to address the challenge of translations) would also be useful.

Question: *How to organize stronger coordination among efforts currently underway in that regard?*

Requests components - The core components that properly documented government requests should contain are generally agreed upon, including: identification of the requester, minimal targeted item, relevant national law invoked, national procedure followed, precise action requested and justifications of proportionality and emergency (when applicable). Flagging tools for users also have diverse but similar requirements.

Question: *Without necessarily aiming towards uniform and standardized forms, how to further document these elements to make it useful for all actors*?

Notification - Notification of users, preferably before decisions are taken, effective and accessible ways for their defense to be heard and the production of rationales for decisions are essential procedural guarantees. A major challenge however is to reconcile this desire for the highest standards possible with the constraints of dealing with extremely large volumes in very limited time, without creating excessive administrative burden.

Question: *Would diversified escalation paths and notification workflows for different types of harm or conditions of emergency be an option to explore*?

Remediation - Beyond the availability of appropriate appeals, mechanisms for the prompt reinstatement of content abusively restricted should be clarified. Additional tools, such as interstitial warnings regarding certain types of sensitive content could also be developed.

Question: *How to document or develop best practices regarding optimal technical modalities?*

---

[7] Some efforts at standardization of labeling were conducted for the Columbia University's Freedom of Expression Database, and the Stanford Center for Internet and Society's Intermediary Liability Database.