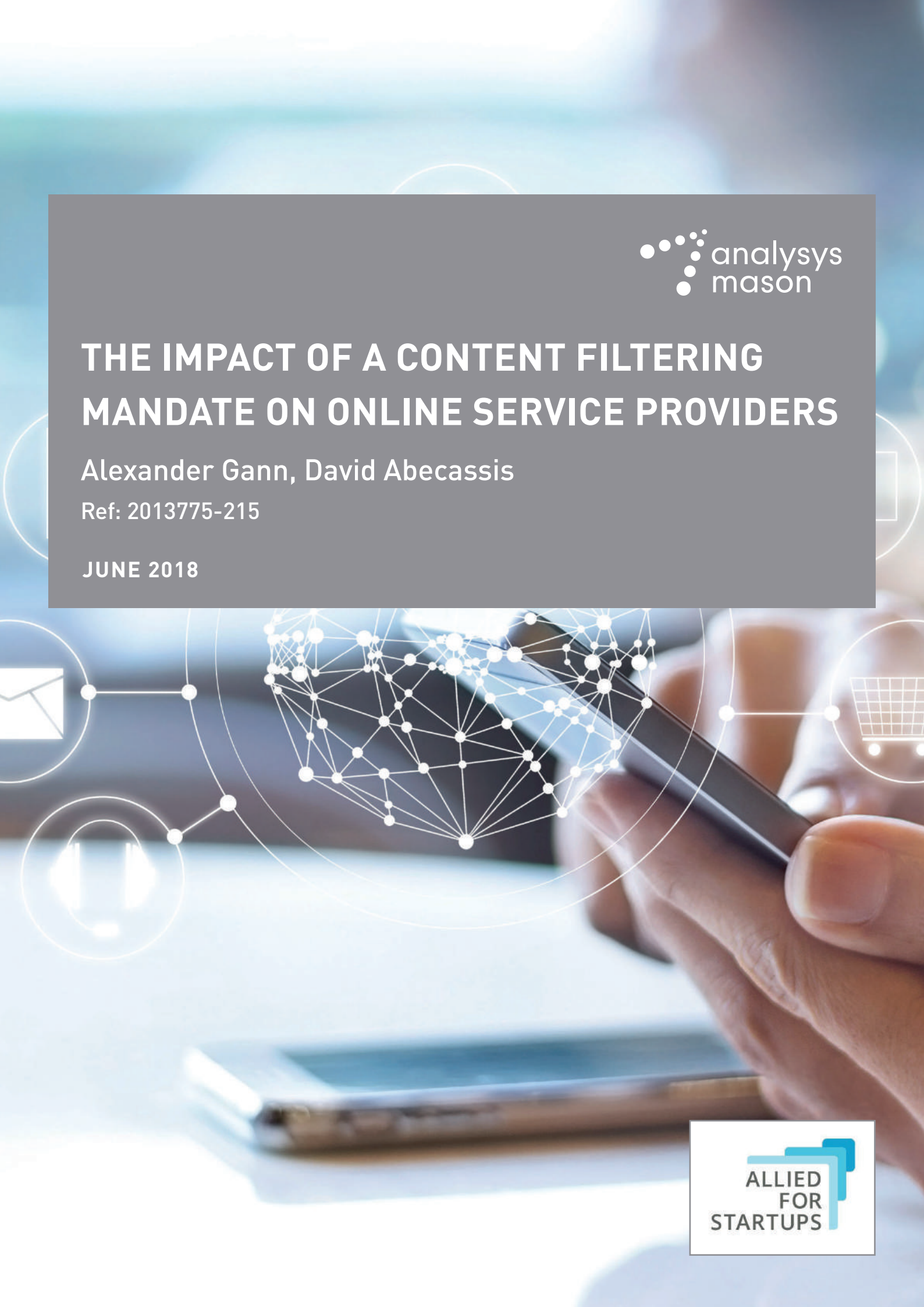

analysys mason

THE IMPACT OF A CONTENT FILTERING MANDATE ON ONLINE SERVICE PROVIDERS

Alexander Gann, David Abecassis

Ref: 2013775-215

JUNE 2018

ALLIED FOR STARTUPS

Introduction

Legislators in the EU have proposed the introduction of mandatory content filters for providers of online services, intended to better protect the rights of copyrights holders and to ensure they receive an equitable share of the value created by those rights. While work has been done on the technical feasibility of such filters, little time has been spent on considering the practicalities of implementing such a mandate and the costs this would impose on the affected parties.

Legislators in the EU have proposed the introduction of mandatory content filters for providers of online services, intended to better protect the rights of copyrights holders. While work has been done on the technical feasibility of such filters, little time has been spent on considering the practicalities of implementing such a mandate and the costs this would impose on the affected parties. The purpose of this study is to explore the likely impact of such an obligation on the providers of online audiovisual services and to quantify the associated direct and indirect costs and benefits. For this purpose, Analysys Mason has drawn on its own research as well as interviews with online service providers, start-ups, producers of content-recognition technologies and legal experts to examine the potential impact of the legislative proposal.

The language of the proposed directive suggests that all types of online service providers will have to proactively filter content, regardless of its nature. The proposed text of the directive mandates platforms to comply using technical measures, although these are not specified. Whilst certain relevant technical components exist for limited types of content, no such solutions exist for the vast amount of content and platforms that appear to fall within the scope of the proposed directive. A content-filtering mandate would impose high costs on those platforms for which limited solutions already exist, while the possibilities of developing additional technologies to comply with the proposed legislation are fraught with difficulty and potentially extremely costly.

- Section 2 broadly describes the proposed legislation and the legal issues raised by the changes.
- Section 3 examines the various technical components that are necessary for content-filtering technologies to work in practice and provides an overview of existing commercial solutions.
- In Section 4, we explore how automated content-filtering technologies could be implemented in practice and what the associated costs are.
- Section 5 concludes.

2 The European Commission proposal for copyright reform

In September 2016, the European Commission published its proposal for a new Copyright Directive,[1] which constitutes a part of its Digital Single Market Strategy[2] and aims to modernise European copyright laws. One of the proposal's stated goals is to tackle what some rights holders perceive as a "value gap". This concept, which has largely been pushed by rights owners in the music industry,[3] refers to the idea that revenues generated using copyright-protected content are supposedly not distributed equitably amongst among creators of content and businesses that rely on the distribution of such content. Traditional music publishers have argued that online service platforms such as YouTube were able to gain an unfair position in rights negotiations, thanks to the "safe harbour" provisions under the e-Commerce Directive,[4] which only require them to remove copyrighted material upon receiving notice from rights holders.

The proposal's Article 13 would oblige "information society service providers that store and provide to the public access to large amounts of works or other subject-matter uploaded by their users" to "take measures to ensure the functioning of agreements concluded with rightsholders for the use of their works" or to "prevent the availability on their services of works or other subject-matter identified by rightsholders through the cooperation with the service providers." [5]

The language of Article 13 suggests a broad application of the mandate that encompasses all information society service providers, including those that distribute user-generated content (UGC), while the absence of a reference to any particular type of content, such as video or imagery, implies that all forms of copyrightable content would fall within the scope of Article 13. While the mandate is intended to apply to those services that make "large amounts of content" available, this is not defined and in practice the amount of content is a poor indicator of the revenue of a platform for example. As a result, there is a real risk that many online service providers (OSPs)[6] that make UGC available would be subject to the proposed mandate in Article 13, and obliged to ensure the functioning of rights agreements or prevent the availability of copyright-protected material on their services.

In practice, Article 13 appears to require the use of content filters, as it is not apparent how OSPs could otherwise comply with the mandate. If they reach an agreement with rights holders, they will have to verify whether uploaded content is covered by these agreements, while the absence of any agreements would still require the deployment of content filters to prevent the uploading of content that is found to be copyrighted and outside of agreements with rights holders.

In May 2018, the European Council published its own position on copyright in the Digital Single Market, which frames the potential obligations on platforms differently to the European Commission's proposal.[7] The Council's position is that OSPs are fully liable for copyright infringements of UGC uploaded to their sites, except in limited circumstances. In paragraph 5 of Article 13 of its position, the European Council states that, in the absence of agreements with rights holders, online platforms are to apply "effective and proportionate" measures to limit or prevent the availability of copyrighted content. Even though the term 'filtering' is not explicitly referenced, rights holders, and ultimately the courts, are likely to interpret this as implying the use of content filters, giving rise to technical and commercial difficulties similar to those caused by the European Commission's proposal.

While the Council's position aims to mitigate the disproportionate costs of a content-filtering mandate for start-ups and other small companies by stating that such enterprises should be "expected to be subject to less burdensome obligations than larger service providers", this exemption is insufficient to reduce costs for start-ups as such businesses would only enjoy a temporary protection under such a clause due to their rapid growth rates.

[1] European Commission, "Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market", 2016. URL: https://ec.europa.eu/digital-single-market/en/news/proposal-directive-european-parliament-and-council-copyright-digital-single-market

[2] Digital single market, 2018. URL: https://ec.europa.eu/commission/priorities/digital-single-market_en

[3] International Federation of the Phonographic Industry, "Global Music Report 2018".

[4] European Commission, "Proposal for a Directive of the European Parliament and the Council on copyright in the Digital Single Market", Article 13, 2016.

[5] Ibid, Article 13.

[6] We use OSP to denote online service providers, which broadly fall under the category of "information society service providers" under the eCommerce Directive.

[7] European Council, "Proposal for a Directive of the European Parliament And of the Council on copyright in the Digital Single market – Agreed negotiating mandate", 2018.

3 Content filters for copyrighted material

For OSPs to comply with Article 13, the European Commission has referenced the imposition of content filters for automated detection of copyrighted content. In order for content filters to be effective, they require several building blocks to be in place: the actual software needed for identifying a unique piece of content, a database of copyright-protected content against which each unique user-generated piece of content can be compared, and potentially agreements with rights holders as to what to do when a match is found. The availability of these building blocks varies widely depending on the type of content, such as audio, video, images, software code, text, and more recent technologies such as 3D printing files.

This section will provide an overview of the technologies that are currently available to analyse different types of digital content, a discussion of existing content databases, as well as a broad overview of existing content-filtering technologies used to identify copyright-protected material.

3.1 Necessary elements for successful identification of copyrighted content

For the successful automated detection of copyright infringement, content-filtering technologies must be available for each type of UGC format. Once UGC has been uniquely identified, an infringement is detected by comparing it against a centralised database, which must be maintained and continuously updated with the co-operation of rights holders and publishers.

Content-identification technology must be available

Currently, content-identification technologies that are able to recognise some content in some contexts exist for video, audio and images. Through these technologies, it is at times possible to apply techniques such as fingerprinting to uniquely identify the relevant content in a timely fashion. In some cases, however, these technologies can be unreliable, ineffective or overly costly, as discussed later in this paper, and may therefore not represent "effective and proportionate" solutions from the perspective of either rights holders or platforms.

Furthermore, other forms of content do not lend themselves to such established identification techniques and are therefore not uniquely identifiable using automated means. For example, files containing designs for 3D printing have several characteristics that currently prevent their unique identification. 3D printing files are large and complex, which means that the time it takes to process a file and determine its unique characteristics is exceedingly long and requires a great deal of computing power. An object can be made to represent a copyrighted design, even though the 3D printing files appear to be completely dissimilar to content-recognition software. Even if it were possible to uniquely determine the identity of a file, small modifications and distortions to the file would make it difficult for a program to compare it against a database of objects.

Rights holders and publishers must provide copyrighted material to centrally accessible databases to allow matching of content

The existence of content-identification technology is a necessary, but not a sufficient condition, as databases with copyrighted content are needed against which online service providers can compare user-generated content. Companies such as Gracenote maintain databases of up to 200 million songs and rely on the co-operation of rights holders such as music labels to submit their catalogues and rules for copyrighted material. For video content, national broadcasters and film and TV production studios submit their content to firms such as INA. Without such databases, it would not be possible to determine whether a unique piece of content is violating copyright or not.

The sheer volume of content as well as the existence of multiple rights holders[9] make this problem a very hard one for all types of content. For industries such as 3D printing, software, or text, centralised databases simply do not exist. When a user submits a 3D printing design, the determination of copyright infringement using automated means would entail the comparison of the design against all known, copyrighted physical objects, which is patently impossible. Depositories for user-submitted software code or written text are similarly unable to avail themselves of content-recognition technologies using centralised databases. The imprecise language of the proposed directive leaves companies in these areas without guidance on what types of works are covered by Article 13. In interviews, operators of platforms that engaged in textual analysis of websites or allowed users to post text expressed concern that Article 13, as currently formulated, would imply they would have to verify that every piece of written content has been checked for copyright violation, which would be technically and commercially impossible.

[1] This encompasses a complex reality: there are many rights holders in absolute terms for some types of content, and this is only growing as user-generated content enables more creators to generate copyrightable content; there can be many rights holders involved in the same piece of content (e.g. in a video, the director, script writer, producer and music providers all have rights); finally there can be many concurrent non-exclusive licensees: important, you can only filter on the basis of an exclusive right; for example, if a broadcaster has a music licence for phono rights from the PPL, that does not entitle it to block content.

Figure 1 provides a stylised example of the various necessary components for content-filtering technologies to work in practice.

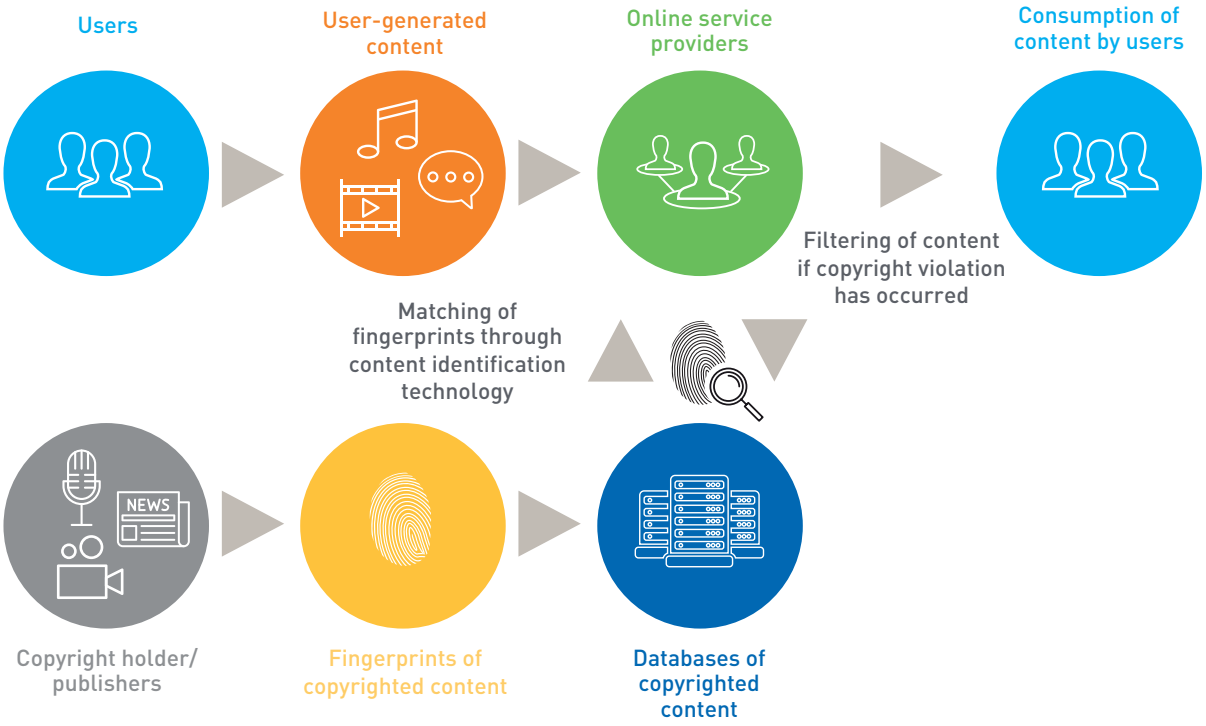

Figure 1: Overview of content-filtering process for copyrighted content [Source: Analysys Mason, 2018]

3.2 Content-filtering technologies

Content-filtering technologies exist and can be applied to certain types of content, but have significant limitations, as they are neither able to perform completely accurately nor applicable to the content of every single rights holder in the world. Looking beyond the limitations of existing filtering technologies, for many types of content, such as 3D printing or software code, no such technology exists.

Types of content filters

Several technologies are available to recognise digital content, which normally means matching an unknown piece of content to a database of known content. These technologies differ by their functionality, the type of content they can be applied to, as well as their accuracy and efficiency. Not all technologies can be used interchangeably to identify all different types of content, and for certain content types there is no readily available technology.

The simplest type of content-recognition technology is based on a content metadata search, which relies on an analysis of data that surrounds the content itself, or "metadata".[10] This data may contain information about the associated file, such as, in the case of a song, the name of the artist, or the duration and title of the song. The metadata of a piece of content can be searched against a database of copyright-protected works, without the need for downloading and analysing the content itself. This approach has significant limitations, however, because metadata can be easily modified by users to obscure information about the content and is often inaccurate or imprecise.

More complex methods offer more robust approaches:

- With **hash-based identification**, a piece of content is represented numerically by a "content hash", a numerical simplification of the original file. The file size of this numerical representation is significantly smaller than that of the original file, which makes it more efficient when comparing the hash value of an uploaded piece of work against a database of hash functions of copyrighted content. However, one of the drawbacks of the technology is that any alteration of the file will also alter the hash value. Hash matching can therefore only detect identical copies of the same file. For example, converting the file to

[10] Evan Engstron and Nick Feamster, "The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools", 2017.

another format, editing its length, or changing the compression will alter the hash value and prevent the detection of a copyright infringement.

- **Fingerprinting** is a more sophisticated technique, which analyses a given piece of content to determine its unique characteristics.[11] Unlike the previously discussed technologies, fingerprinting analyses the piece of content itself, which makes it more robust to modifications and alterations. The inherent characteristics of the piece of content are stored in a smaller fingerprint file, which can be compared to a database of fingerprinted content. In the case of a video file, for instance, the fingerprinted file is independent of resolution and format, and can be used to identify complete videos as well as some short snippets and manipulated versions. Most of the widely used and most advanced content-filtering technologies use some form of fingerprinting.

Algorithms used in content-filtering technologies are specific to the type of content they are designed to recognise. Software that identifies audio content and compares it against a database of audio files cannot identify images or video. While an online platform that only allows the submission of audio files would only have to analyse user-generated content using software developed for audio content recognition, an online platform using user-generated content for video, audio and images would have to use separate filters for each of the content formats.

Case Study 1.Examples of content recognition, matching and filtering technologies

LTU Tech is a technology company that provides image recognition technologies, which are sold as either licensed software or via a hosted platform. Its software allows the indexation of a collection of millions of images in a private database, which can be stored on a server and used to run queries against. Customers of LTU Tech use the software for different purposes, such as the detection of counterfeit goods. Manufacturers of luxury goods submit images of their products and logos, against which LTU Tech compares images from specific shopping sites to identify sales of counterfeit goods.[12]

The European Commission has listed LTU Tech as an example of content-recognition technology that could be used by affected OSPs to comply with Article 13.[13] However, LTU Tech only provides content-recognition software and requires other parties to provide a database of content against which queries can be performed. The use of technology by LTU Tech alone is therefore not a viable way of complying with the mandate.

Audible Magic provides compliance services for web hosting services that rely on user-generated content, using fingerprinting technology to match audio and video content uploaded by users against a database of fingerprints submitted by content owners. Audible Magic will report to the content-sharing site whether a match was found, which can either result in the piece of content not being removed from the site or in the distribution of revenues to compensate the rights holder. Like Gracenote's service, rights holders have to submit content to Audible Magic's Global Content Registry in order for their copyrighted material to be searchable.

Signature, a product developed by the Institut National de l'Audiovisuel (INA), is a software that enables online service providers to detect copies of videos using fingerprinting technology and to either block the content or monetise it. Websites such as Dailymotion, which allows users to upload videos, use Signature to automatically filter content. Signature also relies on a database of referenced content provided by rights holders such as film and television studios, sports organisations and broadcasters.

[11] Dominic Milano, "Content Control: Digital Watermarking and Fingerprinting", 2012.

[12] LTU Tech official website, 2018. URL: http://www.ltutech.com/

[13] European Commission, "Commission Staff Working Document – Impact Assessment on the modernisation of EU copyright rules", p.171, 2016.

Challenges

While content-filtering algorithms have become more accurate over time, they are still not sophisticated enough to detect all matches between uploaded UGC and reference databases, and are also liable to yielding false positives, i.e. indicating a file violates a rights holder's copyright, even if this is not the case. Our interviews with OSPs that use content-filtering technologies suggest that such algorithms are mainly used to flag content that is thought to violate copyright, rather than to replace human judgement outright. According to interviewees, automatic filters could usefully be employed to detect potential copyright violations but were not accurate enough to do away completely with human oversight.

While numerous solutions for audiovisual content allow some level of identification of copyright-protected content, these techniques are imperfect and can result in false positives or false negatives when analysing files.[14] There are few services aimed at identifying copyright-protected content in other content areas. For instance, currently no solution exists for platforms that rely on user-generated content in the areas of 3D printing, software code or text. This is due to several factors, such as the difficulty of applying content-identification technology to the content itself and the sheer quantity of content produced, which makes the creation of a database of copyright-protected material extremely challenging.

3.3 Databases of copyrighted content

To detect copyright-infringing content, rights holders must provide their material to be stored in databases against which OSPs can check the unique fingerprints of UGC. For most types of content, no such central reference databases are maintained, although they are essential to any automated copyright infringement detection mechanism.

While the European Commission references content-recognition technologies, it does not make any concrete proposals regarding this aspect of the technology.

Content-recognition technology enables a piece of unknown content to be compared to a database of files which encompass reproductions of documented works. For example, to determine whether an image is copyright protected, content-identification technology has to be applied to the file in order to analyse it and then compare it against a database of copyright-protected images. In order to do this, rights holders and publishers have to provide files and information on their catalogues of copyrighted content, which then have to be stored in a database than can be accessed by content-identification algorithms, and in a form that is suited to these algorithms.

The establishment of comprehensive databases of copyrighted material therefore appears essential to the applicability of the mandate suggested in the Article 13 proposal. However, it poses significant commercial, technical and legal challenges, which remain largely unaddressed. For each type of content, rights holders must co-ordinate to establish a database and submit their content, in a form that is compatible with content-recognition and -filtering algorithms. In industries where there are relatively few, large rights holders, this co-ordination problem can be more easily overcome than in industries where there are many individual rights holders. There is a relatively low number of globally operating rights holders in the music industry, for example, while rights holders in the world of images are much more numerous. As a result, the existing databases dedicated to storing copyright-protected music, such as Gracenote, are more comprehensive than those dedicated to images, which are generally publisher-specific and fragmented.

Case Study 2. Examples of content filters used in combination with content databases

Providers of content-recognition technology for audiovisual content typically rely on fingerprinting technologies to identify files. **Gracenote** offers several products that are used by companies such as Apple, Amazon or MixCloud to help identify songs uploaded by their users. Gracenote has a database of over 200 million songs against which it can match submissions from its clients and processes 20 billion queries per month. Its clients use its MusicID product to identify music ripped from CDs and files purchased from online stores, as well as using the technology to aid in paying rights holders. For example, music platform Mixcloud, which allows users to upload DJ sets, radio shows and podcasts, uses MusicID to identify songs or extracts of songs within users' mixes to determine how to distribute revenues to the relevant rights holders.

[14] See Engstrom & Feamster, The limits of filtering, 2017, http://www.engine.is/the-limits-of-filtering/, for a more detailed analysis of the technical limitations of existing filtering technologies.

Within the European Union, each member country is left to implement its own copyright law and violations are dealt with at a national level. This could create further complications for the establishment of pan-European content databases, as the exact definitions of copyrightable content differ amongst member states. Content databases would have to reflect national law, potentially giving rise to separate content databases for each country, against which OSPs operating within that territory would have to perform their identification checks. Such a proliferation of content databases would, however, increase the difficulty of operating across borders and would be detrimental to furthering the goals of creating a Digital Single Market within the European Union.

4 Implementation and cost of content-filtering mandate for OSPs

4.1 Implementation of content-filtering mandate by technology

Depending on the industry, the implementation of the content-filtering mandate would either be very costly or technically unfeasible as a result of commercial or technical limitations. Content-identification technology is not readily available for all types of content, while rights-holder co-ordination and technical constraints pose substantial challenges that would have to be overcome to allow the establishment of centralised content databases.

Article 13 refers to "[i]nformation society service providers that store and provide to the public access to large amounts of works or other subject-matter uploaded by their uses", which implies that any OSP that relies on user-generated content is obliged to ensure that copyrighted content is identified using filtering technologies. This does not just encompass companies operating with audiovisual content, such as platforms like Soundcloud, Mixcloud or Dailymotion, but also affects companies in areas such as 3D printing, natural language processing and software code.

As described earlier in this paper, the ability to comply with a content-filtering mandate in keeping with Article 13 differs widely by industry, company and content form. The following section gives an overview of the extent to which companies using different forms of UGC would be able to comply with a content-filtering obligation by examining the available filtering technologies for each content type. Depending on the content type, different content-filtering technologies exist today that allow OSPs to partially comply with Article 13. However, for the vast majority of OSPs, no such solution exists.

Audio content

Online service providers that allow users to upload audio files have several technologies at their disposal to identify content and compare it against database of copyrighted content. Firms such as Gracenote or Audible Magic provide content-identification software and maintain databases in co-operation with rights holders, which allows them to compare queries from OSPs against continuously updated databases. Platforms operating in this domain, such as Soundcloud and Mixcloud, have announced their co-operation with certain content-identification technology providers, which suggests that the technological barrier to implementing this solution is not excessively high.

However, it is not evident that the implementation of these measures would be sufficient to comply with the law. The music industry is characterised by several large rights holders and publishers that control a substantial share of the existing catalogue of music, but there are several thousand smaller rights holders acting globally whose works might not be a part of the databases and therefore not identifiable for OSPs, which might leave OSPs liable if this copyrighted content is distributed via their platforms.

A similar source of uncertainty is the fact, discussed above, that technical solutions are not perfect and will not detect all copyright-infringing content. Content-identification software has been criticised by rights holders and users for not being able to correctly identify all submitted content, which is highly problematic from the point of view of a company relying on them to detect copyright violations.[15]

Video content

Like audio, technological solutions for the identification of copyrighted video content exist to a certain degree and are already deployed by OSPs. Google's Content ID, INA's Signature, and Gracenote's video products are used by YouTube, Dailymotion and other video OSPs to detect copyright-infringing content and to allocate revenues to rights holders. Technical solutions for video exist, but using them to comply with a content-filtering mandate would confront OSPs with the same issues as those of audio content filters. Not all rights holders are contributing their content to copyright databases, and the pace of newly generated video content, particularly that created on

[15] Dr Christina Angelopoulos, "On Online Platforms and the Commission's New Proposal for a Directive on Copyright in the Digital Single Market", p.39, 2017.

technologies other than film and television, is so rapid that maintaining an up-to-date database poses a severe technical challenge. Similarly, while there have been great advances in identification technologies, it is still possible to bypass filters by distorting the content, the speed at which the footage is played, or through other modifications, leaving OSPs liable if copyright-protected content passes the filter.

The establishment of content filters by players such as INA in France as well as providers such as Audible Magic suggests that fragmentation can be an issue in the future if databases are not centralised. Large publishers that have already invested in the creation of databases might not be persuaded to join other databases and share their content, as they will prefer to not incur intermediation costs. OSPs will then be faced with having to check UGC against multiple platforms, which will increase their costs.

Static image content

While image-identification technology is relatively well developed and made commercially available by providers such as LTU Tech, no centralised databases of copyrighted images exist. OSPs operating in this space do employ proprietary techniques to try and identify copyrighted material, but the identification is based on matching a submission against their own catalogue of images or, in certain cases, against catalogues provided by partner companies such as Getty Images. Where automated filters are deployed, these serve to flag content, rather than taking it down automatically, and decisions are reviewed manually. As a result, OSPs rely on notice-and-takedown mechanisms to identify copyrighted material. Implementing a content-filtering mandate for user-generated image content would therefore require the existence of centralised content databases.

Other content

The challenge of using automated content filters to identify copyrighted content is greatest in areas where none of the necessary technology currently exists. OSPs with business models that rely on such user-generated content text submissions, 3D printing files or social media content are neither able to uniquely identify such content nor compare it against centralised databases. As discussed above, a company that allows users to submit 3D printing files does not possess the technology to analyse each submission for its individual characteristics, nor is there a database of physical objects against which it could compare such a file.

Case Study 3. Shapeways

Shapeways is a 3D printing marketplace and service that allows consumers and businesses to buy or submit 3D printing files to be printed by Shapeways' industrial 3D printers. Users can upload their designs for 3D objects to the Shapeways marketplace, where other users can choose to purchase a design, and have it printed by Shapeways. The user who originally created the 3D design receives a share of the total revenues from the sale of the final object.

Shapeways relies on UGC and implements a notice-and-takedown policy to deal with any copyright violations. The firm receives several thousand take-down requests a year, which have to be reviewed individually in order to determine if an infringement has actually occurred. For physical objects, the violation of copyright is less straightforward to determine than in the case of audio or images, and cannot be automated using existing technologies.

Current technology makes it impossible for Shapeways to comply with Article 13, as the technology to uniquely identify 3D printing files does not exist. Furthermore, no reference body of copyright-protected physical works exists against which Shapeways could compare user-submitted 3D designs. It is therefore difficult to see how Shapeways could implement systems that allow it to comply with a content-filtering mandate without severely modifying its business model.

Numerous companies operating in these spaces have reported that they would not be able to comply with a copyright-filtering mandate and that such an obligation would cause their businesses considerable legal uncertainty. In all of these cases, businesses stressed that the current system of notice-and-takedown is the only way for them to deal with copyright infringement, as each case requires a manual review and is highly dependent on the context. Furthermore, rights holders within each industry cannot be characterised as a homogenous bloc, but have differing preferences when it comes to distribution, with some taking a more relaxed approach than others and seeing the diffusion of their content as a way of generating additional revenue.

Case Study 4. Sentione

Sentione provides social listening analytics to companies that wish to monitor their social media and news presence. The company has developed proprietary software that allows it to track organisations and topics discussed on social media platforms, blogs, message boards, as well as other media sources. The software scans websites for content and uses language-processing algorithms to generate sentiment analysis and help organisations understand how they are being perceived online.

As Sentione automatically processes content from other websites, it co-operates with website administrators and publishers by using the robots exclusion standard, which allows websites to determine which part of their content can be searched by web robots. Additionally, they provide the opportunity for publishers to request a take-down of their material from Sentione's platform. In practise, publishers are often Sentione's customers, as it allows them to track the reach of their content and the effect it is having on other platforms.

Complying with an automated content-filtering mandate would currently not be possible for Sentione, as neither the technology for identifying unique user-generated content used in its sentiment analysis software nor a database of copyright-protected content for cross-referencing such content currently exist.

4.2 Costs of content-filtering mandate for platforms

Given the lack of universally available filtering technologies and the fact that the existing technologies are tailored to specific use cases, it is difficult to estimate the overall cost of implementing such systems for a company operating in such a space. An OSP that allows submissions of audio, video and image content would have to use a combination of different content-filtering technologies to uniquely identify each submission before comparing them to databases of copyrighted content. A discussion of costs will therefore only be indicative and incomplete, as the necessary solutions are not commercially available.

Solutions within the audiovisual sector that are tailored towards copyright compliance, such as Audible Magic, are based on subscription models that offer a certain number of queries in return for a monthly service fee. Audible Magic requires a USD2500 set-up fee, as well as a monthly service fee of USD1000 for up to 10 000 queries per month in the case of music and film/TV, which increases up to USD2382 and USD1602 respectively for up to 50 000 queries.[16] These figures need to be put in perspective with the initial capital and average revenues of European start-ups to get a sense of their magnitude. The 2016 European Start-up Monitor, which surveyed 2515 start-ups, found that 22.1% were in the seed stage and were not generating any revenue, while 81.2% of the revenue-generating start-ups reported revenues lower than EUR500 000. The imposition of a content-filtering requirement on such companies would leave them unable to afford the sort of fees charged by filtering technology providers such as Audible Magic.[17]

[16] https://www.audiblemagic.com/compliance-service/#pricing

[17] European Startup Monitor 2016, URL: http://europeanstartupmonitor.com/

Case Study 5. Illustration of costs of content copyright obligation for a start-up

The costs of complying with Article 13 can be illustrated using the example of a hypothetical start-up that operates as an OSP which allows its users to post audio and video content on its platform. In this case, a small- to mid-sized OSP can be assumed to have 3 000 000 users, who upload, on average, two pieces of content per month, resulting in 6 000 000 uploads of UGC to the platform per month.[18] This is conservative when compared to the amount of content uploaded on other platforms, such as Snapchat, where users each upload nearly 50 images or videos per month on average.[19]

For the start-up to comply with Article 13, it would have to identify the unique fingerprint of each piece of uploaded content and compare it against a database of copyrighted content. If copyright databases are not centralised, it will have to compare each fingerprint against several databases.

The OSP could use services priced at rates similar to those of Audible Magic to scan audiovisual content. According to Audible Magic's pricing, the cost per query is c.EUR0.05 for audio and EUR0.03 for video.[20] Taking this figure as a reference and assuming a slightly lower cost of EUR0.04 per query, this start-up would have to spend EUR240 000 per month to comply with a content-filtering mandate.

This total figure can be compared to overall revenues of such a business. Content platforms such as Snap and Facebook reported average revenues of EUR0.58 and EUR7.85 per user for Q4 2017 in Europe, respectively.[21] **Assuming an ARPU of EUR2.00 per user per month** (significantly higher than Snap's), the OSP in this example earns revenues of EUR6 million. **The imposition of a content mandate would therefore increase its costs as a percentage of revenue by approximately 4 percentage points.**

However, if content databases are not centralised and the start-up has to query multiple databases, costs can quickly spiral. Each additional database against which UGC has to be compared would double the cost of complying with the mandate, imposing significant costs on the start-up. **If there were five databases to check,** costs could be about EUR1 200 000, or **20% of revenue.**

As noted above, many start-ups have no, or only low, revenues in the immediate years following their establishment, making it even harder for them to bear the additional costs associated with content filtering. Such a measure would place them at a competitive disadvantage relative to more established players, reducing competition and stifling innovation.

For the technologies where no content-filtering software exists, such as 3D printing, text, scientific writing or images, the costs of implementing such a technology are likely to be function of several factors and to scale significantly with usage. For example, the number of rights holders for image content is much higher than that for music or broadcast content, making the co-ordination between them extremely difficult. Currently, numerous databases for such content already exist as they are maintained by individual publishers and rights holders, but implementing a centralised, cross-national database poses a severe technical and organisational problem. A large amount of server capacity would be required to initially set up a database and would have to be continuously updated with all copyrighted imagery published on the internet. Costs would therefore increase in line with server capacity as well as usage of the database by OSPs, which would rely on API calls to submit user-generated content for matching against stored copyrighted content. In the case of images, usage by OSPs would be extremely high and require an extraordinary number of costly API calls. For example, in 2017, Snapchat saw 210 000 images uploaded per minute, while Instagram recorded 65 000 images uploads per minute, and Twitter 350 000 tweets per minute, many of which contain image content.[22]

Our conversations with OSPs and content technology providers indicate that most saw the imposition of a content mandate outside of very narrowly delimited cases, such as for already large and existing platforms in the audiovisual sector, as technologically unfeasible and prohibitively costly.

[18] For comparison's sake, German social media app Jodel reported user numbers of "several millions" (URL: https://www.zeit.de/campus/2017/04/jodel-app-start-up-studenten-wirtschaftswissenschaften), while Mixcloud is estimated to have 1.2 million active daily users (https://news.crunchbase.com/news/mixcloud-raises-11-5-million-series-djs/).

[19] Statista, Digital Economy Compass 2018; Company annual report.

[20] https://www.audiblemagic.com/compliance-service/#pricing

[21] Company annual reports.

[22] Statista, Digital Economy Compass 2018.

4.3 Costs and implementation issues for rights holders and broader economic costs

Successful deployment of content-filtering technology depends to a large extent on the provision of content referencing databases and their continuous updating. The establishment and maintenance of such databases is dependent on the co-ordination and co-operation of rights holders, which requires significant investment on their part. In areas where rights-holder fragmentation is relatively low, such as in music, rights holders have been able to co-ordinate efforts to establish comprehensive content databases as well as persuading smaller publishers to contribute. Even then, the costs of developing sufficiently precise content-filtering technologies and the requisite databases has taken several years.

For many other forms of content, industries are too fragmented, both within markets and across geographies, to facilitate co-ordination of the creation of the necessary content-filtering infrastructure. Within fragmented industries such as images and scientific publishing, publishers maintain their own databases, which would require OSPs to query multiple databases whenever they receive UGC. In such cases, the cost per query would multiply for OSPs depending on the number of databases against which they would have to check UGC.

In addition to the direct costs imposed by content-filtering technologies, Article 13 is likely to impose significant indirect costs on online service platforms in Europe. During several of our interviews, respondents voiced concerns about Article 13 in its current form and stated that, as they might be unable to comply with the mandate due to the absence of filtering technologies in their industries, European operations would have to be severely restricted in order to minimise liability for potential copyright infringements. This would hamper the growth of such platforms in Europe and place European companies at a disadvantage relative to firms operating in the USA or Asia.

5 Conclusion

The content-filtering mandate as proposed by the European Commission and taken up by the European Council requires technological and commercial solutions that are currently not available for many types of content, and not available to many companies, including in particular the start-ups and scale-ups that are essential to the growth of the technology sector in the EU. The automated recognition of content is not possible for all forms of content that are used by online service platforms, such as 3D printing or software code, and is not likely to be developed in the immediate future. OSPs that use such content would find it impossible to comply with such a burdensome mandate and would be subject to large legal uncertainties related to their compliance with copyright law. In the case where content-filtering technologies exist, such as for audio or video content, they are not accurate enough to correctly detect all types of content and can be circumvented by distorting or manipulating files.

In addition to content-recognition technology, the successful identification of copyright violations requires the existence of content databases against which UGC can be compared. In the case where filtering technologies exist, these databases are at present either fragmented or non-existent. Currently, certain rights holders maintain their own databases, for instance for video content, and harmonisation of such structures across different countries is likely to be beset by difficulties relating to differences in national copyright law as well as co-ordination problems.

As shown above, while certain companies at present deploy limited content-identification technologies, the cost of these for smaller platforms quickly reaches significant levels as a result of universal application and the need for cross-referencing of content against multiple databases for either multiple rights holders or multiple types of content. It is therefore likely that such a mandate would have a negative effect on the provision of digital services in Europe and would significantly hamper the development and growth of existing digital business models. Given the current technological, legal and commercial environment, a content-filtering mandate would appear to be impossible for a vast number of online service platforms to implement, inefficient at dealing with copyright infringements, as well as imposing significant costs on start-ups and medium-sized companies.

Acknowledgments

The present report has been independently drafted by Analysys Mason and commissioned by Allied for Startups. Allied for Startups works with contributions from corporate sponsors to finance its activities, all of which are listed on its website. For this report a financial contribution from Google was received and is herewith disclosed. The content and research question was not subject to any financial contribution. The authors wish to thank representatives from the following companies for participating in interviews for this research: Dubset Media, EyeEm, LTU Tech, PaperHive, Sentione, Shapeways.