WRITTEN STATEMENT OF ADAM HOLLAND AND CHRISTOPHER BAVITZ
EUROPEAN COMMISSION STAKEHOLDER DIALOGUE ON ARTICLE 17
16 DECEMBER 2019

INTRODUCTION

Thank you very much for having us today — we are happy to be able to participate in this Stakeholder Dialogue regarding Article 17 of the European Union Directive on Copyright and Related Rights in the Digital Single Market. We especially appreciate your being able to accommodate our remote participation from here in Cambridge, Massachusetts.

My name is Adam Holland, and I am a Project Manager at Harvard University's Berkman Klein Center for Internet & Society, where I oversee the day-to-day operations of the Lumen database project that is the subject of our remarks today. I am here along with Christopher Bavitz, who is the WilmerHale Clinical Professor of Law at Harvard Law School, one of the faculty co-directors at the Berkman Klein Center, and Lumen's principal investigator.

In these remarks, we plan to offer a short history of Lumen and the ways in which it has served to facilitate transparency in connection with requests directed to websites, by both private sector and government actors, to remove content or links. Using Lumen data, we will offer some observations about the past and current landscape for online takedowns. We will also provide some specific examples of research that has come out of the Lumen database that may be instructive or informative in connection with considerations about implementation of Article 17 of the Copyright Directive in the EU.

In brief, the experience of this project tracking takedowns for nearly two decades demonstrates that:

(1) a legal regime that envisions an opportunity for private or public actors to demand that content or links be removed from the web will inevitably have to reckon with errors and abuses of the system;

(2) no purely technological solution that we have seen to date is capable of comprehensively addressing such errors and abuses; and

(3) transparency about who is requesting that content and links be removed, and the nature of and basis for such requests, is vital in the context of implementation and administration of any takedown regime, including as an ongoing oversight mechanism for identifying and addressing errors and abuse.

LUMEN OVERVIEW

Background and History

Lumen is an independent research project that studies, and facilitates the study of the landscape for online content. This includes requests — based on legal or extra-legal theories — to remove materials (or links to materials) created or uploaded by Internet users.

Lumen operates a research platform that invites rightsholders, Online Service Providers (OSPs), search engines, and other online intermediaries, as well as members of the general public, to share takedown requests that they send and receive concerning online content. Lumen maintains a database containing millions of such notices that have been voluntarily shared with the project by their recipients and senders, and makes those notices available to scholars, journalists, and others for purposes of research and analysis.[1]

Lumen was formed as the Chilling Effects Clearinghouse in or around 2001, in the wake of the United States' 1998 implementation of the Digital Millennium Copyright Act and its safe harbor provisions.[2] The project was founded in recognition of the fact that the DMCA safe harbor in the US would change the landscape concerning availability of online content by incentivizing platforms to remove content in response to copyright

[1] The Lumen database is accessible at https://www.lumendatabase.org.

[2] The DMCA safe harbor is embodied in Section 512 of the United States Copyright Act, 17 U.S.C. § 512.

claims in order to qualify for the safe harbor, and that it would do so through a process that took place largely out of the public eye. Against that backdrop, Lumen's goals are to facilitate research about different kinds of complaints and requests for removal — both legitimate and questionable — that are sent to online publishers and service providers; and to provide as much transparency as possible about such notices in terms of who sends them, why, and to what effect.

Current Operations

No existing law or regulation mandates that removal requests or demands be made public or even available for study. Thus, all of the notices in the Lumen database have been shared voluntarily with Lumen. And, those companies that have chosen to share notices with Lumen decide which notice types, and which data fields, to share. The set of data to which Lumen has access is unfortunately therefore limited. That said, given that the list of companies contributing data includes some of the web's larger online service providers, we believe that the data set serves as at least a rough proxy with respect to global takedown trends.

Currently, the list of companies and institutional senders sharing copies of at least some of their notices with Lumen includes Google, Twitter, Medium, Wikipedia, Wordpress, Kickstarter, and the University of California at Berkeley, among many others.

Balance of Privacy and Research Interests

Throughout its history, Lumen has sought to achieve a balance between bringing greater transparency to the notice takedowns that its database stores and recognizing and respecting the real concerns and privacy interests of the individuals who are sometimes a part of those notices.

Even in its earliest incarnation as the Chilling Effects Clearinghouse, the project made a concerted good faith effort to focus on the key subject material of DMCA complaints (i.e., the rightsholders complaining and the URLs for the content at issue) and to avoid making public the names of individual agents who were merely part of the process (i.e., lawyers and functionaries who help to make requests and send notices). Relatedly, the project made a good-faith effort to minimize or avoid completely the disclosure of personally identifying information such as street addresses or telephone numbers.

In 2015, as the database modernized its underlying software and continued to grow and expand the topical range of its notice corpus beyond copyright, the project made the

decision to use robots.txt[3] to prevent individual Lumen notices from being crawled by Google and other search engines. More recently, motivated by similar concerns, as well as by a desire to more explicitly foreground the research aspects of the project's mission, Lumen has changed the way in which notice pages are by default presented to members of the public visiting Lumen and viewing notices. As of May 2019, in any given notice, URLs that are the subject of a takedown request are no longer readable in their entirety. Instead, only the domain and top-level domain ("TLD") of each URL is visible. Researchers, journalists, policymakers, and others can request credentials for the site, with which they may see full URLs and file attachments. These are not decisions that the project made lightly, and they are decisions that caused consternation within the community of Lumen users. But, we believe that they have allowed Lumen to strike the aforementioned balance.

Representative Statistics

As of December 2019, the Lumen database contains approximately twelve million takedown notices and receives between five and seven thousand new notices per day, the majority of which are DMCA (copyright-based) notices. These notices represent requests to remove approximately four-and-a-half billion URLs.

Depending on how they are categorized, between twenty and twenty-five companies have chosen to submit copies of at least some of the notices they receive with Lumen. All but one of the companies submitting notices share DMCAs notices, while a smaller subset share a wider variety of the takedown requests they receive. Of DMCA notice recipients, Google submits the largest volume of notices to Lumen by far[4], followed by Twitter, and then a long tail of the remainder of companies. Lumen's database does have a small

[3] A robots.txt file indicates whether user agents (i.e., web-crawling software applications) are or are not allowed to crawl a web page. See. e.g., https://www.robotstxt.org/orig.html.

[4] The results of a search for all Lumen notices with Google as the submitter can be found at https://www.lumendatabase.org/notices/search?utf8=%E2%9C%93&submitter_name=Google&sort_by=

number of copies of notices that have been provided by individuals, usually in conjunction with an unusual fact pattern or obvious violation of DMCA requirements.[5]

With respect to those companies and individuals sending and receiving notices, Lumen's database contains close to five hundred thousand (500,000) different senders of notices, who are sometimes, but not always, the rightsholders whose material is at issue. These rightsholders number approximately three hundred thousand (300,000). By contrast, all of the notices in the database have been sent to approximately eight thousand (8,000) recipients, a statistic that speaks to the presence of large-scale institutional recipients like Google within the full dataset, and to the long tail of OSPs affected by takedown notices.

TRENDS

We wanted to highlight a few trends that we have observed over time, in terms of both what we see in individual notices submitted to the database today as opposed to years ago, and in terms of the overall volume of notices. For example:

- In terms of **overall volume of notices shared with Lumen per month**, Lumen over the past year or two, Lumen has been in the range of approximately 150,000 to 200,000.
- In terms of **overall volume of notices shared with Lumen**, Lumen observed an order of magnitude increase in the volume of submissions from January 2009 to January 2011, another similar increase from 2011 to 2013, then a 100% increase from 2013 to 2014, another 100% from 2014 to 2016, and finally a 50% increase from 2016 to 2018, before leveling off to current numbers. It can be difficult to separate signal from noise in these numbers and even harder to attribute them to particular causes. But, it is clear that Lumen has observed an increase in the number of notices it receives from an average of one (or fewer) per day in 2002, to a steady five- to seven-thousand per day so far in 2019. It took over ten years for Lumen to receive its one-millionth notice but only a little over a year to receive its two-millionth, less than one year for the third million, and only eight months for

[5] See, e.g., https://www.lumendatabase.org/notices/19500204. (A DMCA takedown sent by a public radio station to a website that had written an article commenting on one of the station's recent programs.)

the fourth million. 2018 represented the first time that overall DMCA volume decreased since the project began tracking.

- In terms of the **number of URLs addressed in any given notice**, Lumen has similarly observed a significant increase over time. When Lumen began to collect DMCA notices, each such notice typically contained only one or two URLs, often accompanied by a personalized letter. More recent notices often contain many more URLs, frequently over one-thousand or more in a single notice, and occasionally as many as twenty-thousand. The British Phonographic Industry ("BPI"), for example, is one of the largest senders by volume in the database. Lumen's data show that BPI has requested the removal of more than 200 million URLs from Google alone, at an average of 650 URLs per notice in total, with a much higher average for more recent notices. BPI used 274,810 DMCA notices to request the removal of its first 100 million URLs, (an average of ~384 URLs per notice) but fewer than 55,00 notices to request the removal of the next 100 million (an average of more than 1,800 URLs per notice).

- Increases both in total notice volume and in the number of URLs per notice appear to be linked to a significant degree to the **rise in the use of automated techniques for sending and receiving notices**, a trend that our data suggest really took off in early-2012.[6] Google, typically a bellwether of larger trends, saw notice volume increase from only a few requests per week in 2008 to one request to remove an item from search results every 0.08 seconds in 2014, a rate unachievable by human-scale effort.[7] Urban, Karaganis and Schofield note that:

> "For some OSPs, this automation increased the annual number of notices they received to hundreds of thousands or even millions of requests. Some OSPs responded by sacrificing human review of the vast majority of takedown requests and deploying their own automated processing methods to accomplish takedown more efficiently.[8]

---

[6] See, e.g., "Ernesto," "Google Asked to Remove 1 Million Pirate Links Per Day," TorrentFreak, Retrieved December 14, 2019 from https://torrentfreak.com/google-asked-to-remove-1-million-pirate-links-per-day-140820/.

[7] Id.; https://smallbiztrends.com/2015/05/fraudulent-dmca-takedown-requests.html

[8] Urban, Jennifer M. and Karaganis, Joe and Schofield, Brianna, Notice and Takedown in Everyday Practice, p. 114 (March 22, 2017) UC Berkeley Public Law Research Paper No.

There has also been a similar increase over time in the **rise of notice-sender intermediaries** — i.e., parties other than the copyright or other claimant, sending notices on the claimant's behalf.[9] As just a few examples of these, Remove Your Media LLC, on behalf of its clients, has asked Google to remove more URLs than all but two other companies -- nearly four million URLs-- and most of those since 2015.[10] Counterfeit Technology, a rights-management company, has sent over one million takedown notices to various OSPs, mostly Google, on behalf of a wide variety of rightsholders, all since 2015. [11] Web Sheriff, a company based in the United Kingdom providing intellectual property, copyright and privacy rights protection services, has sent over sixty-one thousand of the notices in Lumen, with approximately a third of those being sent in the last 12 months.

- Finally, one of the great benefits of a project like Lumen is the fact that providing individual notice data at a granular level allows Lumen users to check, on a link-by-link basis, whether a given notice was sent to vindicate a legitimate claim or was sent in error (whether inadvertently or deliberately). The database thus complements the sort of aggregate transparency reporting about takedowns in which many online service providers engage today. That said, developing a comprehensive and accurate sense of error rates at an aggregate level can be a challenge. Lumen itself is not in a position to analyze each of the billions of URLs in its database, either in real-time, or retrospectively. A number of researchers, however, have relied on Lumen data to examine notice errors and trends.

2755628. Available at SSRN: https://ssrn.com/abstract=2755628 or http://dx.doi.org/10.2139/ssrn.2755628.

[9] See, e.g. Id, p.114 ("an ever-escalating arms race fought with millions of automated notices and revolving offshore domains").

[10] Google Transparency Report, "Reporting Organization Page for Copyright Reporter 1504, Remove Your Media LLC," available at https://transparencyreport.google.com/copyright/reporters/1504

[11] The results of a search for all Lumen notices submitted to Lumen by the company Counterfeit Technology can be found at https://www.lumendatabase.org/notices/search?utf8=%E2%9C%93&sender_name_facet=Counterfeit.Technology

RESEARCH HIGHLIGHTS

Among relevant highlights of research involving Lumen data are the following examples:

- Along with other colleagues at the Takedown Project,[12] **Jennifer Urban** of Berkeley Law has been carrying out empirical and theoretical research into takedown regimes and their outcomes, relying in part on a dataset of millions of Lumen notices. In one 2016 paper — *Notice and Takedown In Everyday Practice* (2016)[13] — the authors found, among other conclusions, that 11.5 % of Google Image-related takedown notices possessed characteristics that weighed in favor of the content's qualifying for one of the legal exceptions to copyright infringement under United States law (such as educational uses). More than half of these were requests to remove material from news organization websites.

- Restricting his inquiry to United States court orders and associated documents, **Eugene Volokh** of University of California at Los Angeles School of Law has identified over 120 examples within a set of approximately 700 court orders containing evidence of some kind of false statement or fraud. Professor Volokh's work has already led to at least one investigation and lawsuit by US law enforcement.[14]

- Canadian law professor and academic **Jon Penney**, of Osgoode Hall Law School, University of York, Toronto (presently a Research Affiliate of Harvard's Berkman Klein Center for Internet & Society and visiting scholar at Harvard Law School), conducted empirically-grounded qualitative and quantitative research on takedown notices addressing the "chilling effect" that the existence and use of

[12] See https://www.thetakedownproject.org.

[13] Urban, Jennifer M. and Karaganis, Joe and Schofield, Brianna, Notice and Takedown in Everyday Practice (March 22, 2017). UC Berkeley Public Law Research Paper No. 2755628. Available at SSRN: https://ssrn.com/abstract=2755628 or http://dx.doi.org/10.2139/ssrn.2755628.

[14] Volokh, Eugene, "Texas AG's office accuses 'reputation management company' of procuring fraudulent libel takedown lawsuits," Washington Post, (September 12,2019) Available at https://www.washingtonpost.com/news/volokh-conspiracy/wp/2017/09/12/texas-ag-accuses-reputation-management-company-of-procuring-fraudulent-libel-takedown-lawsuits/

takedown enforcement regimes can have on the speech and online participation of various groups, including Twitter and Google Blog users.[15]

- In his 2013 paper, *Who Watches the Watchmen?" An Empirical Analysis of Errors in DMCA Takedown Notices*,[16] **Daniel Seng** of the National University of Singapore applied data parsing techniques to a dataset of half a million takedown notices and more than fifty million URL takedown requests sent to Google, up until 2012. His research found that 8.3% of all takedown notices sent did not fully comply with the statutory functional formalities, and that at least 1.3% of the takedown requests exhibited "substantive" errors.

- In their paper *Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown*,[17] **Sharon Bar-Ziv and Niva Elken-Korin** of Sapir Academic College, School of Law and the University of Haifa - Faculty of Law, respectively, systematically analyzed a large-scale random sample of approximately 10,000 removal requests sent to Google Search regarding allegedly infringing materials on .il websites. Coding the notices with the Takedown Project's coding engine[18] revealed that only 34% of of the DMCA notices actually raised copyright issues, while 66% pertained to other claims such as libel or privacy, which for the researchers raised serious concerns regarding the integrity of online copyright

---

[15] Penney, Jonathon, Privacy and Legal Automation: The DMCA as a Case Study (September 1, 2019). Stanford Technology Law Review, Vol. 22, No. 1, 412. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3504247

[16] Seng, Daniel Kiat Boon, 'Who Watches the Watchmen?' An Empirical Analysis of Errors in DMCA Takedown Notices (January 23, 2015). Available at SSRN: https://ssrn.com/abstract=2563202 or http://dx.doi.org/10.2139/ssrn.2563202. ("The most plausible (and most enlightening) explanation for this serious problem is that the owners and reporting agents are looking for infringing materials online by merely checking for the presence of certain search terms on third party sites . . . .")

[17] Bar-Ziv, Sharon and Elkin-Koren, Niva, Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown (July 15, 2018). Connecticut Law Review, Vol. 50, 2017. Available at SSRN: https://ssrn.com/abstract=3214214

[18]Urban, Jennifer M. and Karaganis, Joe and Schofield, Brianna, Notice and Takedown in Everyday Practice (March 22, 2017). UC Berkeley Public Law Research Paper No. 2755628. Available at SSRN: https://ssrn.com/abstract=2755628 or http://dx.doi.org/10.2139/ssrn.2755628.

enforcement, specifically that the removal and blocking of access to online materials took place without any legal oversight, allowing some notice senders to misuse the system to restrict the availability of content online.

CONCLUSION

One of the key reasons for Lumen's existence is a belief that good data informs good policy. Drawing on its experiences, and grateful for the opportunity to share what it has learned, Lumen urges those involved in implementation of Article 17 to make decisions based on data, and to ensure that any forward-looking implementation schema encourage data sharing. Those who send and receive takedown notices (on one hand) should be incentivized to share copies of those notices with researchers, journalists, policymakers, and the public at large (on the other hand). Mechanisms that provide for transparency will allow government and regulatory bodies to develop robust and effective policies to govern situations in which content and links are required to be removed, while facilitating the sort of scrutiny and oversight that ought to characterize any legal framework that impacts the flow of information online.