# A Content-Linking-Context Model for "Notice-and-take-down" Procedures

Pei Zhang
School of Electronics and Computer
Science
University of Southampton
Southampton, UK
pz2g12@soton.ac.uk

Sophie Stalla-Bourdillon School of Law University of Southampton Southampton, UK S.Stalla-Bourdillon@soton.ac.uk Lester Gilbert
School of Electronics and Computer
Science
University of Southampton
Southampton, UK
L.H.Gilbert@soton.ac.uk

## **ABSTRACT**

The US Digital Millennium Copyright Act (DMCA) of 1998 [1] adopted a notice-and-take-down procedure to help tackle alleged online infringements through online service providers' actions. The European Directive 2000/31/EC (e-Commerce Directive) [2] introduced similar liability exemptions, but did not specify any take-down procedure. Many intermediary (host, and online search engine) service providers even in Europe have followed this noticeand-take-down procedure to enable copyright owners to issue notices to take down allegedly infringing Web resources. However, the accuracy of take-down is not known, and notice receivers do not reveal clear information about how they check the legitimacy of these requests, about whether and how they check the lawfulness of allegedly infringing content, or what criteria they use for these actions. In this paper, we use Google's Transparency Report as the benchmark to investigate the information content of take-down notices and the accuracy of the resulting take-downs of allegedly infringing Web resources. The analysis of copyright infringement is limited to the five scenarios most frequently encountered in our study of Web resources. Based on our investigation, we propose a Content-Linking-Context (CLC) model of the criteria to be considered by intermediary service providers to achieve more accurate take-down.

#### **CCS Concepts**

General and reference  $\rightarrow$  Cross-computing tools and techniques  $\rightarrow$  Empirical studies

## **Keywords**

Copyright; Notice-and-take-down; CLC Model; Google Transparency Report.

# 1. INTRODUCTION

The emerging Web technologies and online services have brought new challenges to copyright enforcement on the Web. Internet intermediaries such as Internet access providers, content hosts and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

WebSci '16, May 22-25, Hannover, Germany

Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 978-1-4503-4208-7/16/05...\$15.00

DOI: http://dx.doi.org/10.1145/2908131.2908171

publishers, and link providers, play an important role in the distribution and communication of online content. They are subject to increasing obligations to monitor allegedly illegal activities undertaken through their platforms, despite the fact there is still a debate over whether, or to what extent, Internet intermediaries ought to have such duties imposed upon them [3]. The DMCA is the first statute to create limitations on the liability of Internet intermediaries on copyright infringement by imposing certain regulatory duties on them. It adopts a notice-and-take-down procedure for host providers and information location tools such as search engines. In European law, there is no equivalent harmonised procedure being discussed at the Commission level, although similar liability-exemption rules are set forth in the e-Commerce Directive (Articles 12 to 15). Some EU Member States have, however, adopted a notice-and-take-down procedure for copyright infringement [4].

The DMCA does not require intermediary service providers to check the allegedly infringing content to decide whether it is infringing. Instead, it only requires that the content be removed "expeditiously" if the notification substantially complies with Section 512(c)(3). This mechanism has been criticised by many legal researchers because of its major focus on copyright owners' interest and over-protection [5] [6] [7]. Under EU/UK law, it is still unclear whether intermediary service providers have to assess the lawfulness of the allegedly infringing content even in cases in which the allegedly infringing content is not manifestly infringing [8]

In practice, many intermediary service providers such as Google, Twitter and Dailymotion have followed notice-and-take-down procedures. Google has taken a step further to assess take-down requests so as to determine if an infringement has occurred. Because the notice-and-take down procedure implemented by Google for content available in Europe/UK is the same as the one implemented for content available in the US, and because the implementation of the notice-and-take-down procedure by Google has been directly triggered by adopting the DMCA, it makes sense to examine the procedure in the light of the DMCA to fully understand how it works in practice. In order to ensure the accuracy of take-down, it is also important to know the criteria used to examine the allegedly infringing Web resources and the workflow for using such criteria.

Based on a literature review of legal materials and analysis of current practices, this paper presents a Content-Linking-Context (CLC) Model for copyright related criteria used in assessing content/webpages which are requested to be removed in notices. There are three main components defined in the model. Content is a set of criteria used to compare the similarity between the allegedly infringing work and the original copyright work. Linking is a set of

criteria to assess through what method the allegedly infringing work is accessible on a website. Context is a set of criteria to illustrate the likelihood a website contains allegedly infringing works. This model intends to support decision making processes triggered by notifications and implemented by online link providers such as search engine providers and index service providers. It could also be used by anti-piracy service providers such as Muso, Degban, and AudioLock.Net in order to help them filter allegedly infringing websites when they send out take-down notices.

# 2. ANALYSIS OF CURRENT PRACTICES

#### 2.1 State of Claimed Web Resources

To understand more thoroughly the notices and the reported infringing web resources, we analysed the Google Transparency Report, specifically the "request by copyright owners to remove search results", since this report is openly available and provides comprehensive information in respect of webpages associated with potentially infringing content.

According to Google's Transparency Report, 831,185 notices containing over 300 million URLs (used to locate the allegedly infringing content) were received in 2014 in relation to Google Search. Figure 1 shows an example of the copyright claims in a notice. We can see that copyright owners can make several "claims" which contain information about the title, type, and description of the copyright work, original URL, and allegedly infringing URLs.

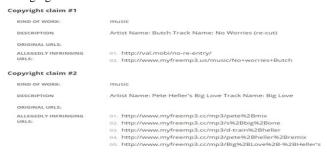


Figure 1. Copyright claims in each notices sent through Web

We chose one month's notices received by Google dated from September to October 2014. The reason we chose this time period is that our experiment started around the beginning of October 2014, and the latest notice data we could get at that moment was dated from September. For each day, we picked up the first notice received in every hour. And in every notice, two URLs from the first and second claims were selected to make sure the URLs were chosen randomly. In total, 730 URLs were obtained. Among the 730 URLs, 202 pages were not found (IP restriction, 404 error¹, etc.). The following analysis is based on the 528 pages retrieved.

The URLs point to various types of copyright work. Figure 2 shows the different types of copyright works that were claimed to have been infringed and their percentage in the total of the URLs examined. We can see that Music/Audio represents the largest proportion of alleged copyright infringing work on the Web. Many websites offer online play functions and supply links for downloading. These music works can be streamed online or downloaded through file sharing websites. At the same time, over half of notices were sent by the right holders in the music industry.

There are five types of website which can broadly be said to participate in infringement activities. They are online streaming websites, online reading websites, One-Click Hosters [9], index websites, and P2P communities. Online streaming websites enable content, including music/audio and film/video, to be played online. The source could be hosted by the website itself or be embedded from a different host. The second type of website, online reading websites, applies only to books. Books are displayed in text or image format which allows users to read online freely. The third type is One-Click Hoster sites, such as zippyshare<sup>2</sup>, which allows users to upload large files and exchange them by sending corresponding download links to intended recipients of the files. The fourth type is websites offering index services. This type of website searches for content online and indexes corresponding downloadable links. It usually indexes links to different One-Click Hosters. The last type is P2P communities. P2P communities usually supply peer-to-peer download services. The most common P2P services are hosting .torrent files, supplying an index of .torrent files, and running bit torrent tracker servers. Figure 3 shows the percentage of different types of reported infringing websites.

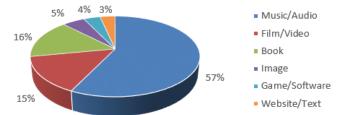


Figure 2. Type of copyright work that claimed to be infringed

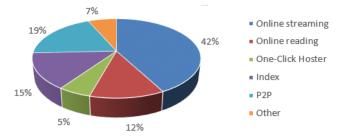


Figure 3. Different types of infringing websites

# 2.2 Discussion of Google's Practice on Noticeand-take-down

Google receives a large number of copyright notices every day. Google assesses these notices and the associated URLs to decide whether to remove them. Google releases only simple information about how it assesses take-down requests [10]. One fact known is that Google has adopted a Trusted Copyright Removal Program (TCRP) to help with these assessments. Notice senders who participate in TCRP are believed to be "reliable high accuracy submitters", compared to "non-sophisticated submitters" who issue many "incomplete or abusive" notices [11]. The exact details of the program and how it operates are, however, relatively secret [12]. Seng believes the program is an automated method that allows notice senders to submit large numbers of take-down requests to Google, which Google processes rapidly [13]. No detailed information has been published either about the criteria considered in the decision making process or about how the lawfulness of the content is checked.

<sup>&</sup>lt;sup>1</sup> http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html

<sup>&</sup>lt;sup>2</sup> http://www.zippyshare.com/

A reasonable assumption is that domain-driven analysis plays an important role in the take-down process implemented by Google. From the Google Transparency Report and its website, we can see that Google has been doing extensive data analysis on domain names<sup>3</sup>. The Transparency Report website lists the number of URLs that were reported under the same domain name during a time period, the number of URLs that were already removed under the same domain name, and the number of notice-senders who reported the same domain although they had reported different URLs etc. As a result, the decision to take down is more likely to be according to a top-level domain name suspicion instead of an assessment of the exact content for each URL. Taking the domain vmusice.net<sup>4</sup> as an example, between 8th August 2012 and 8th February 2015, Google received 40,372 notices containing 3,236,150 URLs under this domain. Because vmusice.net is a top domain specified, Google's automated program has a high takedown rate of URLs under that domain. The extent to which Google goes further to assess the exact content under each single URL is still unknown. Technically, it is much easier for a system to just compare domains instead of the actual content in the webpages that

From a legal point of view, this method is relatively safe and it follows, to some extent, the practice defined in Section 512(g)(1) DMCA, which indicates that a service provider will not be liable for infringement if the taking down action is based on the "good faith" disabling of access to material that is claimed to be infringing. So if a domain is highly suspected of containing infringing content, intermediary service providers will be acting in "good faith" by removing any URLs under that domain without needing to examine every reported URL.

It is arguable whether the domain-driven method is sufficient to ensure reasonable take-down accuracy. Under EU/UK law, there is not a good-Samaritan exemption, which would mean that accuracy is a significant issue.

# 2.3 Linking Issues on the Web

Linking issues on the Web have triggered a heated debate for legal professionals. An early paper by Deveci [14] defined different types of links and raised some copyright issues associated with linking.

In the US case *Perfect 10, Inc. v. Google Inc* [15], the Ninth Circuit agreed that hyperlinks and framing are not infringing copyright since Google could not "supervise or control" the third-party websites linked to from its search results. It is arguable, however, whether Google would still not be liable assuming Perfect 10 had given Google actual knowledge of specific infringements (e.g. specific URLs for infringing images).

In the recent *Nils Svensson and Others v Retriever Sverige AB* case [16], an interesting question was raised as to whether hyperlinks are covered by the right to communicate works to the public [17]. The CJEU holds that hyperlinks to protected works which are already freely available online do not infringe copyright. In another case, *BestWater International GmbH v Michael Mebes and Stefan Potsch* [18], the CJEU holds that embedded linking from another freely available website does not constitute an infringement of the right of communication if the work concerned is neither directed at a new public nor communicated by using specific technical means different from that used for the initial communication [19]. We

define two types of link in this paper depending on where the copyright work is located.

- Simple link. A clickable hyperlink in HTML which leads visitors to another web resource.
- Embedded link. Many webpages contain links to content in other domains. A webpage can embed content from another domain by using a HTML tag. For example, the page "http://example.org/index.html" contains an audio file from http://test.org by using the link <audio src="http://test.org/music1.mp3">. In this example, the audio file from test.org is directly embedded in the webpage "index.html" on example.org and users are not explicitly notified that the music is from another domain. Content can also be embedded using an <iframe> tag. Specifying a URL using the "src" attribute in the <iframe> tag will direct the browser to fetch the webpage the URL points to and display it in the current webpage. Similarly, users are not explicitly notified that the content is from another domain. This method of embedding is also called "framing".

# 3. CONTENT-LINKING-CONTEXT (CLC) MODEL

# 3.1 Methodology

The objective of the research is to build a Content-Linking-Context Model for analysing copyright infringement on webpages in order to help make better decisions on take-down. To build this model we have followed a three-step methodology.

**Step one**: We undertook a literature review of legal materials from different jurisdictions and current practices in order to identify consensual infringement and non-infringement scenarios. Based on this literature review, we constructed five scenarios as listed below: four infringement scenarios and one non-infringement scenario. In order to construct these five scenarios we adopted a conservative view of copyright laws. A conservative view (for a US example see [5]) was needed to address uncertainties and simplify the analysis. More precisely, we adopted a broad definition of exclusive rights and in particular given the persistence of uncertainties in the field we assumed that even if an act could be considered as being outside the scope of copyright owners' exclusive rights (such as the right to communicate the work to the public), actual knowledge of the presence of infringing material on its system or network on the part of the online service provider (excluding mere conduits) would trigger liability, be it on the ground of copyright liability theories or other liability theories. In addition, we excluded transformative uses of copyright works from our analysis and assumed that partial reproductions of copyright works always amounted to a taking of the originality of the copyright works.

- a. Hosting an exact copy of a copyright work without authorization. In this scenario, the website operator hosts the copyright work without the permission of the copyright owner, and usually puts it in the domain of their website for viewing or downloading. We thus assume there is an infringement in this case.
- b. Hosting a partial copy of a copyright work without authorization. We define a partial copy of a work as a section of the copyright work which does not have any further additions, and which is a substantial copy. We thus assume there is an infringement in this case.

-

<sup>&</sup>lt;sup>3</sup> http://www.google.com/transparencyreport/removals/copyright/d omains/?r=all-time

<sup>&</sup>lt;sup>4</sup> http://www.google.com/transparencyreport/removals/copyright/d omains/vmusice.net/, data is captured on 8<sup>th</sup> February 2015

- c. Supplying links (simple or embedded) to an exact copy of a copyright work where the copy is unauthorized. In this scenario, the website operator provides links for users to view/download unlawful content, and the operator has actual knowledge of the fact that the link is to unlawful content. We thus assume there is an infringement in this case.
- d. Supplying links (simple or embedded) to a partial copy of an unlawful work. This scenario is similar to scenario c, however, instead of giving access to an exact full copy, users are only able to view part of the unauthorized copy. We thus assume there is an infringement in this case.
- Supplying links (simple or embedded) to work made publicly available by the copyright owner. We assume there is no infringement.

**Step two:** In order to investigate whether the most encountered scenarios in practice are covered by the scenarios listed in Step one, we examined the notices in relation to the formats and patterns of reported infringing webpages. From Figure 3, 34% allegedly infringing websites stream content and 10% provide reading content hosted locally or embedded from external websites. All five scenarios created in Step one refer to this type of webpage. Twenty three percent of allegedly infringing websites are host providers, and scenarios a and b refer to this type of website, while 12% offer index services (scenarios c, d and e) and 16% provide peer-to-peer content (all five scenarios).

**Step three**: We derived 3 categories of criteria to be considered in order to determine whether there was an infringement in each of these scenarios and ultimately whether a take-down action would be legitimate. The categorization of content, linking, and context was based on whether the criteria of copyright infringement referred to the website content, the links to it, or the metadata context of the content and the website.

# 3.2 CLC Model

Our model was limited in the following ways:

- The model uses the two types of links aforementioned: simple and embedded
- 2. The model deals with the five scenarios identified earlier.
- 3. Only music work is considered in the CLC Model as a starting point, because allegedly infringing music represents the largest proportion of removal requests on the Web (57% in Figure 2).
- 4. We consider that the principle of exhaustion does not apply to the supply of works online for music. There might be some exceptions in certain systems with regard to certain types of work such as software in the European Union [20], but we assume this is not the case for music. We will therefore not attempt to capture and represent the principle of exhaustion in our CLC model.
- Although the accuracy of Google's domain-driven method needs further discussion, it does reflect the level of suspicion of an URL. We use it as a factor to indicate the likelihood of infringement in CLC model.

A Content-Linking-Context Model which contains 10 criteria (C1 to C10) is proposed to indicate different factors we have considered when verifying allegedly infringing web resources in a notice. The model is explained below.

• Content. Allegedly infringing content on the webpage to which a URL points needs to be compared with the original copyright work in order to decide on the similarity between them. Criteria C1 and C2 indicate whether the reported content exists on the webpage, and C3 indicates how much the reported content is similar to the original work (by audio comparison).

- **C1: URL accessibility.** Whether the web resource identified by the URL is still accessible. It is possible that the URL is no longer valid.
- **C2:** Content existence and accessibility. Whether the allegedly infringing content is accessible without special authorization, such as login and IP restrictions. In the case of HTTP redirection, the final redirected page is examined.
- **C3:** Work (Audio) comparison. If a copy of the work is accessed, its similarity to the original work, whether in whole or part. Both the alleged infringing file and the original copyright music file are used for comparison. There are some technical libraries and open source tools available to compare the two files and give a percentage on how much they match each other.
- Linking. Allegedly infringing content could be directly accessed (and played) on the webpage (C4) or downloadable by users (C5). Criteria C6 and C7 reflect the requirement that the types of link need to be examined in order to reveal the ownership of the content and whether the source is authorized.
  - **C4: Online streaming.** For music, whether the website offers an online-playing function.
  - **C5: Downloadable.** Whether the content can be downloaded directly.
  - **C6:** Link type of online streaming resources. When an online streaming function is offered, whether the resource is hosted on the current domain, or is embedded from another domain.
  - **C7:** Link type of downloadable resources. When a download function is offered, whether the resource is hosted on the current domain, or is linked from another domain for download.
- Context. While criteria in Content and Linking can in theory lead to a clear decision of copyright infringement on the Web, in practical instances, however, it may not be so clear. For example, the allegedly infringing music cannot be downloaded or be listened to online when the webpage is viewed (for technical reasons, e.g. temporary broken links), but the decision of taking down by notice receivers still needs to be made. In this case, "Context" information such as whether metadata (C8, C9) of the content appears in the webpage, and whether the current website is highly suspected to contain copyright infringement work (C10), will be used in the decision making process. In addition, if the allegedly infringing content is embedded from/linked to other external website instead of being hosted on the current reported one, C10 assesses whether the original domain is suspected to contain unlawful content.
  - **C8: Title of copyright work.** Information about the title of the music
  - **C9: Performer of the copyright work.** Information about the person who performed in the music.
  - **C10: URL suspicion.** Google Transparency Report data of URLs that have been claimed to have infringed content is compared to the current URL domain name to find out how many claims have been made under that domain name. This criteria reflects the level of suspicion of a URL.

Figure 4 illustrates the classes and their associations in the CLC model. The Request class represents a removal request and each Request contains one to many WebResources indicated by URLs. The Context consists of criteria about the metadata matching and URL suspicion. The Content class can be either a HostedContent or LinkedContent. LinkedContent means even though the content is displayed within the current WebResource, the content is fetched from a URL other than the URL representing the current WebResource. The TypeOfDelivery class means the content can be delivered by OnlineStreaming, or Downloadable. The LinkedContent will associate with an instance of the Linking class. Depending on the type of the linking, a Linking instance can be one

of SimpleLink or EmbeddedLink. Compared with LinkedContent, HostedContent indicates the content delivered is hosted on the current WebResource's URL.

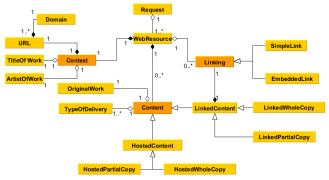


Figure 4. Static Content-Linking-Context conceptual design

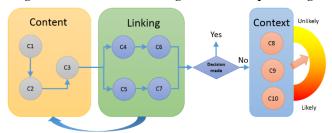


Figure 5. Dynamic Content-Linking-Context illustration

Figure 5 illustrates a dynamic workflow using the CLC Model. If positive answers have been given to C1 and C2 when a removal request is made, allegedly infringing content is compared with original content (C3). At the same time, the Linking criteria identify how the content is displayed (C4, C5) and where the content source is located (C6, C7), so as to further answer the questions of how likely there is a copyright infringement and eventually whether to take it down. In some circumstances, there is no clear answer to copyright infringement by analysing Content and Linking criteria, so C8 to C10 are checked to facilitate any decision on infringement.

# 4. CONCLUSION AND FUTURE WORK

How to reform the notice-and-take-down procedure is hotly debated by legal professionals. Applying proper criteria to assess Web resource in removal requests in order to support notice receivers' decision making process is essential to improve the procedure. We designed a CLC Model to represent 10 criteria and indicate how these criteria operate for the analysis of allegedly infringing Web resources.

The purpose of CLC Model is to help verify copyright infringing activity on webpages, preferably in an automatic manner. Obviously, strictly speaking only judges are well placed to make a decision on the lawfulness of available Web resources. In consequence, the output of the CLC Model will be a score to indicate a likelihood of infringement with a view of supporting the decision making process and not replacing it. The next step for our future work, is to design a decision tree and an algorithm according to the CLC Model.

# 5. REFERENCES

- [1] Digital Millennium Copyright Act, H.R. 2281, 105th Congress. 1998.
- [2] Directive 2000/31/EC of the European Parliament and of the Council on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (June 8, 2000), OJ L 178, 17.7.2000, p. 1–16.

- [3] S. Stalla-Bourdillon, "Online monitoring, filtering, blocking... what is the difference? Where to draw the line?," in International Association of IT Lawyers, Copenhagen, DK: International Association of IT Lawyers, 2012.
- [4] "First Report on the application of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market," 2003.
- [5] J. M. Urban and L. Quilter, "Efficient Process or Chilling Effects-Takedown Notices under Section 512 of the Digital Millennium Copyright Act," St. Cl. Comput. High Tech. LJ, vol. 22, p. 621, 2005.
- [6] J. H. Reichman, G. B. Dinwoodie, and P. Samuelson, "Reverse Notice and Takedown Regime to Enable Pubic Interest Uses of Technically Protected Copyrighted Works, A," Berkeley Tech. LJ, vol. 22, p. 981, 2007.
- [7] J. Cobia, "Digital Millennium Copyright Act Takedown Notice Procedure: Misuses, Abuses, and Shortcomings of the Process, The," Minn. JL Sci. Tech., vol. 10, p. 387, 2008.
- [8] S. Stalla-Bourdillon, "Sometimes one is not enough! Securing freedom of expression, encouraging private regulation, or subsidizing Internet intermediaries or all three at the same time: the dilemma of Internet intermediaries' liability," J. Int. Commer. Law Technol., vol. 7, no. 2, 2012.
- [9] T. Lauinger, M. Szydlowski, K. Onarlioglu, G. Wondracek, E. Kirda, and C. Kruegel, "Clickonomics: Determining the Effect of Anti-Piracy Measures for One-Click Hosting.," in NDSS, 2013.
- [10] Google, "How Google Fights Piracy," 2013.
- [11] R. Tushnet, "PTO/NTIA: notice and takedown-Improving the Operation of the Notice and Takedown System," 2013. [Online]. Available: http://tushnet.blogspot.co.uk/2013/12/ptontia-notice-and-takedown.html.
- [12] M. Leiser, "The copyright issue and censorship threat buried within Google's transparency report," 2013. [Online]. Available: http://www.thedrum.com/news/2013/12/23/copyright-issue-and-censorship-threat-buried-within-googles-transparency-report. [Accessed: 25-Feb-2015].
- [13] D. Seng, "The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices," Virginia J. Law Technol. Forthcom., 2014.
- [14] H. A. Deveci, "Hyperlinks Oscillating at the Crossroads," CTLR-OXFORD-, vol. 10, no. 4, pp. 82–94, 2004.
- [15] Perfect 10, Inc. v. Amazon.com, Inc. and A9.com Inc. and Google Inc. 508 F.3d 1146 (9th Cir. 2007). 2006.
- [16] CJEU C-466/12 Nils Svensson et al v Retriever Sverige AB, 13 February 2014 ECLI:EU:C:2014:76.
- [17] E. Arezzo, "Hyperlinks and Making Available Right in the European Union--What Future for the Internet After Svensson?," IIC-International Rev. Intellect. Prop. Compet. Law, vol. 45, no. 5, pp. 524–555, 2014.
- [18] CJEU C-348/13 BestWater International GmbH v Michael Mebes and Stefan Potsch of 21 October 2014 ECLI:EU:C:2014:2315..
- [19] E. Rosati and O. Löffel, "That BestWater order: it's up to the rightholders to monitor online use of their works," 2014. [Online]. Available: http://ipkitten.blogspot.fr/2014/10/thatbestwater-order-its-up-to.html. [Accessed: 12-Dec-2015].
- [20] CJEU C-128/11 Usedsoft GmbH v Oracle International Corp, 3 July 2012 ECLI:EU:C:2012:407.