UNIVERSITY OF SOUTHAMPTON
Faculty of Physical Sciences and Engineering
School of Electronics and Computer Science

A Content-Linking-Context Model and Automatic Copyright Verification in the Notice-and-take-down Procedures

By Pei Zhang

November 2017

A thesis submitted for the degree of Doctor of Philosophy

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

A Content-Linking-Context Model and Automatic Copyright Verification in the Notice-and-take-down Procedures

by Pei Zhang

The US Digital Millennium Copyright Act (DMCA) of 1998 adopted a notice-and-take-down procedure to help tackle alleged online infringements through online service providers' actions. European Directive 2000/31/EC (e-Commerce Directive) introduced a set of liability exemptions similar to the one found in the DMCA, but did not specify any take-down procedure. Many intermediary (hosts and online search engines) service providers, even in Europe, have followed this notice-and-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, whether and how they verify the lawfulness of allegedly infringing content, and what criteria they use for these actions. Google's Transparency Report is used as the benchmark to investigate the information content of take-down notices and to assess the accuracy of the resulting take-downs of allegedly infringing Web resources. Based on the investigation, a Content-Linking-Context (CLC) Model which identified the criteria to be considered by intermediary service providers to achieve more accurate take-down is proposed. The technical issues by applying the CLC Model to an automation system to automatically assess Web resources and produce a series of analytic results and, eventually, a 'likelihood of infringement' score are investigated. The CLC Model is validated by experienced copyright experts, all of whom have a good level of agreement regarding the usage of the criterion and the infringement score generated in the CLC Model. The automation system is evaluated by users and the results confirm that, for specific types of Web resources, the system helps to bring users' decisions closer to those of the experts.

Table of Contents

List of Figures

List of Tables

Declaration of Authorship

I, Pei Zhang, declare that this thesis entitled

A Content-Linking-Context Model and Automatic Copyright Verification in the Notice-and-take-down Procedures

and the work presented in it are my own and has been generated by me as the result of my own original research. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Part of this work have been published as:
   Zhang, Pei, Stalla-Bourdillon, Sophie and Gilbert, Lester (2016) A content-linking-context model for "notice-and-takedown" procedures. Web Science 2016 Proceedings of the 8th ACM Conference on Web Science, Germany. 22 - 25 May 2016. , pp. 161-165. (doi:10.1145/2908131.2908171)

Signed: ………………………………………………………………………………………

Date: …………………………………………………………………………………………

Acknowledgements

I would like to thank my supervisors Lester Gilbert and Sophie Stalla-Bourdillon for their support, guidance and suggestion towards the completion of this PhD thesis.

I am grateful to the participants who generously gave their time to take part in my experiment.

Last but not least, I would like to thank my parents and husband who constantly encourage me and thank my lovely son who is always there to bring happiness to me.

Nomenclature

Ajax	Asynchronous JavaScript and XML
API	Application Program Interface
CJEU	Court of Justice of the European Union
CLC Model	Content-Linking-Context Model
CSSOM	CSS Object Model
DMCA	Digital Millennium Copyright Act
DNS	Domain Name System
DOM	Document Object Model
DRM	Digital Rights Management
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
ISP	Internet Service Provider
OSP	Online Service Provider
P2P	Peer to Peer
TCRP	Trusted Copyright Removal Program
UGC	User Generated Content
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium

Glossary of the CLC Model Criteria

Criteria 1 (C1)	URL accessibility
Criteria 2 (C2)	Content existence
Criteria 3 (C3)	Work (Audio) comparison
Criteria 4 (C4)	Online access
Criteria 5 (C5)	Online playable
Criteria 6 (C6)	Download access
Criteria 7 (C7)	Downloadable
Criteria 8 (C8)	Link type of online accessing resources
Criteria 9 (C9)	Link type of downloadable resources
Criteria 10 (C10)	Title of copyright work
Criteria 11 (C11)	Performer of copyright work
Criteria 12 (C12)	URL suspicion

Glossary of URL Types

URL Type 1 Neither the metadata of the work such as title, author, publication time etc. are found on the webpage, nor the actual copyright work are found on the webpage

URL Type 2 The metadata of the work such as title, author, publication time etc. are found on the webpage, but the webpage does not supply any interface for users to get access to the content, so the actual work cannot be accessed by users

URL Type 3 The metadata of the work such as title, author, publication time etc. are found on the webpage, the webpage offer access interface for users but the content is not accessible

URL Type 4 The metadata of the work such as title, author, publication time etc. is found on the webpage, the work is hosted under the current webpage and it is accessible

URL Type 5 The metadata of the work such as title, author, publication time etc. is found on the webpage, the work is linked from other website and is accessible through the current webpage

Chapter 1 Introduction

1.1 Research Motivation and Purpose

Emerging Web technologies and online services have brought new challenges for copyright enforcement on the Web (Elkin-Koren 2014). As many researchers have commented, copyright issues on the Internet is "the most inflamed issue in current intellectual property" (Cornish, Llewelyn, and Aplin 2010). This research focuses on copyright verification in the context of notice-and-take-down processes. Although an increasing amount of data has recently been published by notice receivers about their notice-and-take-down practices, there is a lack of systematic analysis of notice-and-take-down lifecycles, such as how notice receivers check the lawfulness of allegedly infringing content, what techniques they use to make their decisions to take down content and what is the degree of take-down accuracy. Analysing and summarising these data to understand the patterns and characteristics of allegedly infringing Web resources is crucial to improve notice receivers' assessment processes as well as take-down accuracy. This research thus aims to contribute to this analysis.

Internet intermediaries such as Internet access providers, content hosts, 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 that there remains a debate regarding whether, or to what extent, Internet intermediaries ought to have such duties imposed upon them (Stalla-Bourdillon 2012a). The DMCA is the first statute to create limitations on the liability of Internet intermediaries specifically for copyright infringement by imposing certain regulatory duties on them. It adopts a notice-and-take-down procedure for host providers and to a lesser extent information location tools such as search engines. It requires them to perform several take-down steps when they receive removal notices. In European law, there is no equivalent harmonised procedure, although similar liability-exemption rules have been set out in the e-Commerce Directive (Articles 12 to 15) and the need to harmonise the notice-and-action has been discussed and debated by the EU legislature, including the Commission (Kuczerawy 2015). Some EU Member States, however, have adopted a notice-and-take-down procedure for copyright infringement (*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).

The DMCA does not require intermediary service providers to check the allegedly infringing content before making a decision as to 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 (Urban & Quilter 2005; Reichman et al. 2007; Cobia 2008). 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 whereby the allegedly infringing content is not manifestly infringing (Stalla-Bourdillon 2012b).

The DMCA has been selected as the theoretical background of the study because de facto it operates as a benchmark for online service providers. In any case, the prescriptions regarding the content of the notifications appear compatible with EU law and in particular the e-Commerce Directive. To be sure, the purpose of this thesis is not to solve private international law issues. It starts from the assumption that it is possible to identify internationally consensual infringement scenarios and aim to derive key criteria to describe these scenarios.

In practice, many intermediary service providers such as Google, Twitter and Dailymotion have implemented notice-and-take-down procedures. However, in this research, only Google Transparency Report on Copyright is considered because only Google has published a report including raw data and detailed information such as what the specific allegedly infringing content is, who the notice senders are and what the specific allegedly infringing URLs are. Google also assesses take-down requests so as to determine if an infringement has occurred. At this point it is fitting to point out that the notice-and-take down procedure implemented by Google for content available in Europe/UK is the same as that implemented for content available in the US. With this in mind, and because the implementation of the notice-and-take-down procedure by Google has been directly triggered by the adoption of the DMCA, it makes sense to examine the procedure with reference to the DMCA to fully understand how it works in practice. In order to ensure the accuracy of take-down, it is important to know the appropriate criteria used to examine the allegedly infringing Web resources and the workflow for using such criteria.

According to the study of Google's practice in the notice-and-take-down procedure, requests are made for hundreds of thousands of webpages to be viewed and examined for takedown every day by Google Search. In the face of such a large number of take-down requests, more automatic and evolving mechanisms are needed to support the decision-making process. Indeed, of the upmost importance is how to develop an automatic system which can efficiently extract feature values from so many webpages and to generate analytic results in view of supporting the online service providers' assessment process.

Through a literature review and preliminary study of Google's practice, this research aims to:

- Propose a model for copyright-related criteria used in assessing content/webpages which are requested to be removed in take-down notices.
- Develop an automatic system to dynamically apply the model to support online service providers in assessing allegedly infringing Web resources.

1.2 Research Questions and Contributions

According to the research purpose, the research questions are addressed as follows:

- Question 1: What is an appropriate model that can be used to assess allegedly infringing content on webpages in the notice-and-take-down procedure?
    - Sub-question 1.1: What is the current state of allegedly infringing Web content and notice-and-take-down practice?
    - Sub-question 1.2: How can the model be developed?
        - Sub-question 1.2.1: What criteria should be considered in the model to assess whether a webpage contains copyright infringement content?
        - Sub-question 1.2.2: What is the workflow of the criteria in the model?
    - Sub-question 1.3: Is the model valid?
- Question 2: What is an appropriate automation system for applying the model to automatically assess allegedly infringing content on webpages in the notice-and-take-down procedure?
    - Sub-question 2.1: To what degree can the model be automatically implemented in the system?
    - Sub-question 2.2: How good is the automation system at supporting the assessment?

The present research seeks to make three main contributions to the field, all of which are summarised below:

- Conduct an empirical study to understand the patterns of copyright infringement on webpages and investigate the take-down accuracy of online service providers.
- Propose and build a Content-Linking-Context (CLC) Model which is composed of 12 copyright-related criteria and the workflow of these criteria in order to assess allegedly infringing Web resources.
- Explore the possibility of automation of the CLC Model and develop an automation system to support online service providers in assessing allegedly infringing Web resources.

1.3 Structure of the Thesis

In the following sections of this thesis, Chapter 2 provides the background of the current copyright issues on the Web, including current Web applications used for perpetrating online infringement, Internet Intermediaries' liability, DMCA notice-and-take-down procedure and Google's practice, and current copyright enforcement technologies. Chapter 3 discusses a preliminary study in order to more thoroughly understand the notices and the patterns of reported infringing Web resources whilst also investigating the take-down accuracy of Google. Based on the literature review and preliminary study, Chapter 4 introduces a Content-Linking-Context (CLC) Model to analyse the existence of copyright infringing content on webpages. Chapter 5 validates the CLC Model through an expert validation experiment. Chapter 6 discusses the Web technologies used to automate each criterion in the CLC Model and analyses the development of an automation system to support the infringment assessment process. Chapter 7 introduces the user-based evaluation of the automation system. Chapter 8 summarises how the research questions are answered during the study, discusses the research challenges and implications, concludes the research and outlines future work.

Chapter 2 Literature Review

The fast development of the Internet and Web technologies have enabled broader distribution and communication of useful information to others. At the same time, it also have brought new challenges to copyright enforcement through the Internet. Internet intermediaries such as Internet access providers, content hosts, and link providers, play an important role in the distribution and communication of online content. It is still a debatable question whether some liability or regulatory duties should be imposed on Internet intermediaries. The Digital Millennium Copyright Act (DMCA) of 1998 (*Digital Millennium Copyright Act, H.R. 2281, 105th Congress* 1998) created limitations on the liability of Internet intermediaries for copyright infringement and adopted a "notice and take-down" procedure for certain types of online service providers such as host providers and to a more limited extent information location tools. They are also exploring and implementing different technologies to apply this procedure. In order to have a comprehensive understanding of the copyright infringement lifecycle in the context of notice-and-take-down, in this chapter, firstly copyright concerns in the context of the Web, particularly the linking issues, are discussed and, secondly the DMCA notice-and-take-down procedure and Google's practice are reviewed. Finally, technologies to detect copyright infringement and automatic copyright enforcement systems are discussed.

2.1 Copyright in the Context of Web

Broadly speaking, copyright in the sense of author right deals with the rights of creators in respect of their original works. Authors are granted a bundle of rights including economic rights and moral rights. Economic rights protect authors' economic interests and can be assigned or licensed against payment of authors. Moral rights are awarded to authors to protect extra-patrimonial interests such as the ability to be recognised as the true author of the work, or the ability to control subsequent modifications of the work and prevent derogatory treatments (Cornish, Llewelyn, and Aplin 2010). In this sense, authors are as a matter of principle first owners.

The Berne Convention achieved an agreement on copyright, recognized and acted as minimum standards on an international level (Ricketson 1987) (Ricketson and Ginsburg 2005). Under the Berne Convention, the term "protected works" covers:

- Literary works such as novels, poems, plays, reference works, newspapers, and computer programs.
- Films, musical works, and choreography.
- Artistic works such as paintings, drawings, photographs, and sculpture.
- Architecture.
- Advertisements, maps, and technical drawings.

Economic rights include: a right of translation (Article 8), a right of reproduction (Article 9), a right of public performance (Article 11, 14), a right of communication (Article 11, 11bis, 11ter, 14), a right of distribution (Article 14), a right of broadcasting (Article 11bis), and a right of adaptation (Article 12). Moral rights (Article 6bis) include a right of attribution (to claim authorship), and the right of integrity (to object to certain modifications and other derogatory actions).

2.1.1 Copyright infringement on the Web

Copyright concerns have existed for as long as the means to make copies have (Wallach 2001). Copyright was first developed in the sixteenth and seventeenth centuries to facilitate the production and distribution of printed text. Later, the use of magnetic tape made recording and re-recording of audio/video possible and the introduction of digital media such as CDs and DVDs made copying easier. At that time, the high costs of reproducing products operated as a deterrent to copyright infringement (Murray 2010). Since the 1990s, the fast development of computer and Internet technologies brings huge challenges to traditional principles of copyright. Copyright issues on the Internet have been referred to as "the most inflamed issue in current intellectual property" (Cornish, Llewelyn, and Aplin 2010). Books, music, films, etc., are frequently digitised and can be easily transmitted through a broad range of channels. As Rowland et al.(Rowland, Kohl, and Charlesworth 2010) commented, "copying material from the vast information source that is the Internet is a trivial matter. Similarly, the technology also makes it a trivial matter to make existing copyright works available on the Internet."

Infringement is widespread on the Web because the Web is open, unmanaged, and essentially the domain of 'free' goods (Hargreaves 2011). Various Web applications enable large-scale distribution and sharing of digital copyright works. Google and Performing Right Society (PRS) jointly published a report in 2012 to investigate websites that were thought, by major rights holders, to be significantly facilitating copyright infringement. Based on the type and operation of these sites, six major copyright infringement business models were identified in their study. These were: Live TV Gateway, Peer to Peer (P2P) community, Subscription Community, Music Transaction, Rewarded Freemium, and Embedded Streaming (Google and PRS for Music 2012).

Live TV Gateway predominantly offer links to streams of live free-to-air and pay TV. These sites also provide links to download games and eBooks, as well as other content in lower proportions. The content is centrally hosted in a different location from the site. P2P community facilitates downloading of content via P2P or distributed servers. These websites have facilitated a great number of online infringements (Sung and Huang 2014) by using P2P file sharing software. Famously, Napster (Lemley and Reese 2004), which was of the first generation of P2P systems developed for sharing music on the Web, was sued for copyright infringement under reproduction and distribution rights of the copyright protected works (A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004 2001). Today, although P2P file sharing has been extensively developed and no longer requires a central server to maintain the database index or perform user management, it still needs a 'seed' file to track the source file in order to set up the downloading connection. Websites such as thepiratebay.se centrally supply an index of such 'seed' sources (BitTorrent files) for users to complete the P2P transmission (Bridy 2011). Large amounts of content indexed on such websites are believed to be copyright works. Subscription Community and Music Transaction websites are very similar, where users pays a subscription fee for a range of content types or buy music, games and eBooks to download from the site's own servers. These types of websites, which illegally offer copyright works for download or for streaming, are also a major source of copyright infringing materials (Feiler 2012). Rewarded Freemium service enable users to provide content for others and these users are rewarded for their contribution. For example, some websites provide free online storage and upload services. They encourage users to upload more content by awarding them more free storage space and faster download speed. Embedding Streaming service provides hosting where users can upload content, and where others can stream the content from. They allows users to embed content on their sites and on third party sites. Web 2.0 dominant applications such as Wikipedia, YouTube, and Facebook focus more on user-interaction services which allow User Generated Content (UGC) to be widely shared. These platforms have become a popular forum for exchanging audio, photos, video and other UGC. By one estimate, over 65,000 videos are uploaded to YouTube and 100 million videos are viewed daily (Latham, Butzer, and Brown 2008). While these platforms can help small businesses or artists to publish their work with low cost, users often upload, reproduce or share infringing content.

Content on the Web can be communicated or distributed through linking. And the link is an essential concept on the Web. As the aforementioned infringing websites and services, they do not host the content and they supply links which enable users to access the content. What have been seen on the webpage does not necessarily belong to the current webpage. However, the content of the webpage, including what external resources this webpage want to link to, is determined by the owner of the webpage. In this sense, to connect the basis of Web architecture to different types of copyright infringement, not only the content of the Web resource needs to be examined, but further, different types of links that make the content available need to be considered. In the following section, issues of linking on the Web from technical and legal perspectives are discussed.

2.1.2 Linking on the Web

According to the Architecture of World Wide Web W3C Recommendation ("Architecture of the World Wide Web, Volume One"), all the resources on the World Wide Web have an identifier (URI) to point uniquely to them, and every resource may have one or more representations (Figure 1). The representations could be an HTML document (webpage) in different languages, an image, or a video file depending on the format of the resource. To link Web resources to each other, one resource may contain many outgoing links or URLs to other Web resources. In this sense, the URL of the resource tells where the resource can be found and the representation of the resource decides the content of the resource. When a URL is typed into a Web browser (also called user agent), the browser will retrieve the document identified by the URL and render the document to the visual content that can be seen (or be heard if it is an audio). During this process, the browser will also retrieve the resources that this Web resource links to and render it together with the current webpage as a whole. For example, in an HTML document, an <img> tag that points to an image from another domain can be embedded.

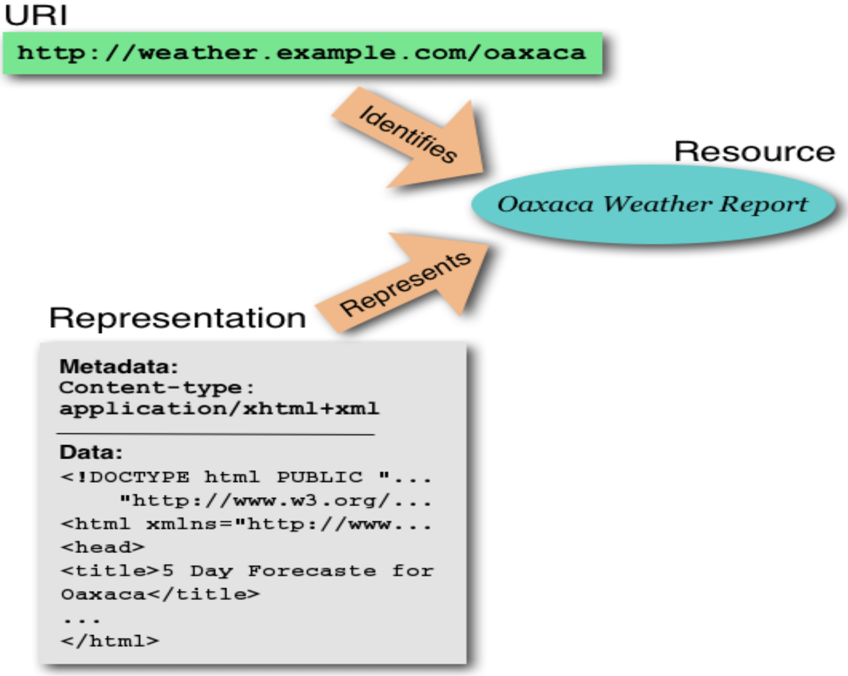


Figure 1. Relationship between identifier, resource, and representation

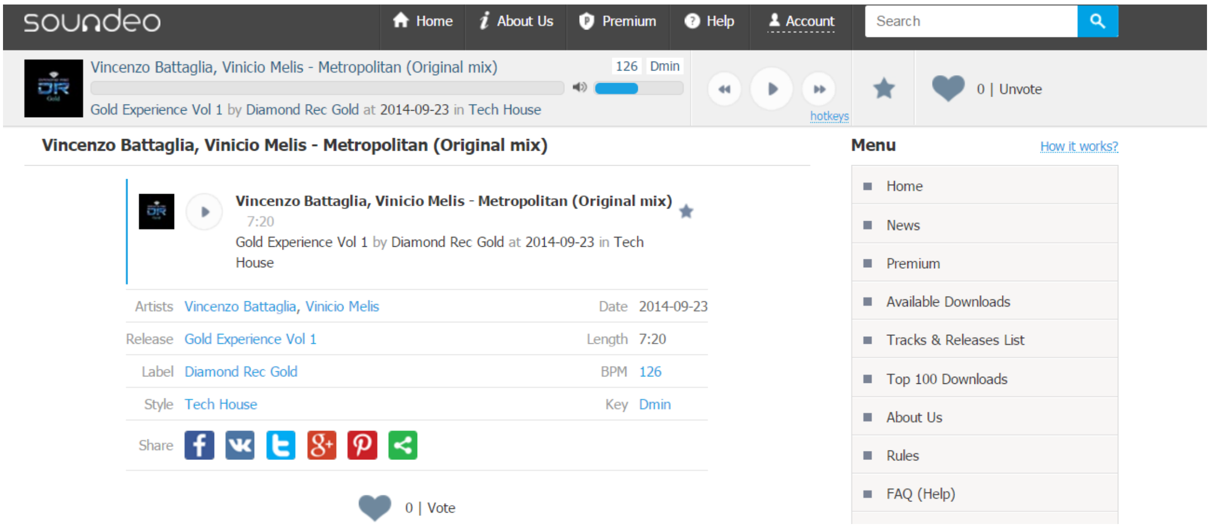
According to the W3C Recommendation ("HTML5- A Vocabulary and Associated APIs for HTML and XHTML"), links are a conceptual construct that represents a connection between two resources, one of which is the current document. In HTML, one type of link is linking resources such as CSS or JavaScript files to augment the current webpage. Another type of link are hyperlinks which link to other resources that are exposed to the user by the user agent so that the user can cause the user agent to navigate to those resources. This is the definition of links in HTML. However, in practice, because of various implementations of technology, links have broader meanings. Of course, links can be implemented in the same webpage or different webpages in the same domain, but here only cross domain links are discussed.

- Simple link and deep link
  "Simple link" and "deep link" refer to hyperlinks in HTML. A simple link is a clickable link which will lead visitors to other Web resources. One special type of simple link is called a Deep Link. Using a deep link, visitors will be directly pointed to the webpage within the website instead of the landing page. Thus, the visitors can view the content they are interested in without going through the hierarchy of a website. For example, a search engine could index a deep link http://example.com/path/page of the website http://example.com/. Then visitors can view the content of http://example.com/path/page without visiting the homepage first and having to click the links on the homepage to get to this page. Another example is the URL YouTube offers to view a video in the form of http://youtube.com/watch?v=videoid.
- Embedded link
  Many webpages contain content that links directly from other domains. In such cases, various types of links will be embedded in the webpage to tell the browser the content source. The

webpage can embed an image directly linked from another domain. For example, the http://example.org/index.html can contain an image from http://test.org using the HTML code <img src="http://test.org/img1.jpg">. In this example, the image from test.org is directly embedded in index.html and users are not explicitly notified that the image is from another domain. The same situation can also be applied to audio and video files, which can be embedded using <audio> or <video> tag in HTML.

A webpage can be embedded using an <iframe> tag. Specifying a URL using the "src" attribute in <iframe> tag will let the browser fetch the webpage the URL points to and display it in the current webpage. This method of embedding is also called "framing." In HTML, framing is a method to arrange and display content, and the content of different frames are independent of each other, i.e. they can be fetched from different domains. A difference from the previous example, where the link points to an image or video file, is that the content in the frame is usually treated as a webpage. However, similarly to the previous example of <img> tag, <iframe> also needs a link pointing to the resource that will be displayed in the frame. So both cases can be considered as embedded links.

One scenario in which embedded links are used is the embedding of music. On the page http://soundeo.com/track/vincenzo-battaglia-vinicio-melis-metropolitan-original-mix-5723680.html which is shown in Figure 2, users can listen to a sample of the music *Metropolitan*. Users also can control the playing of the music, such as pause and stop. From users' perspectives, the music is playing on the current website, but the content is embedded from another domain, beatport.com, using an HTML 5 <audio> tag.



Vincenzo Battaglia, Vinicio Melis - Metropolitan (Original mix) 126 Dmin

Gold Experience Vol 1 by Diamond Rec Gold at 2014-09-23 in Tech House

Vincenzo Battaglia, Vinicio Melis - Metropolitan (Original mix)

Vincenzo Battaglia, Vinicio Melis - Metropolitan (Original mix)
7:20
Gold Experience Vol 1 by Diamond Rec Gold at 2014-09-23 in Tech House

Artists	Vincenzo Battaglia, Vinicio Melis	Date	2014-09-23
Release	Gold Experience Vol 1	Length	7:20
Label	Diamond Rec Gold	BPM	126
Style	Tech House	Key	Dmin

Figure 2. An example of webpage that embed music from a different domain

Another scenario involving embedded links is YouTube. YouTube offers services to facilitate the embedded links. Compared to scenario 1, contents hosted on YouTube are generated by uploaders and not much expertise is needed to embed them on another webpage. Users can simply add a line similar to this <iframe width="560" height="315" src="https://www.youtube.com/embed/AbCdEfj" frameborder="0" allowfullscreen></iframe> to their webpage to embed YouTube videos.

Linking issues have triggered a heated debate among the legal community. An early paper written by Hasan A. Deveci (Deveci 2004) defined the different types of links and believed links bring a number of unresolved issues and raised some copyright concerns associated with linking, such as, "deep linking might bypass advertisements", "framing might not reveal the ownership of the page called up", and "a search tool might breach the terms and conditions for use of the site in question". Several legal cases from different jurisdictions are reviewed in the following paragraphs.

In the US case Perfect 10, Inc. v. Google Inc (Perfect 10, Inc. v. Amazon.com, Inc. and A9.com Inc. and Google Inc. 508 F.3d 1146 (9th Cir. 2007) 2006), the Ninth Circuit agreed that hyperlinks and framing were not infringing copyright since infringing websites existed before Google and would continue to exist without Google. 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 *Nils Svensson and Others v Retriever Sverige AB* case (CJEU C-466/12 Nils Svensson et al v Retriever Sverige AB, 13 February 2014 ECLI:EU:C:2014:76), an interesting question was raised as to whether hyperlinks are covered by the right to communicate works to the public (Arezzo 2014). The CJEU held that "the public targeted by the initial communication consisted of all potential visitors to the site concerned, since, given that access to the works on that site was not subject to any restrictive measures, all Internet users could therefore have free access to them". Accordingly, the links did not make the articles available to a "new public" and therefore the consent from the journalists was not required. In this case, the articles were published in the original website with the consent of copyright owners. A further question is, if the articles were published in the original website without the consent of copyright owners, would any third-party link that connected to the articles still be legal?

In another case, *BestWater International GmbH v Michael Mebes and Stefan Potsch*(CJEU C-348/13 BestWater International GmbH v Michael Mebes and Stefan Potsch of 21 October 2014 ECLI:EU:C:2014:2315), it was considered whether embedded linking from another freely available website ought to be considered as a communication to public. CJEU believed that the answer to the question could be found in the Svensson case. The CJEU held 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 (Rosati and Löffel 2014). In this case, the claimant claimed that its video was uploaded onto YouTube without permission. So *BestWater* leaves the question whether "embedding copyrighted videos is not copyright infringement, even if the source video was uploaded without permission" open (Van der Sar 2014).

In the GS Media BV v Sanoma Media Netherlands BV and others case (CJEU C-160/15 GS Media BV v Sanoma Media Netherlands BV and Others, 8 September 2016 ECLI:EU:C:2016:644), the CJEU decided that hyperlinks to a third-party website on which protected works were made available without consent of the rights holder constituted a communication to the public if the person placing those links knew this consent was not given. There is a presumption of knowledge that the consent is not given if the linker pursues a financial gain. However, this presumption can be rebutted so it is not absolute.

In the recent Stichting Brein v Jack Frederik Wullems (Filmspeler) case (CJEU C-527/15 Stichting Brein v Jack Frederik Wullems, 26 April 2017 ECLI:EU:C:2017:300), the CJEU held that the sale of a multimedia player which enables films that are available illegally on the internet to be viewed easily and for free on a television screen could constitute an infringement of copyright. The court believed that the defender has full knowledge of the intervention of his multimedia player adds-on which gave "access to a protected work to his customers and does so, in particular, where, in the absence of that intervention, his customers would not, in principle, be able to enjoy the broadcast work".

In the Stichting Brein v Ziggo BV and XS4All Internet BV case (CJEU C-610/15 Stichting Brein v Ziggo BV and XS4All Internet BV, 14 June 2017 ECLI:EU:C:2017:456), the decision "encompasses different types of platforms and operators with different degrees of knowledge of the character – lawful or unlawful – of the content made available therein" (Rosati 2017). The CJEU held that the Pirate Bay had knowledge of the fact that its platform was being used by its users to infringe copyright and there is a communication to the public "by the operator of a website, if no protected works are available on that website, but a system exists … by means of which metadata on protected works which are present on the users' computers are indexed and categorised for users, so that the users can trace and upload and download the protected works".

From this analysis it is found that in the EU, particular legal solutions have been developed through a set of complex cases, which is difficult to comprehend for the layman. One key question, which does not seem to have been resolved yet, is whether and to what extent intermediary providers can be held responsible for the activities of their users. Should they be responsible for the infringement action did by the users through their platform? Some concerns are discussed in the following section.

2.1.3 Internet intermediary liability

Reproduction of copyright works and make them available to public through different Web services, are major concerns for copyright owners and lawmakers. In the digital environment, copyright owners tend to sue those who facilitate the infringement of others instead of suing actual infringers (Lemley and Reese 2003). Seeking an injunction against an intermediary whose services are used by third parties to infringe an intellectual property right is proved to be an important tool for copyright owners (Rosati 2017). This is particularly the case in the context of the Web because of both the difficulty in tracing individual infringers and insignificant economic compensation when infringers are found.

David Lindsay categorized four types of intermediary: Telecommunications carriers, Internet Service Providers, Content hosts, and BBS operators (Lindsay 2000). Diane Rowland defined different categories of Internet intermediary based on the role they played in the online communication or transaction chain (Rowland, Kohl, and Charlesworth 2010): Connectivity, Navigation, Commercial and social networking, and Traditional commercial intermediaries and facilitators.

Based on the Internet and Web architecture, Internet intermediaries are divided into three main categories and these types of providers are defined to specifically fit the CLC Model.

- Internet access providers. These mainly provide connection to the internet and include cable companies, internet service providers (ISP) and backbone telecommunication providers.
- Host providers. Some intermediaries host news, documents, music and videos on their own servers. Examples are traditional host providers such as Fileserver, which provide users with online storage and upload/download services. Other host providers are Web 2.0 platforms such as social networking and commercial service providers like YouTube, Facebook, and eBay, etc., which also provide hosting for content that is generated by the users themselves.
- Link providers. Acting like search engines, they index online content and simplify navigation. Some other sites primarily offer links to stream live content or downloads. The content is hosted in a different location which is not administrated by the site itself.

When protecting copyright on the Internet, two alternatives to pursuing expensive copyright enforcement charges against individual users can be used: targeting intermediaries involved with the transmission of material via the Internet, or establishing technological means of restricting access to the copyright-protected material (Lindsay 2000). Internet intermediaries' liabilities are determined by reference to factors such as their actual or constructive knowledge, their relative control over the activity, and the financial benefit gained from those activities (Rowland, Kohl, and Charlesworth 2010).

The question whether a regulatory duty should be imposed on Internet intermediaries and to what extent the duty should be applied is still hotly debated. In the UK, the *Newzbin2* case (Twentieth Century Fox Film Corp v BT [2011] EWHC 1981 (Ch)) is an interesting case study. The Newzbin2 website offers search functions for a wide range of content, some of which is copyright infringing. The High Court ordered the Internet access provider (BT) to deploy a content filter technology to block the website. In the *Sabam* (CJEU C-70/10 Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM), 24 November 2011 ECLI:EU:C:2011:771) and *Netlog* (CJEU C-360/10 Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV, 16 February 2012 ECLI:EU:C:2012:85) cases, the Court of Justice of the European Union (CJEU) held that ISP could not be ordered to install a system of filtering of all electronic communications and blocking certain content in order to protect intellectual property rights. Similarly, the court held that social networks like Netlog "cannot be obliged to install a general filtering system, covering all its users, in order to prevent the unlawful use of musical and audio-visual work." In the Newzbin2 case, the High Court imposed a duty to use the filtering technology for two main reasons. First, the technology was already being used by the Internet access provider for a different purpose and was not excessive; and second, the judge believed there was no general monitoring of all the data of all users. In the Sabam and Netlog cases, the injunction or filtering process were not imposed because these actions relied upon the inspection of the content from all the service subscribers. There is not a big difference between upload filters and blocking injunctions at least technically. The measures apply to the entire user basis in both cases (Stalla-Bourdillon 2012a).

Based on different roles in the Internet activity, what procedure should Internet intermediaries apply to limit their liability? DMCA Section 512 gives some answers. The DMCA is the first piece of legislation to specifically create limitations on the liability of Internet intermediaries for copyright infringement. It introduces a notice-and-take-down procedure for host providers and to a lesser extent information location tools such as search engines. In European Union law, similar liability-exemption rules inspired by DMCA are set forth in the e-Commerce Directive. Although there are still significant gaps in the resulting framework, EU is making efforts in the harmonization in the area of IP enforcement.

Some EU Member States have adopted a notice-and-take-down procedure for copyright infringement. Still, in practice the DMCA notice-and-take down procedure remains the benchmark and arguably its main prescriptions relating to the content of notifications are compatible with the EU e-Commerce Directive, although others such as 512(3)(g)(D) relating to consent to the jurisdiction of Federal District Court for any judicial district in which the service provider may be found to be problematic.

2.2 Digital Millennium Copyright Act of 1998 (DMCA) "Notice-and-take-down" Procedure and Google's Practice

2.2.1 DMCA notice-and-take-down

Section 512 of the US Digital Millennium Copyright Act (DMCA) of 1998 creates limitations on the liability of online service providers for copyright infringement when engaging in certain types of activity. It categorises several Internet Intermediaries and adopts a "notice and take-down" procedure to impose legal duties on some of them. Section 512(a) gives Internet service providers who provide transmission and routing safe harbour from their users' infringements as long as they employ standard technical measures to prevent repeating infringement. Section 512(b) gives cache service providers protection if they respond expeditiously by removing or disabling access to infringing material when certain conditions are met and are subject to court-ordered injunctions to remove infringing material. In order to be protected by safe harbour, Section(c) requires host providers to perform the following steps when they received the removal notice:

1. Take down the infringing material "expeditiously".
2. Inform the alleged infringer the material has been removed.
3. If a counter-notification has been sent, forward any counter-notification back to the original complainant. After 10 to 14 days, if the Internet service providers are not informed of a lawsuit by the complainant, they can reinstate the material.

Section 512 (d) requires information location tools such as search engines to "expeditiously" remove the infringing links from their index. However, they do not need to notify the alleged infringer (i.e. the content provider) the link has been removed.

The "notice-and-take-down" procedure operates in a complex manner. If Internet intermediaries receive a notice, they must take down the material "expeditiously". In addition, if they receive a counter-notification and do not receive notice from the complainant about further court action, they will restore the materials.

Urban & Quilter used an empirical approach to look at the notice and takedown landscape, and collected data about the number and type of notices that were sent up to 2005 (Urban and Quilter 2005). They found that while s.512 constituted a quick way to police copyright on the Internet, some concerns arose regarding the "notice and take-down" procedure. There were a certain amount of invalid notifications which contained inaccurate information. Alleged infringers are subject to having their expressive materials removed before they receive a notice of complaint. Even when they can send a counter-notification, the material must stay down for at least 10 to 14 days according to the statute. "The effect may be to substantially burden expressive and other individual rights" (Urban & Quilter, 2005, p.637).

Section 512 (g) (1) creates limitation on liability for taking down generally if a service provider has "good faith disabling of access to, or removal of, material or activity claimed to be infringing or based on facts or circumstances from which infringing activity is apparent, regardless of whether the material or activity is ultimately determined to be infringing". Because the vast majority of "notices" are likely never subject to judicial scrutiny, the "good faith" may result in over-protection (Reichman, Dinwoodie, and Samuelson 2007) (Cobia 2008). Karaganis and Urban's research shows that the number of take-down requests increased dramatically in the last few years because they are sent by automated system. "An important question is whether automated notices do, in fact, reliably target infringing material" (Karaganis and Urban 2015).

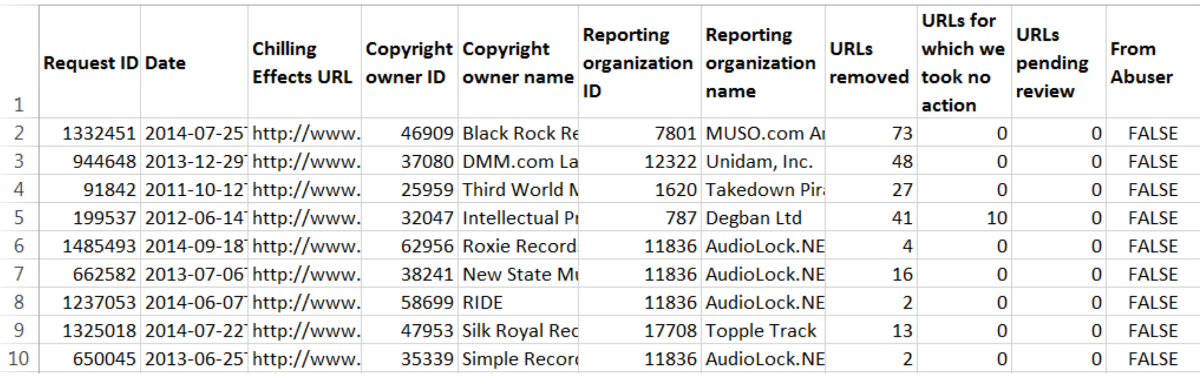
How to reform the "notice and take-down" procedure by law is a complicated issue which is still discussed by policy and law makers. In European law, similar liability-exemption rules are firstly outlined in the e-Commerce Directive. The European Commission also reviewed the rules on the intermediary liability by commencing a "Notice and Action" initiative (Kuczerawy 2015). Recently a copyright directive which aims at imposing upload filters has been proposed (Stalla-Bourdillon 2016). It is still in debate that 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 (Stalla-Bourdillon 2012b).

In practice, this procedure is used as a benchmark by many Web operators. For example, the website filestube.com owned by a Polish company applied the DMCA notice-and-take-down procedure. On certain pages on their website, they removed infringing content and stated that "Content has been removed on the author's request in accordance with the DMCA policy." Google also applies the notice-and-take-down procedure. Following Google's practice on the notice-and-take-down procedure will be discussed to understand how it is implemented in practice by online service providers.

2.2.2 Google's practice on notice-and-take-down

Google's take-down procedure starts when they receive notices sent by copyright owners, or their representatives, requesting that they remove potential copyright infringing content from Google's search results. Section 512(d)(3) DMCA requires "information location tools" such as Google to act expeditiously to remove the alleged infringing content when they receive take-down requests. The DMCA does not require them to verify whether there is an infringement. Google, however, has gone a step further on this procedure and deployed technologies and human intervention to assess take-down requests so as to determine if an infringement has occurred. If the URL is removed from search results, a "counter-notification" procedure will be activated to enable content owners to oppose the decision.

In a notice, information about copyright work, such as authorship, kind of work, title and infringing URLs is given. According to Google's Transparency Report, 831,185 notices, which contain over 300 million infringing URLs were received in 2014. On average, around 2,270 notices and around 821,900 URLs were received every day. Google assesses these notices and URLs to decide whether to remove the URL from its search results. Google does not release more information about how it assesses their take-down requests. One thing has been known is that Google has adopted a Trusted Copyright Removal Program (TCRP) (Google 2013) to help with the making of 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 (Tushnet 2013). However, the exact details of the program are shrouded in relative secrecy (Leiser 2013). Seng (Seng 2014) believes the programme is an automated method that allows notice senders to submit large numbers of take-down requests to Google, which Google would process rapidly via this programme. But no more information was published either about how it checks the validity of these requests or about how it checks the lawfulness of the content.



	Request ID	Date	Chilling Effects URL	Copyright owner ID	Copyright owner name	Reporting organization ID	Reporting organization name	URLs removed	URLs for which we took no action	URLs pending review	From Abuser
1											
2	1332451	2014-07-25	http://www.	46909	Black Rock Re	7801	MUSO.com Ar	73	0	0	FALSE
3	944648	2013-12-29	http://www.	37080	DMM.com La	12322	Unidam, Inc.	48	0	0	FALSE
4	91842	2011-10-12	http://www.	25959	Third World N	1620	Takedown Pir	27	0	0	FALSE
5	199537	2012-06-14	http://www.	32047	Intellectual Pr	787	Degban Ltd	41	10	0	FALSE
6	1485493	2014-09-18	http://www.	62956	Roxie Record	11836	AudioLock.NE	4	0	0	FALSE
7	662582	2013-07-06	http://www.	38241	New State Mu	11836	AudioLock.NE	16	0	0	FALSE
8	1237053	2014-06-07	http://www.	58699	RIDE	11836	AudioLock.NE	2	0	0	FALSE
9	1325018	2014-07-22	http://www.	47953	Silk Royal Rec	17708	Topple Track	13	0	0	FALSE
10	650045	2013-06-25	http://www.	35339	Simple Recor	11836	AudioLock.NE	2	0	0	FALSE

Figure 3. Shortcut of request.csv from Google's Transparency Report

Figure 3 shows the removal request sent by copyright owners/representatives and Google's take-down action. When a take-down notice is received by Google, a decision is usually made within 6 hours ("Google Transparency Report - FAQ" 2016a). Google will remove the URLs in its index if it decides there is a copyright infringement on the webpages. However, there are instances in which Google will refuse to remove the webpages if the request is inaccurate or intentionally abusive. Google lists several examples of such cases ("Google Transparency Report - FAQ" 2016b). In these instances, decisions will be made quite quickly. However, there are other cases where the decision cannot be made within 6 hours which will be listed as "pending" in the transparency report. It is assumed that the programme has difficulties to analyse those webpages and human interaction is needed to make a decision. Usually, the pending status will be resolved within days following which a final decision will be made.

Although it is not clearly mentioned in Google's report, it is found that domain name analysis plays an important role in the decision-making process. From the Google Transparency Report and its website, it is believed that Google has been doing extensive data analysis on domain names ("Google Transparency Report - Explore the Data" 2016), which is a high-level summary of decisions on the allegedly infringing action. For example, the Transparency Report website lists the number of URLs that were reported under the same domain name during a period, the number of URLs that were already removed from the same domain name, and the number of notice-senders who reported the same domain although they could have reported different URLs, etc.

Firstly, from a technical point of view, this method is relatively simple. Taking the domain vmusice.net ("Google Transparency Report - Domain" 2015) as an example, between 8th August 2012 and 8th February 2015, Google received 40,372 notices containing 3,236,150 URLs under this domain. There were 188 around URLs per week. Because vmusice.net is a top domain specified, Google's automated program has a high take-down rate of URLs under that domain. To what extent Google goes further to assess the exact content under each single URL is still unknown. Technically, it is much easier for a system just to compare domains instead of the actual content in the webpages that URLs point to.

Secondly, this method is relatively safe from a legal point of view, and it follows, to some extent, the practice defined in Section 512(g) (1) DMCA. Section 512(g) (1) DMCA 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 to contain infringing content, online service providers will act in "good faith" to remove any URLs under that domain without bothering to examine every reported URL. However, the questions that arise in conjunction with this are: (1) is the domain-driven method sufficient enough to make sure a reasonable take-down accuracy? and (2) does it result in "over take-down"?

IPL reviewed Google's take-down procedure and published a report (IPL 2013) in 2013. It calculated the take-down accuracy in one day (30/03/2013) and predicted the effects an enforced time limit would have on accuracy. The accuracy in that single day was 0.998, which is very high, and increases to 0.9995 with a longer time limit. However, this number needs to be further discussed because IPL's definition of accuracy considers a take-down decision to be inaccurate when a removed URL is reinstated. This only happens when Google receives a counter-notice. Compared to a large number of notices, the number of counter-notices remained very small. And the counter-notice procedure is not working very effectively in practice. Some explanations are found in the IPL's report and Urban's study (Urban, Karaganis, and Schofield 2016): the content provider being unaware the URL has been taken down, the content provider not understanding the law or the counter-notice procedure, and the content provider not being sufficiently interested in the content of the URL to issue a counter-notice.

2.3 Copyright Enforcement Technologies

One way for copyright owners to seek protection for their work is to establish technological means of restricting access to the copyright-protected material. Many pieces of software and systems which also are named algorithmic law enforcement systems have been designed to help detect infringement. For example, YouTube provides a Content Verification Program (Kim 2012) to help copyright holders search infringing content and issue multiple take-down requests. Vimeo also claims that all infringing materials will be removed according to Digital Millennium Copyright Act. Google has generated Transparency Reports on Copyright Removal Requests to publish relative copyright notification data. In the following section, three important copyright protection technologies implemented in the Internet and Web context are reviewed. And the algorithmic copyright enforcement systems are discussed.

2.3.1 Digital Rights Management (DRM)

Although the term DRM has a range of meanings, its narrower definition is: systems that restrain a work with technological lockbox by encryption, and use of the work requires some specific software or hardware acting as a gatekeeper to determine which uses proceed and which are blocked (Felten 2003). Apple's Fairplay, Amazon DRM and Adobe PDF Merchant are all DRM systems that have been used or are in current use. DRM systems mainly implement three technologies (Lannella 2001):

- Cryptography. The protected content is encrypted and packaged for distribution, and will be decrypted in a secure end-user environment. Encryption uses an algorithm and a key to scramble the digital content. The key for decryption to recover the original content is provided to legitimate consumers (Subramanya and Yi 2006).
- Content Identification Technologies. Digital signatures or digital watermark are inserted into the content to serve as a proof of ownership identity (Ku and Chi 2004). The content identification technology will be reviewed in detail in next section.
- Rights expression language. Rights expression languages is used to express the terms and conditions of content usage in an unambiguous manner. In the license generation process, a license stating the rights and conditions of content usage is generated. The license also contains a key required to unlock the protected content.

DRM aims to tackle piracy and regulate users' access control of copyright works. However, on one side, most DRM systems are accused of not considering the legitimate exceptions such as fair use/fair dealing, the lack of which results in negative impacts on civil society (Justice et al.). On the other side, some researchers believe that streaming content either through subscriptions or on-demand "pay per view" will become the dominant model, instead of purely controlling user access to downloaded content ("Digital Rights Management (DRM): Media Companies' Next Flop" 2006).

2.3.2 Content Identification

Content identification technologies are used in the area of online anti-piracy and file usage tracking. It contains three key enabling technologies: watermarking, fingerprinting, and content identifier standards.

Digital watermarking, or watermarking for short, is used to examine a file in order to verify its identity. It involves modifying the 'noise' portion of the content in a file so that it contains some identity data, called the payload, in such a way that the user's perception of the content is not impaired (Rosenblatt 2008). Special hardware or software examines the file, searches for the watermark and extracts the payload. Digital watermarking techniques were mainly applied to digital still images, but were not very effective, and some content owners had concerns over losses in perceptual quality (Singh and Chadha 2013).

Fingerprinting is a more recent technology than watermarking. The premise of fingerprinting is the examination of a file (audio, video) and identification of the content using the perceptual characteristics of the file itself across file formats, codecs, bitrates, and compression techniques. On the current market, Audible Magic Company’s SmartID and CopySense technologies (“Media Identification” 2017) are both based on the fingerprinting techniques. YouTube’s Content ID also relies on fingerprinting to sample an uploaded file and compare it against a database of reference files (Stone and Helft 2007).

The third technology, Content Identifier Standards, regulate how to identify online/digital content. It also reveals the content’s properties such as Registration Authority and Content Type etc. There are many standards because of the subtle complexities of content identification. Examples of Content Identifier Standards, for example, ISAN (International Standard Audiovisual Number), UMID (Unique Material ID), ISWC (International Standard Musical Work Code) and DOI (Digital Object Identifier).

2.3.3 Content Blocking

Content Identification technology works on the verification of end-user products, which is located in the Application layer in TCP/IP protocol stack (Frystyk 1994). Content Blocking technology is focused on the Network and Transport layer in TCP/IP protocol stack and mainly implemented by the Internet Service/Access provider and network operators. There are three basic methods of blocking content (Clayton 2005). They are packet dropping, DNS poisoning and content filtering.

- Packet dropping. The IP (Internet Protocol) addresses of the websites to be blocked are listed, and any packet request to these IP addresses will be discarded. The main problem with packet dropping is over-blocking because it denies access to all the Web content on a particular IP address.
- DNS poisoning. Domain Name Systems (DNS) enable the translation of a domain name into a corresponding IP address. DNS poisoning system will fail to return the correct IP address of the blocked sites. It also suffers from over-blocking, but it doesn’t block other domains that are hosted on the same machine.
- Content filtering. URL blocking is one method. It will not only check the header of a packet but also the contents of the packet. Deep Packet Inspection technology (Bendrath and Mueller 2011) could be performed to examine the content of the packet. So content filtering will only block specific contents of the website instead of blocking the entire website. CleanFeed (Richardson 2004) is a filtering technology that examines all headers from all packets of all users so that packets destined for the blocked IP addresses are redirected and examined again to precisely locate their destinations. Content filtering involves monitoring of all users and applying packet inspection technology; it is considered to be intrusive to users’ privacy and fundamental rights and liberties in some cases.

2.3.4 Algorithmic copyright enforcement systems

The technologies discussed in previous section are independent technical methods to identify authentication of copyright work or to block access to copyright work. In this section, systems that implement legal principles to automatically make or support decision on copyright infringement will be reviewed. This type of system is also called algorithmic copyright enforcement system.

Algorithmic enforcement for online behaviour was predicted by information law scholars such as Joel Reidenberg and Lawrence Lessig. Reidenberg discussed that the development of technology and communication networks had established a set of technological standards named “Lex Informatica” for policy and rule makers (Reidenberg 1997). Lessig proposed the term “code is law” to describe how algorithms can substitute law in regulating certain behaviours (Lessig 2009). While technology has been employed to support law enforcement in many areas, algorithmic copyright enforcement often involves the implementation of flexible legal standards, which is quite challenging for a machine. Design and build algorithmic systems of law enforcement is a complicated task, which requires knowing what cognitive frames as well as social, political, economic and legal motivations shaped the choices made by those who design them (Kesan and Shah 2003). The legal domain has its own specific characteristic: a particular question requires some deduction or inference before an appropriate answer can be given. As many researchers mentioned, issues in copyright law involve discretion, including deciding the degree of originality, and deciding what amount to “substantial similarity” to establish infringement.

"Reaching a determination about these flexible issues largely depends on a qualitative process of assessment and balancing that ought to be made on a case-by-case basis"(Perel and Elkin-Koren 2016).

One typical application of legal algorithmic enforcement system is a legal expert system. There are two types of legal expert systems according to the way in which such a system represents the law that it contains (Australian Administrative Review Council 2003). The two ways are through constructed knowledge and learned knowledge. The main type of legal expert systems that use constructed knowledge is the rule-based system. It applies the legislations, model rules and automates the process of investigating those rules. It relies on deductive reasoning based on "if A, then B" rules (Aikenhead 1995) (Schauer 1991). Case-based reasoning and neural networks are typical systems using learned knowledge. Ashley summarized that all case-based reasoning "employs some methods for generalizing from cases to support indexing and relevance assessment and evidences two inference methods: constraining search by tracing a solution from a past case or evaluating a case by comparing to past cases" (Ashley 1992). Some examples of legal expert system such as SHYSTER, Split-Up, ASHSD-II. SHYSTER is a case-based legal expert system which "produces its advice by examining, and arguing about, the similarities and differences between cases" (Popple 1993). Split-Up is a rule-based reasoning system in order to predict the outcome of property disputes according to Australia family law (Zeleznikow and Stranieri 1995). ASHSD-II (Advisory Support for Home Settlement in Divorce) is a hybrid legal expert system that explores rule-based and case-based reasoning in the area of matrimonial property disputes (Pal and Campbell 1998).

Nowadays, major online service providers such as Google, Facebook and Twitter use algorithmic copyright enforcement to filter, block and disable access to allegedly infringing content automatically, with little or no human intervention (Urban, Karaganis, and Schofield 2016). A typical example is YouTube's Content ID system (Miller 2010). Content owners/copyright owners can submit their content to YouTube. A fingerprint of this content is created and saved in a database by YouTube. Once a user uploads a content to YouTube, the content is compared against the database to find out whether there is a match to copyright work (Charrington 2013).

2.4 Summary

Development of digital technologies, particularly the Internet and the Web, bring new challenges to copyright protection. The Internet and Web supply an open platform which enables free communication and sharing of information online. Nowadays, linking issues have caused hot debate for legal professionals such as whether hyperlinks and embedded links should be considered as communication to public, and in what context these links should be treated as communication to public. From technical perspective, using technical methods to clarify where and how the allegedly infringing content is linked is a major concern in this research which will be discussed in the later chapters.

As an essential component of online activities, Internet intermediaries such as Internet access providers, content hosts and publishers, and link providers, are more and more frequently issued with regulatory duties to gather information in order to identify unlawful content or modify online behaviour to prevent or terminate illegal activities. Different copyright enforcement technologies such as DRM system, content identification and blocking technologies, and algorithmic copyright enforcement systems are also developed and employed by internet intermediaries to deal with copyright infringement in the Internet era.

While it is still not clear what regulatory duties Internet intermediaries should be imposed on, many online service providers are following the notice-and-take-down procedure. A typical example is Google who implements this procedure and has published a related Transparency Report regarding requests to remove content due to copyright. When Google receives a removal request, it examines the Web resources for potential infringement claimed by copyright owners to decide whether to take it down. However, the report does not disclose the criteria used to decide whether the content should be taken down. It is important to understand what the criteria to assess the related Web resources are. A further question that can also be asked is which technologies can be implemented to support the assessing process, and how this implementation of technology can be fitted with the process, particularly in the context of large amount of take-down request? In the next chapter, using Google transparency report as a benchmark, a preliminary study to find out the patterns of infringement through webpages will be discussed.

Chapter 3 Preliminary Study on Google's Transparency Report on Copyright

Chapter 2 provides the background of the current copyright issues on the Web. To understand more thoroughly the notices and the reported infringing Web resources, this chapter will analyse the Google Transparency Report, specifically the "request by copyright owners to remove search results". This report is openly available and provides comprehensive information in respect of webpages associated with the potentially infringing content. In Section 3.1, 730 URLs with alleged copyright infringing content from the Google Transparency Report were viewed and examined. In Section 3.2, the results of manually labelled URLs are discussed.

3.1 Study on URLs

3.1.1 Selection of URLs

The Google Transparency Report, is a useful repository which provides information in respect of Web resouces associated with allegedly copyright infringing content. By analysing Google Transparency Report data which includes data relating to allegedly infringing webpages, a detailed categorisation and summarization of the formats or patterns of reported infringing webpages, and the ways in which these webpages are infringing will be discussed.

Google's Copyright Transparency Report lists metadata of each notice received every day since March 2011, and it links to the Chilling Effect database (called Lumen database now) (Seltzer 2001), which contains detailed information on each notice. The database stores the content of each notice, including the copyright owner, the kind of work, the title of the work, and potential infringing URLs that identify the location of the copyright infringement of that work on the Web. Through the connection of the Google Transparency Report and the Lumen database, it is known exactly which URL has been taken down by Google from its search results.

As mentioned before, over 300 million URLs were requested to be removed by Google in 2014. A set of sample URLs were chosen to be manually reviewed. The sample size firstly needs to be determined because "inappropriate, inadequate, or excessive sample sizes continue to influence the quality and accuracy of research" (Barlett, Kotrlik, and Higgins 2001). The following conventional formula to calculate the sample size is used (Daniel and Cross 2013).

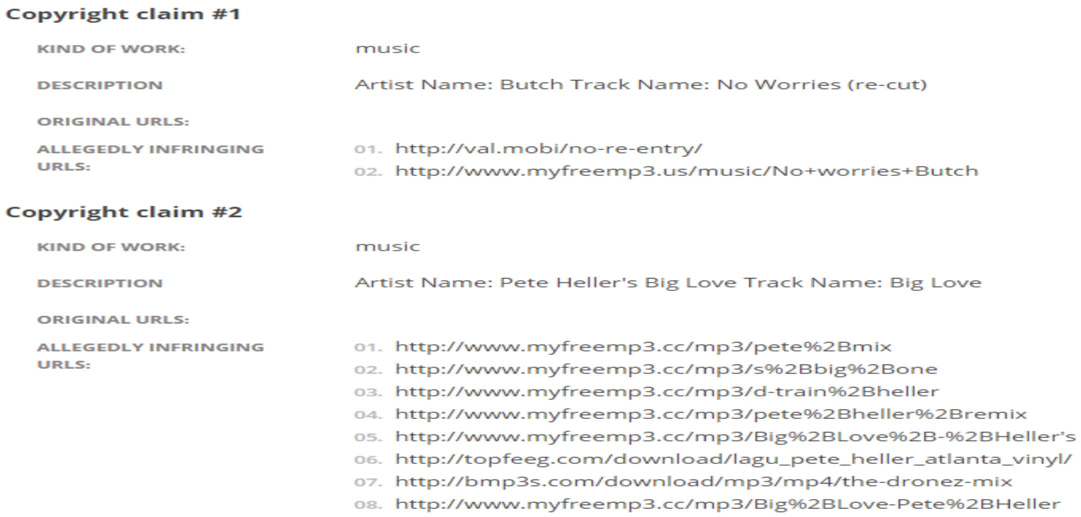
$$n = \frac{Z^2 \times \hat{p} \times (1-\hat{p})}{c^2}$$

Where

- Z is the Z-score, confidence level of 95% is chosen, the corresponding Z-score is 1.96.
- $\hat{p}$ is the prior judgment of the accuracy of Google's take-down. As mentioned in Chapter 2, IPL's report gave an estimation of the accuracy, however, it needs further verification. To make sure the sample size is big enough, 0.5 is chosen as the $\hat{p}$ value because mathematically f(x) is largest when x equals 0.5 in this formula $f(x) = x \times (1-x)$ where $0 \leq x \leq 1$.
- c is the margin of error, and 0.05 is chosen here.

The sample size is 384.

Because of the possibility that many webpages might have been taken offline for various reasons, such as broken URLs or already taken-down by Web operators, therefore, in the dataset more than 384 URLs are captured in order to make sure the actually assessed number of webpages is big enough.



Copyright claim #1

KIND OF WORK:	music
DESCRIPTION	Artist Name: Butch Track Name: No Worries (re-cut)
ORIGINAL URLS:	
ALLEGEDLY INFRINGING URLS:	01. http://val.mobi/no-re-entry/ 02. http://www.myfreemp3.us/music/No+worries+Butch

Copyright claim #2

KIND OF WORK:	music
DESCRIPTION	Artist Name: Pete Heller's Big Love Track Name: Big Love
ORIGINAL URLS:	
ALLEGEDLY INFRINGING URLS:	01. http://www.myfreemp3.cc/mp3/pete%2Bmix 02. http://www.myfreemp3.cc/mp3/s%2Bbig%2Bone 03. http://www.myfreemp3.cc/mp3/d-train%2Bheller 04. http://www.myfreemp3.cc/mp3/pete%2Bheller%2Bremix 05. http://www.myfreemp3.cc/mp3/Big%2BLove%2B-%2BHeller's 06. http://topfeeg.com/download/lagu_pete_heller_atlanta_vinyl/ 07. http://bmp3s.com/download/mp3/mp4/the-dronez-mix 08. http://www.myfreemp3.cc/mp3/Big%2BLove-Pete%2BHeller

Figure 4. Copyright claims in each notice sent through Web form

Figure 4 indicates that in each notice, 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. One month's notices received by Google dated from September 24th to October 23rd, 2014 were chosen. The reason this period was chosen was that the experiment started around the beginning of October 2014, and the latest notice data that can be got at that moment was dated from 24th of September. For each day, the first notice received in every hour was picked up. And in every notice, two URLs from 1st and 2nd claims were selected to make sure URLs were chosen randomly. In total, 730 URLs were selected, which have been removed by Google from its search index.

3.1.2 State of URLs

Among the 730 URLs, there were 202 URLs that were not found, which usually returned 404 error ("Status Code" 2016) when they were visited. The website hosting server would typically generate a "404 Not Found" webpage when a user attempted to follow a broken or dead link. Among the 202 URLs, 97 URLs were from file sharing websites that were also named One-Click Hosters (Lauinger et al. 2013), such as uploaded.net and zippyshare.com. These URLs were broken or dead links when they were visited, and browsers were redirected to pages showing that contents under those URLs no longer existed. A reasonable deduction was that the content of these websites had already been removed by website administrators because of file expiration or copyright infringement. Google updated and published its data every day. However, it took a couple of days or longer for the Chilling Effect project to process and publish Google's data in detail. Because the 202 URLs extracted from Chilling Effect database were not found, the analysis results were based on the remaining 528 URLs which were removed by Google.

These URLs were retrieved from different notices, which were sent by different senders. There were four different types of senders found. They were: Individual, Industry, Agent, and Collective Management Organization. Individual is defined as a person who is very likely to be the personal copyright owner. Industry is usually a company who might be the copyright owner or its representative such as a law firm. Agents are usually anti-piracy service providers such as Muso and Degban, who normally offer online tools or services for their clients so that they can easily issue DMCA notices to related parties. Collective Management Organization is defined as a collaborative association whose members have the same interest such as British Phonographic Industry (BPI) and International Federation of the Phonographic Industry (IFPI).

Of these 528 URLs, 18 were sent by individuals, 55 were sent by industrial companies, 40 were sent by Collective Management Organizations, and 415 were sent by Agents. The dataset shows that a lot of notices were sent by Agents, which constituted approximately 79% of the total notices analysed.

The URLs point to various types of copyright work. Figure 5 shows the different types of copyright works that were claimed to have been infringed and their percentage in the total of the URLs examined.

The figure indicates 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 the notices were sent by the music industry. For books, some infringing websites make copyright books, particularly comic books, available for online reading. Some websites link to other Web hosts for users to download books. These three groups together make up 88% of the alleged copyright infringing webpages.

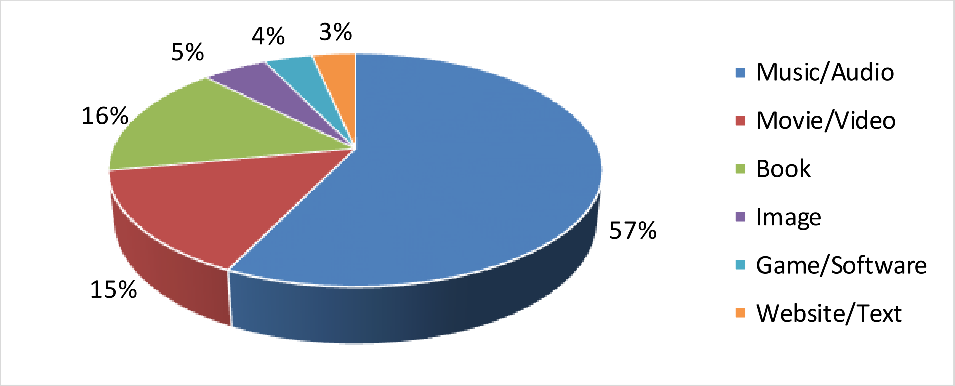


Figure 5. Type of copyright work that claimed to be infringed

3.1.3 Categorisation of URLs

Considering URLs features such as whether the allegedly infringing work exists on the webpage, whether the work can be accessible and how the work can be accessible, the URLs reviewed are divided into the following types:

- Type 1: Neither the metadata of the work such as title, author, publication time etc. are found on the webpage, nor the actual copyright work are found on the webpage.
- Type 2: The metadata of the work such as title, author, publication time etc. are found on the webpage, but the webpage does not supply any interface for users to get access to the content, so the actual work cannot be accessed by users. A typical example is music review websites, they give introduction and comments of music, but they do not offer any function which enable users to get access to the music.
- Type 3: The metadata of the work such as title, author, publication time etc. are found on the webpage, the webpage offer access interface for users but the content is not accessible. The content in this case is not hosted by the current website, instead, it is linked from different websites. The link is changed or broken, or the content is removed by other websites, this explains why users cannot get access to the content.
- Type 4: The metadata of the work such as title, author, publication time etc. is found on the webpage, the work is hosted under the current webpage and it is accessible.
- Type 5: The metadata of the work such as title, author, publication time etc. is found on the webpage, the work is linked from other website and is accessible through the current webpage.

3.2 Results Analysis

In the following section, detailed results after analysis of the 528 URLs will be discussed.

3.2.1 Validly removed URLs

There are 431 URLs that are considered to be validly removed. The main reason for the preliminary judgement is as follows. Having examined those webpages manually, it was found that copies of copyright work were made available or accessible through these webpages. Among those 431 URLs, some websites host infringing work themselves, some supply links to other illegal websites, and some embed links from illegal websites in another domain. The detailed results are discussed below.

- **Types of infringing websites**. There are five types of websites which can broadly be said to participate in infringement activities. They are online streaming websites, online reading websites, One-Click Hosters, index websites, and P2P communities. Online streaming websites enable content, including music/audio and movie/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. Through a simple Web interface, this type of website 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 searches simple 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 the index of .torrent files or running bit torrent tracker servers. Figure 6 shows the percentage of different types of reported infringing websites.

According to the results, 42% are online playing websites which offer users opportunities to play music, movies, and videos online. The number of One-Click Hosters is small (5% out of 431 URLs in total). Even though there are much more One-Click Hoster cases in the total 730 URLs, the files usually have been removed by the owners or no longer existed at the time when they were visited. As a result, these URLs are considered to be non-existent URLs. The number should be much bigger than that. P2P websites (19%) constitutes the second largest source of allegedly infringing content. So far as the remaining websites are concerned, 15% are indexing websites, which may supply unauthorized download links to illegal content, and 12% are online reading websites, which offer free online books, many of which are comic books.

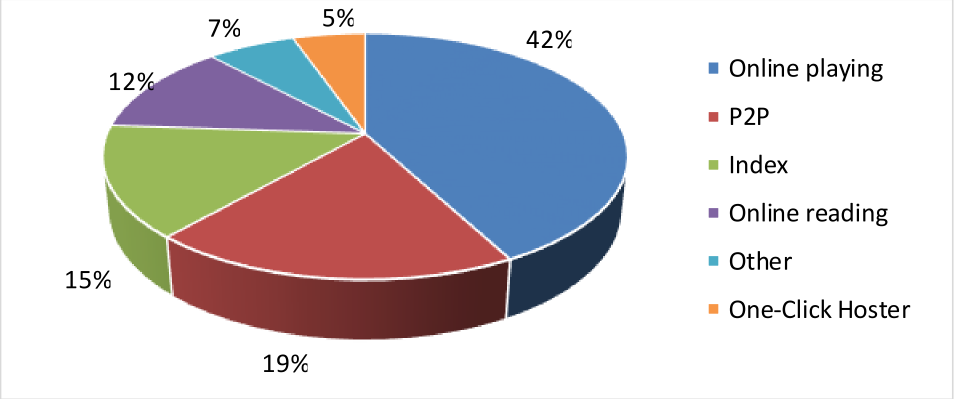


Figure 6. Different types of infringing websites

- **Whole or partial copy of a work.** Among the 431 webpages that the URLs point to, six webpages contain a partial copy of a copyright work. Partial copies of copyright works are samples of the original work. For example, a piece of music may have a complete duration of five minutes, the first two minutes of which can be extracted as a sample file, which can then be made available for users to play on a particular webpage. If the user is interested in this music, they can follow links on the webpage to download the whole copy of the work.
- **Online playing/reading, and downloading function.** Among the 431 webpages that the URLs point to, 234 webpages offer online playing/reading function. 226 webpages (not including those with P2P functions) offer downloading functions. If the P2P downloading is included, 308 webpages offer downloading functions. The content is either hosted by the current website for downloading or can be downloaded through simple links to other hosted websites. There are overlaps between these two groups of URLs because some websites offer both online streaming and downloading services.
- **Link type of online playing/reading content, and Link type of downloading resources.** Among the 234 webpages which offer online streaming/reading functions, 145 webpages embed contents from other domains; and the remaining 89 webpages host the content under the current domain.

Among the 226 webpages which offer downloading functions, 68 webpages use simple links which direct users to other websites to download the sought content. 64 webpages host the content themselves and allow users to stay on the current page to finish downloading. The remaining 94 webpages also allow users to stay on the current page to finish downloading, however, the content is not hosted by the current website.

Based on how the copyright work can be accessed, Figure 7 shows the percentage of different categories. Among the 431 URLs, 32% are categorized into embedded links, which means the infringing sources displayed on the current webpage are hosted from different domains. These domains belong to file sharing websites or cloud services, which supposedly host these contents illegally, i.e. without the consent of the copyright owners. 25% are directly hosting copyright work and have user interfaces which display these works to users. 12% are supplying simple links which link users to other websites to view or download copyright work. 16% are peer-to-peer websites which may host .torrent files, supply index of .torrent files or supply bit torrent tracker servers. 3% are supplying both embedded links and simple links. Therefore, generally speaking, most of the websites analysed do not host copyright work on their own servers, but use a variety of methods to link contents from different websites.

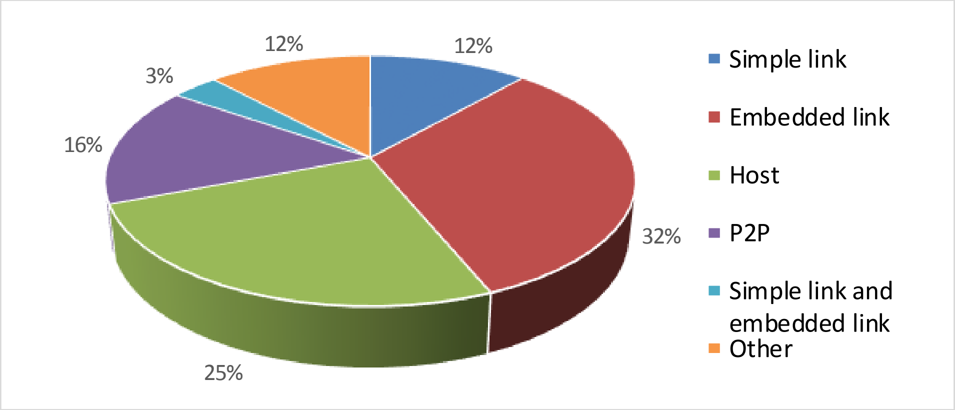


Figure 7. Different ways that copyright work is accessed

3.2.2 Invalidly removed URLs

7 out of 528 URLs are considered as clearly invalidly removed URLs, which makes up only 1.5% of those considered. The reasons for this characterisation are the following. Among these seven webpages, one is a website which offers plagiarism checking service for copying websites. This webpage is not considered to contain any copyright work or copy thereof, so it should not be removed. Six webpages contain the details of the metadata of pieces of music, such as artist, length, and release date. They also have online streaming functions to play samples of music. These samples are embedded from the authorized source BeatPort, who has developed an API, which allows Web applications to embed its sample music. None of these six webpages provides download functions at all. Some of them provide links to authorized legal websites such as iTunes or BeatPort.

3.2.3 Uncertain URLs

Among 528 existed URLs, 90 URLs removed by Google have been classified as uncertain. They correspond to 3 specific situations as Figure 8 illustrates it.

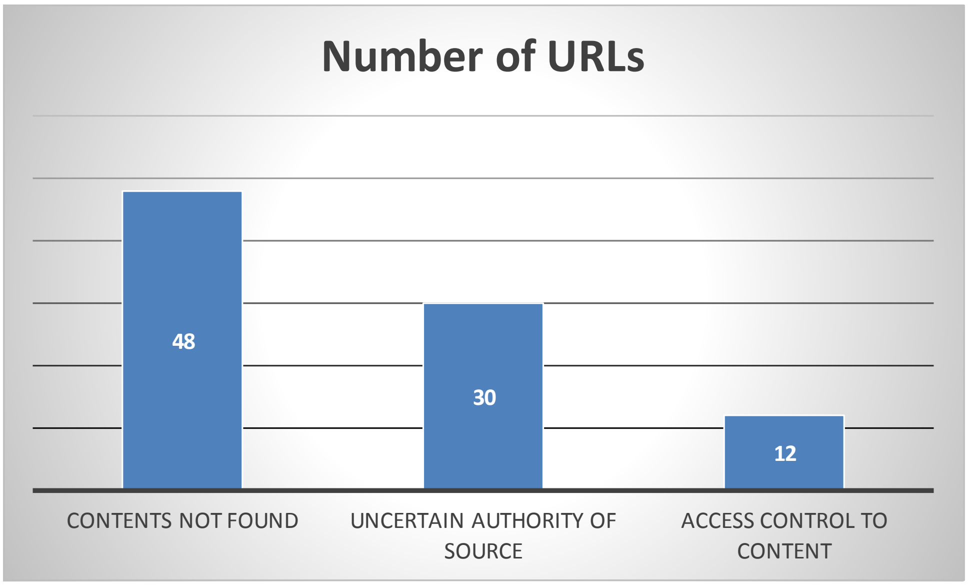


Figure 8. Uncertain URLs

- Contents are not found: For some notices, sent by copyright owners or their representatives, it was found that the title of the copyright work was missing; or the allegedly infringing content could not be found under the requested URLs. Among the 90 URLs, there were 48 URLs for which the allegedly infringing content could not be found. Another fact found was that 46 out of the 48 URLs were sent by agents (19 of them were requests from Total Wipes Music Group). The other two were sent by collective management organisations.
- Authority of the source is uncertain: Many websites embedded YouTube videos or music from SoundCloud. If the embedding was allowed by YouTube or SoundCloud, it was still not sure whether the content uploaded to those websites was authorized by the copyright owner or not. In this case the decision does not simply rely on the fact that the source is not authorised because the notification has been sent by the right holder. Another uncertain situation arises where some websites use images, but the usage is claimed as not authorized. Based on Google's Transparency Report, some websites of this description were removed, but some were not removed. In general, it is difficult to investigate the authorization of usage of images. So these cases were categorised in the situation of uncertain authority of source.
- Access control to view content: In some cases, the copyright owner claims that some URLs contain infringing content. But actually the content under those URLs are not visible because of access control measures such as logging in. In this case, whether these URLs should be taken down or not is not clear, so these URLs are listed as uncertain URLs.

3.2.4 Take-down accuracy

Considering the uncertain cases discussed in previous section, no conclusions are made on whether the take-down actions operated by Google for those URLs were valid or not. Therefore, it is reasonable to list different modes of "accuracy" calculation. The calculation of accuracy is also related to the actual action that are measured. Detailed analysis is listed below in Table 1. As mentioned before, there were 202 URLs that were not accessible in the UK or were broken links or the content under the URL had already been taken down when they were visited. They are not included in the analysis. So the base dataset includes existing URLs, the amount of which was 528. There were 90 among these 528 URLs that were categorized as "uncertain". Table 1 shows that if different calculation methods are selected, i.e. combinations of uncertain situations, the accuracy will change accordingly.

Table 1. Take-down accuracy

	Invalid removed URLs	Webpage accessible but content doesn't exist	Uncertain source authority	Access control to content	Accuracy
Amount of URLs	7	48	30	12	528
Calculation 1	√				98.7%
Calculation 2	√			√	96.4%
Calculation 3	√	√			89.6%
Calculation 4	√		√		93.0%
Calculation 5	√	√	√		83.9%
Calculation 6	√	√		√	87.3%
Calculation 7	√		√	√	90.7%
Calculation 8	√	√	√	√	81.6%

In Calculation 1, 7 URLs are considered to have been removed invalidly. In Calculation 2, 19 URLs (7 plus 12) are considered to be removed invalidly. And so, all the URLs in the "uncertain" category are considered to have been removed invalidly in Calculation 8. As a summary, even though there are different methods of calculations, the accuracy of take-down actions in the Google Transparency Report is between 81.6% and 98.7% after the manual examination of 528 URLs.

3.3 Summary

Through the study of practical take-down notices sent by copyright owners or their representatives, the format of allegedly infringing Web resources, the presentation of these resources, and the methods of making these resources accessible are investigated. Music/Audio represents the largest proportion of allegedly infringing work on the Web. These works are presented by online playing/streaming or downloading. Most of the allegedly infringing websites supply links to enable users easily get access to these works. Certain amount of websites are host providers which host these works for user's access.

Through the manual analysis of 528 URLs that have taken down by Google, they are categorized into validly removed URLs, invalidly removed URLs, and uncertain URLs, which result in an estimation of Google's take-down accuracy. However, the details of how Google assesses the allegedly infringing Web resources and makes the decision are still not known. Consequently, from the literature review in Chapter 2 to understand the copyright related issues in the context of the Web, and from the preliminary study in this chapter to explore the patterns of copyright infringement in practice, a model with a serial of criteria and workflow to examine copyright infringement on the Web will be developed.

Chapter 4 Content-Linking-Context Model

4.1 Introduction

Chapter 3 analysed the patterns and characteristics of allegedly infringing Web, and the next step in this research is to develop a model for copyright related criteria which will be applied to analyse the allegedly infringing Web resources that are requested to be removed. Then how these criteria are connected to operate as workflow to verify copyright infringement is explored.

The purpose of the model is to support the verification of allegedly copyright infringing material on webpages, preferably in an automatic manner. Obviously, strictly speaking only judges are competent to make a decision on the lawfulness of available Web resources. However, private actors such as online service providers are being asked to react upon allegedly infringing content before the issuance of a court order. There is still some uncertainty as to whether these private actors should automatically react upon satisfactory notifications or not. In consequence, the output of the model will be a score to indicate a likelihood of infringement with a view of supporting the decision making process and not necessarily replacing it.

In this chapter, Section 4.2 discusses the research methodology to build a serial of criteria for analysing allegedly infringing content on webpages which constitute the base components of the Content-Linking-Context (CLC) Model. Section 4.3 discusses dynamic work flow of CLC Model. Section 4.4 presents the process and algorithm to produce a score reflecting the probability of copyright infringement on a webpage by applying the CLC Model.

4.2 Methodology

To build the Content-Linking-Context Model for accurately analysing copyright infringement content on webpages, four steps have been followed to work as a circle to achieve the creation of the CLC Model.

Figure 9 shows the steps of the methodology.

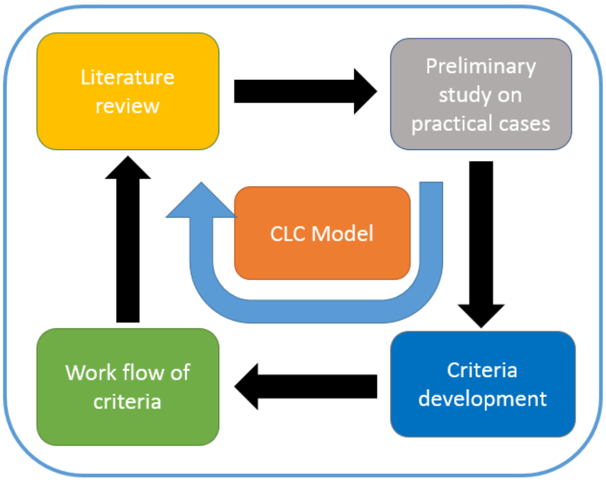


Figure 9. Methodology diagram

Step 1: A literature review of legal materials from different jurisdictions and current notice-and-take-down practices was undertaken in order to identify consensual infringement and non-infringement scenarios. Based on those study, five scenarios were constructed as listed below: four infringement scenarios and one non-infringement scenario. In order to construct these five scenarios a conservative view of copyright laws was adopted. A conservative view was needed to address uncertainties and simplify the analysis. Firstly, exceptions such as fair use are not considered at this stage for the CLC Model. Exceptions such as fair use require the development of a sophisticated model for norm representation which is not within the scope of this thesis but which is crucial to ultimately ensure a high degree of accuracy for notice-and-take-down procedures. Fair use in particular is a complex system consisting of multitude necessitating case by case analyses (Beebe 2008) (Samuelson 1993). The legal test has also evolved over time and across jurisdictions. In any case, it is likely that full automation is not possible to detect fair uses as the Lenz case illustrates it (Lenz v. Universal Music Corp., 801 F.3d 1126 2015). In addition, the results of the empirical study described in Chapter 3 show that exceptions were not relevant for the URLs analysed as no transformative work were encountered. As a result a broad definition of exclusive rights was adopted, and in particular given the persistence of uncertainties in the field it is 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, transformative uses of copyright works were excluded from the analysis and it was 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. Thus it is assumed there is an infringement in this case.
b. Hosting a partial copy of a copyright work without authorization. A partial copy of work is defined as a section of the copyright work which does not have any further additions, and which is a substantial copy. Thus it is assumed there is an infringement in this case.
c. Supplying links (simple or embedded) to an exact copy of a copyright work where making available of the copy is unauthorized. In this scenario, the website operator provides links for users to view/download unlawful content, and the online service provider is informed through notification that the link is to a content, where making available of the content has not been authorised. Thus it is assumed there is an infringement in this case or at the very least, a takedown should happen.
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. Thus it is assumed there is an infringement in this case or at the very least, a takedown should happen.
e. Supplying links (simple or embedded) to work made publicly available by the copyright owner. It is assumed there is no infringement.

Step 2: In order to investigate whether the most encountered scenarios in practice are covered by the scenarios listed above, the notices were examined in relation to the formats and patterns of reported infringing webpages as discussed in Chapter 3. From Figure 7, 25% allegedly infringing websites are host providers, and scenarios *a* and *b* refer to these types of websites. Forty-seven percent of allegedly infringing websites provide links (12% simple link, 32% embedded link, 3% both simple and embedded link) to copyright works. According to Figure 5 , most of these works are music, videos or eBooks. The allegedly infringing websites supply links to these content which enable users to get access to them. Scenarios *c* and *d* refer to these types of websites. Nineteen percent of websites provide peer-to-peer services and all five scenarios refer to these type of websites.

Step 3: Three categories of criteria were derived 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.

Step 4: A dynamic work flow was worked out to connect the three categories of criteria which identify the order of using these criteria. The basic method is using the criteria in Content category to check whether the allegedly infringing content is copyright work. And then use the criteria in Linking category to check where and how the content is accessible by users. Criteria in Context category work interactively with the other two categories of criteria to indicate the likelihood of infringement.

4.3 Criteria Development

A list of criteria is created to indicate the different factors that should be considered in the five scenarios listed in Step 1. These factors work as labels to help categorise URLs and understand the characteristics of Web resources. Making the copyright protected work accessible or downloadable on the Web are the main triggers for the characterisation of copyright infringement. From a technical perspective, a common method on the Web to make work accessible or downloadable is through linking. Consequently, linking issues such as where the work is linked from and how the work is linked are also important components in the model. The model was limited in the following ways:

1. The model uses the two types of links aforementioned: simple and embedded.
2. The model deals with the five scenarios identified in previous section.
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 5).
4. The principle of exhaustion does not apply to the supply of works online for music. There might be some exceptions to the principle of exhaustion in certain jurisdictions such as concerning software in the European Union (CJEU C-128/11 Usedsoft GmbH v Oracle International Corp, 3 July 2012 ECLI:EU:C:2012:407), but this is not the case for music. Therefore the principle of exhaustion will not be captured and represented in the CLC Model.
5. Although the accuracy of Google's domain-driven method needs further discussion, it does reflect the level of suspicion of a webpage. It is used as a factor to indicate the likelihood that the webpage contains copyright infringing content.

4.3.1 Criteria creation

Twelve criteria (C1 to C12) are proposed to indicate different factors that should be considered when verifying allegedly infringing Web resources in a notice.

- **C1: URL accessibility**. Whether the Web resource identified by the URL is still accessible. From the study in Chapter 3, it is found that a certain amount of webpages (URLs) do not exist anymore or they are already taken down by host providers. So the URL accessibility is checked at the beginning.
- **C2: Content existence.** When the Web resources identified by the URL are reviewed, whether the alleged infringing content can be found on the webpage. This criterion co-works with criteria C10 and C11. In the previous 528 webpages analysis in Chapter 3, 42 webpages were found contain no music content that was claimed as allegedly infringing content by copyright owners. At the same time, the context information such as the title and performer of the music cannot be found on the 42 webpage either. So from a technical perspective, it is believed that the context information can be used as a first and reliable checking step to determin whether a content exists or not.
- **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. The detailed technologies to compare music work with original copyright music work will be discussed in Chapter 6.
- **C4: Online access.** For music, whether the website offers an online-playing function. This criterion focuses on how the music work is accessible. Different answers to this criterion will lead to different results of the infringement assessment process. For example, a website that does not offer any function to play the music work may have lower infringement possibility than a website where a music work can be played online.
- **C5: Online playable**. Whether the music can be successfully played online. This criterion aims to figure out whehter the music work is eventually accessible in order to do further music comparison. In some cases, there is online access function (e.g. play button) offered by a website, but the music work is not actually successfully played. So different criteria will be operated instead of doing content comparison.
- **C6: Download access**. Whether the website offers a download function that enables the user to download the music. Apart from online playing method, another common method on the Web to make copyright work accessible is through download. So criteria considering

music download such as C6 and C7 are designed. The reason that C6 and C7 are developed is the same as C4 and C5.

- **C7: Downloadable**. Whether the music can be downloaded successfully.
- **C8: Link type of online accessing resources.** When an online accessing function is offered, whether the resource is hosted on the current domain, or is embedded from another domain.
- **C9: 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. C8 and C9 aims to tell how the resources are retrieved and delivered under the current webpage. The original source where the resources come from can thus be identified.
- **C10: Title of copyright work.** Information about the title of the music.
- **C11: Performer of copyright work**. Information about the person who performed the music.
- **C12: URL suspicion**. The likelihood that the current website contains allegedly infringing content. Google Transparency Report data of URLs that have been claimed to have infringing 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.

4.3.2 Criteria categorisation

The 12 criteria are divided into different categories: Content, Linking, and Context. As stated in the section of methodology, the categorization is 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. The criteria and the CLC model are explained below.

- **Content.** Allegedly infringing content on the webpage to which a URL point needs to be compared with the original copyright work in order to decide on the similarity between them. Those criteria are defined as “Content”. 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).

- **Linking.** As discussed in Chapter 2, there are different accessibility modalities of Web resources: the content displayed on the webpage is not necessarily hosted on the same domain of the webpage. In this case, the characterisation of infringement does not only depend upon an analysis of to the actual content displayed on the webpages, but also upon the inclusion of links in the webpages. Criteria C4 to C9 indicate how the allegedly infringing content is delivered and presented on the webpage. Allegedly infringing content could be directly accessed (and played) on the webpage (C4 and C5) or downloadable by users (C6 and C7). Criteria C8 and C9 reflect the requirement that the types of link need to be examined in order to reveal the source of the content and ultimately whether the initial source is authorized. These criteria are classified as “Linking”.

- **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 (C10, C11) of the content appears in the webpage, and whether the host website is highly suspected to contain copyright infringement work (C12), will be used in the analysis 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, C12 assesses whether the external domain is suspected to contain unlawful content. In Chapter 2 it was confirmed that Google had used Context information in their decision-making process, and they are useful for decision making, especially when the process needs to be automated. For example, Google has taken down a number of URLs from myfreemp3.cc, as it illegally offers download links to many music records. So a presumption could be that further claims on the URLs from this domain are quite likely to be valid. However, a simple conclusion cannot be made based on statistical assumptions and actually the content of any such claims needs to be examined to assess its validity. Consequently, Content, Linking and Context criteria should work interactively which will be discussed in the next section.

Figure 10 illustrates the concept design of Content-Linking-Context (CLC) Model, which explains the classes and their associations in CLC Model. The Request class represents a notice-and-take-down request, and each Request contains one to many WebResources indicated by URLs. In the CLC Model, the Context, Content, and Linking of each WebResource are examined. The Context consists of criteria about the metadata matching (title and performer matching) and URL suspicion. The Content class can be either a HostedContent or LinkedContent. The LinkedContent means even though the content is displayed within the current WebResource, the content is fetched from another URL other than the URL representing the current WebResource. The TypeOfDelivery class means the content can be delivered by OnlinePlay or Download. For C10 and C11, the LinkedContent will associate with an instance of Linking class. Depending on the nature of the linking, a Linking instance can be one of SimpleLink or EmbeddedLink. Compared with LinkedContent, HostedContent indicates the content delivered as the response to the request of the WebResource's URL.

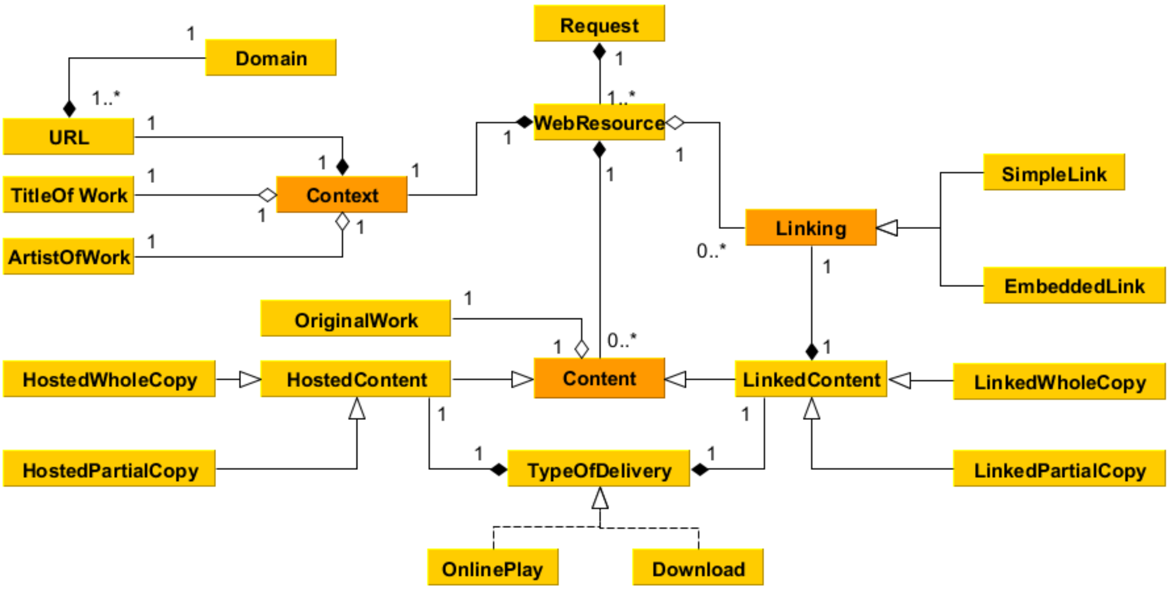


Figure 10. Content-Linking-Context conceptual design

4.4 Criteria Workflow

The conceptual design oversees the CLC model from a static point of view. Figure 11 illustrates a dynamic workflow using the CLC Model. C1 is going to investigate whether the URLs in the notice are publicly available. There are occasions on which a URL will point to a webpage, but the content of the webpage is not accessible as has been claimed in the request. It may due to various reasons. For example, the content of the webpage may have been changed by the website owner between the times that the request sender attempted to access it and when other people access it. Another example is that in some file sharing applications, users can set the files to be shared during a certain period. C2 is used to examine those cases. As discussed in Section 4.3.1, C2 also co-works with C10 and C11 to identify whether a content exists or not. 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, C6, C7) and where the content source is located (C8, C9), so as to further answer the questions of how likely there is a copyright infringement. In some circumstances, it is difficult to derive a clear answer as to whether there is a copyright infringement by analysing the Content and Linking criteria. For example, a website supplies functions to facilitate music to be played online or downloaded, but the music cannot be successfully played or downloaded. In this case, C10 and C11 are used to indicate whether the content exists, and C12 is further checked to indicate whether the website is suspected to contain copyright infringing material. In addition, if the allegedly infringing content is embedded from/linked to another external website instead of being hosted on the current reported one, C12 assesses whether the external domain is suspected to contain infringing content.

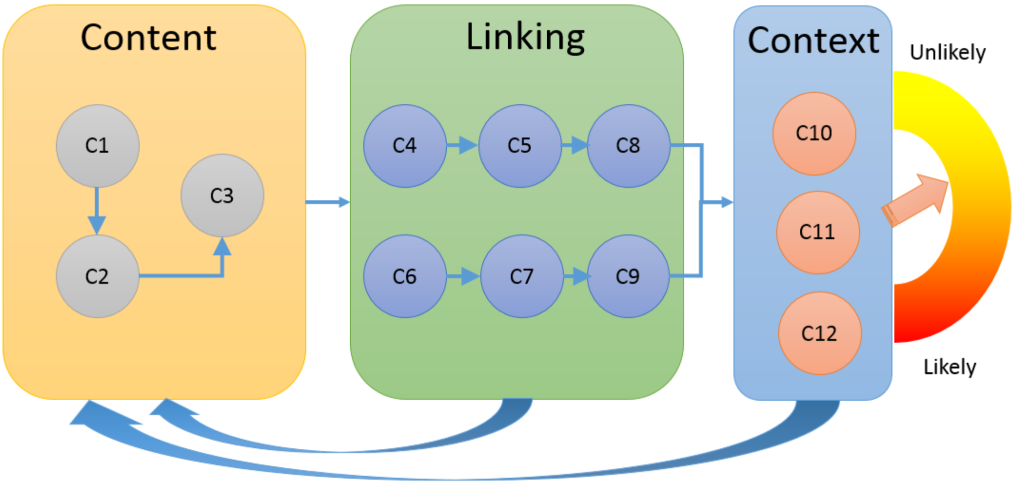


Figure 11. Content-Linking-Context dynamic illustration

4.5 Infringement Score Generation

The purpose of CLC Model is to help verify copyright infringing activity on webpages, preferably in an automatic manner. 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. In last section, the dynamic workflow of CLC Model highlights the idea of how to apply the 12 criteria to examine the allegedly infringing Web resources. In this section, the detailed steps of using each criterion and the algorithm to calculate the final infringement score are explained. Figure 12 shows the activity diagram of the process to produce a score reflecting the probability of copyright infringement on a webpage.

1. C1 is firstly checked to indicate whether the URL is accessible. If the webpage (URL) is not accessible (negative answers to C1), the output result will show that further assessment cannot be completed because of the issue of URL accessibility. This result is indicated as R1 in Figure 12.
2. If the webpage is accessible, C2 (co-works with C10 and C11) is checked to determine whether the content exists on the webpage. As explained earlier, the context information is used to inform this determination. A negative answer to C2 terminates the assessment and scores the probability of infringement as 0 because the content does not exist at all. This result is indicated as R2 in Figure 12. A positive answer to C2 leads to a consideration of C6. Table 2 shows the strategy used to decide the answer to C2.

Table 2. Decision table on answer to C2

Answer to C10	Answer to C11	Final answer to C2
Yes	Yes	Yes
Yes	No	Yes
No	Yes	No
No	No	No

3. If neither a download access function nor an online access function can be found on the webpage (negative answers to C6 and C4), the assessment terminates and scores the probability of infringement as 0 because the webpage does not supply any method to make the copyright work available or accessible. The result is indicated as R3 in Figure 12.

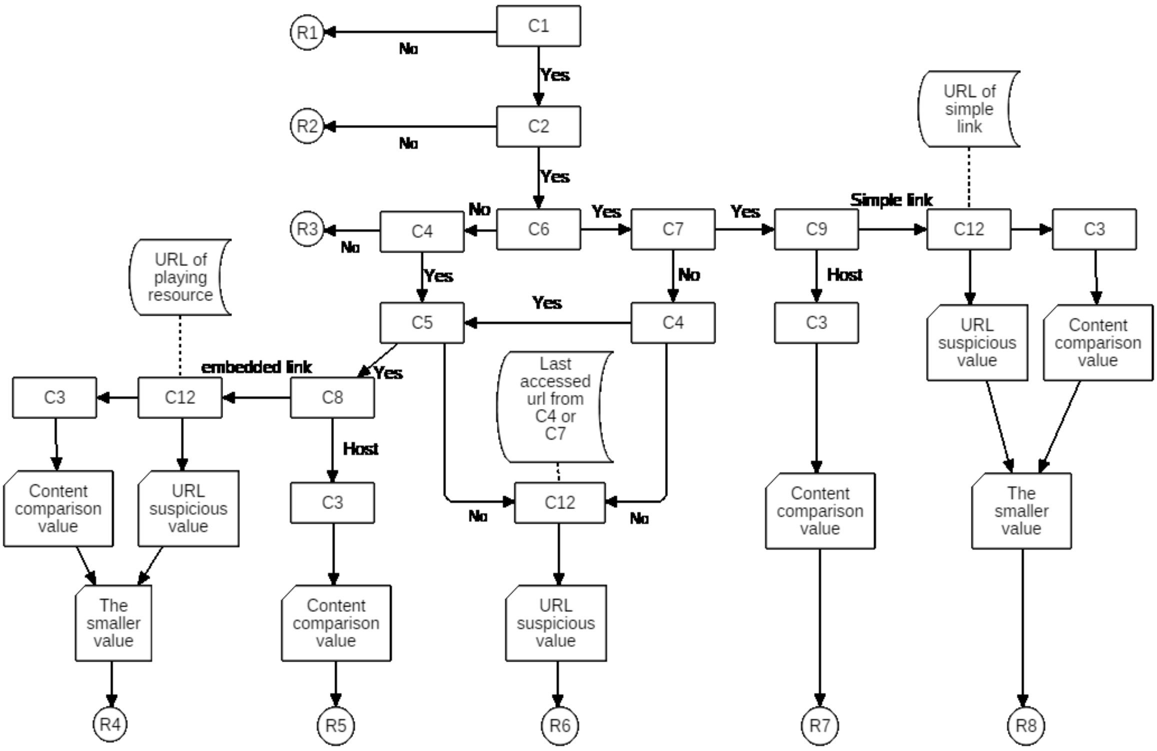


Figure 12. Activity diagram of assessment process

4. If there is no download access function (negative answer to C6), but there is an online access function (positive answer to C4), and the content can be played online (positive answer to C5), whether the content is hosted on the current website or is embedded from external website is checked (C8).
   a) If the content is embedded from an external website, the URL suspicion of the external website (C12) is calculated as well as the similarity between the content and the original copyright content. The score is the smaller of the two values, which is indicated as R4 in Figure 12. When C3 is smaller than C12, although the source where the content comes from is suspicious, the content similarity is lower than this and the probability of infringement is scored accordingly. Conversely, when C3 is bigger than C12, although the content is quite similar to original copyright work, the source is less suspicious and the probability of infringement is scored accordingly.
   b) If it is hosted on the current website, the similarity between the content and the original copyright content is checked (C3). The value of the content similarity is the score given for the probability of infringement. The result is indicated as R5 in Figure 12.
5. If there is a download access function (positive answer to C6), but the content can neither be downloaded (negative answer to C7) nor be accessible online (negative answer to C4), the download URL's suspicion (C12) is given as the probability of infringement score. Similarly, if only an online access function is found on the webpage (negative answer to C6, and positive answer to C4), but the content cannot be played online (negative answer to C5), online access URL's suspicion (C12) is given as the probability of infringement. The result is indicated as R6 in Figure 12.
6. If there is a download access function, and the content can be downloaded (positive answers to C6 and C7), whether the content is hosted on the current website or is linked from external website (C9) is checked.
   a) Similar to step 4a), if the content is linked from an external website, the URL suspicion of the external website (C12) is calculated as well as the similarity between the content and the original copyright content. The score is the smaller of the two values which is shown as R8 in Figure 12.

b) Similar to step 4b), if it is hosted on the current website, the similarity between the content and the original copyright content is checked (C3). The value of the content similarity is the score given for the probability of infringement. The result is indicated as R7 in Figure 12.

4.6 Summary

Applying appropriate criteria to assess Web resources in the context of removal requests in order to support notice receivers' decision making process is essential to improve the notice-and-take-down procedure. A four-step method was followed to design and develop a Content-Linking-Context Model. The CLC Model comprises 12 criteria and indicates how these criteria operate for the analysis of allegedly infringing Web resources. They are divided into three categories which constitute the three main components 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 whether a website is suspected to contain allegedly infringing works. The three categories of criteria work interactively to examine allegedly infringing Web resources step by step and eventually generate an infringement score to indicate the likelihood of infringement. In the next chapter, the expert validation of the CLC Model through quantitative and qualitative methods is investigated.

Chapter 5 CLC Model Validation

Chapter 4 proposed the CLC Model and explored the development process of criteria and workflow to generate the infringement score in the model. This chapter describes an expert review study aimed at validating the CLC Model. Section 5.1 discusses the methodology used to implement the expert validation. Section 5.2 explores the process of conducting the expert validation. Section 5.3 presents the data analysis following the expert validation. Section 5.4 further explains the results from Section 5.3, before finally Section 5.5 summarises the findings of the expert validation.

5.1 Methodology

A model should be developed for a specific purpose and its validity determined with respect to that purpose (Sargent 2005). It is impossible to define an absolute notion of model validity divorced from its purpose (Barlas 1994). Validation involves assessing the accuracy of the model's representation of the real system. Model validation is defined as "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" (Schlesinger 1979). Validation refers to the processes and techniques that the model developer, model customer and decision makers jointly use to assure that the model represents the real system to a sufficient level of accuracy (Carson 2002). Most validation methods involve both quantitative and qualitative processes, such as expert reviews, inspections, walkthroughs, data flow etc. (Barlas 1996).

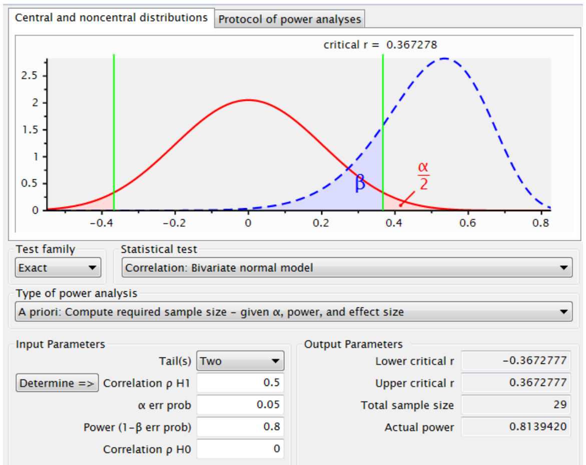
Expert validation is employed in this study because there exists no other model which addresses comprehensive criteria and which can assess allegedly infringing Web resources in the notice-and-take-down procedure before eventually generating a score to indicate the probability of infringement; as such, the model is new and must be validated. Using qualitative data from an experts review during a validation process offers useful feedback regarding the quality of the newly developed model (Newman, Lim, and Pineda 2013). Quantitative data generated by experts is related to measurement, which largely works in a complementary manner with qualitative data (Jick 1979). Therefore, in this study, both qualitative and quantitative methods are used for expert validation and review.

During the validation process, it was first essential to confirm the criteria used in the CLC Model. Secondly, the infringement score generated by the model needed to be validated through comparison with the experts' score. An electronic document questionnaire was given to experts. The participant experts in the study were located in different places, and thus it was easy for them to obtain the questionnaire and return it. The electronic questionnaire also gave them adequate time to consider their answers (Brace 2008). Quantitative data from experts was collected and analysed to investigate the level of agreement on the usage of criteria between experts and the model when different allegedly infringing webpages are viewed and examined. At the same time, qualitative data was obtained to explain the meaning of the quantitative data gathered. It also helped to identify new criteria suggested by experts which were not covered in the CLC Model.

5.2 Expert Validation

5.2.1 Selection of webpages

To investigate the level of agreement on the usage of criteria for assessing the allegedly infringing content on webpages, a certain number of webpages were presented to experts. These webpages were selected from real take-down notices received by Google. As mentioned in the preliminary study, a large number of take-down requests, including hundreds of thousands of webpages, are received by Google every day. A sample of these webpages had to be selected. The sample size needed to be determined firstly because it has an important impact on the quality of solutions (MacCallum et al. 1999). G* Power (Faul et al. 2007) was used to calculate the sample size through a correlation test. Figure 13 shows the result calculated by G*Power, where correlation coefficient $\rho0$ of the null hypothesis is 0, and the expected $\rho1$ correlation coefficient is 0.5. The value of Type I error α is determined as 0.05 and statistical power is 0.8, both of which are considered conventionally acceptable values (Cohen 1988; Fox and Mathers 1997).



Test family: Exact

Statistical test: Correlation: Bivariate normal model

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): Two
Correlation ρ H1: 0.5
α err prob: 0.05
Power (1–β err prob): 0.8
Correlation ρ H0: 0

Output Parameters

Lower critical r: –0.3672777
Upper critical r: 0.3672777
Total sample size: 29
Actual power: 0.8139420

Figure 13. Sample size calculation by G*Power

The calculation result from G*Power shows that the sample size is 29, thus meaning that 29 webpages are suggested for viewing and examination by experts. In this study, 29 webpages were chosen which met the following conditions:

- They are related to copyright music work.
- For each type of URL listed in Section 3.1.3, there are a certain number of webpages which belong to it.
- They cover all five scenarios listed in Section 4.2.

5.2.2 Selection of experts

Experts' education background, research and working experience were the main principles for the selection process. Because the purpose of the study is to validate and review the criteria developed in the CLC Model and the infringement score generated by the model, the experts needed to have a good working knowledge of intellectual property law, and preferably of the notice-and-take-down procedure. A recruitment post describing the study and requirements was published on 1709 Blog (Rosati 2016). Although there is no agreed-upon number of experts for a panel conducting the validation process, the most frequently suggested number is between three to five experts with different backgrounds (Jørgensen 2004).

Four experts were eventually selected to participate in the study. Two of them held the jobs of lawyer and academic researcher respectively, and were both working in the field of Information Technology and Intellectual Property Law. One of them was an IT/IP lawyer and the other had 20 years' working experience in advising and directing intellectual property policy and management. Three of them were located in the UK, while the final respondent was located outside of the UK. As aforementioned, the CLC Model relies upon consensual infringement scenarios and therefore it is assumed that the location of the experts was indifferent. It is sure nevertheless that the experts were coming from jurisdictions with a strong tradition of copyright or author right.

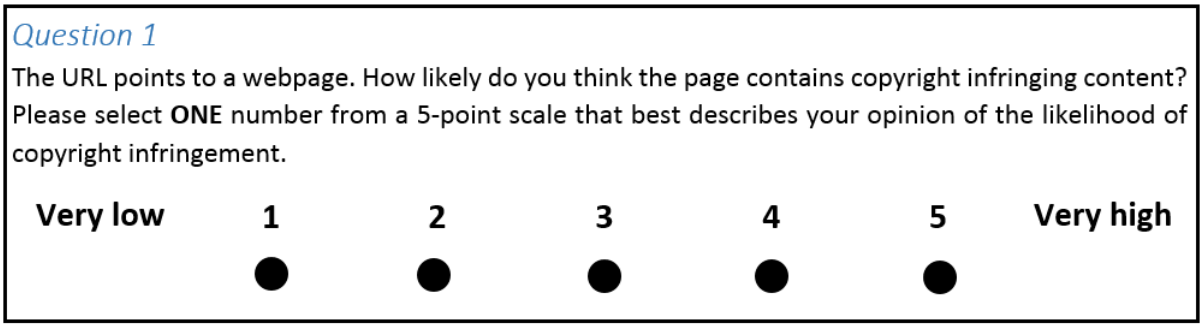
Before the start of the study, expert participants were contacted by email and provided with a detailed explanation regarding the purpose and process of the study; they were also sent an overview of the questionnaire, and instructions on how to complete said questionnaire.

5.2.3 Designing questionnaire

In order to validate the criteria and infringement score, the questions in the questionnaire were split into two parts. The first section of questions considered experts' rating score of the likelihood of infringement, while the second group of questions considered experts' opinions on the usage of criteria when they view and examine different webpages.

Part 1 comprised just one question, which was a closed question (Schuman and Presser 1979) asking the experts to give their rating on the likelihood of infringement. Closed questions are easy and quick for participants to fill in, and useful for generating statistical results in quantitative research (Coombes

2001). The likelihood of infringement was represented on a five-point Likert scale. Shorter scales such as five-point scales work effectively to measure attitudes or decisions (Dawes 2012). As such, the five-point scale was used in this study. Figure 14 shows question 1, which asked about experts' opinions on the infringement rating. Score 1 means the likelihood is very low, while a score of 5 means the likelihood is very high.



Question 1

The URL points to a webpage. How likely do you think the page contains copyright infringing content? Please select **ONE** number from a 5-point scale that best describes your opinion of the likelihood of copyright infringement.

Very low	1	2	3	4	5	Very high
	●	●	●	●	●	

Figure 14. Question example – 5-point scale

The questions in the second part of the questionnaire were a combination of closed and open questions. They were listed in a table pertaining directly to criteria usage. While 12 criteria are defined in the CLC Model, the questionnaire presented 9. All the webpages given to the experts existed, and thus C1 was not presented. Because C2 co-works with C10 and C11, and both C10 and C11 are always used as a first step to check content existence, C2 was not presented. C8 and C9 relate to the technology of the link type (host, simple link, embedded link) with which the experts may not be familiar; however, they may use a criterion related to the source of copyright work. Thus, criteria C8 and C9 were combined into one in the questionnaire.

There were two sets of questions in the second part. Question Set A sought to establish whether, generally speaking, the experts used the criteria, while question Set B asked whether the experts actually used the criteria when examining each webpage. As such, answers from Set B could further explain whether or not an expert used a criterion, while they may have given different answers to the Set A questions. For example, a webpage may contain some context information of a copyright music work such as title and performer, but the music content indeed cannot be accessed because the webpage supplies no function through which users can play or download the music. In this case, experts may still answer yes when asked Set A questions, such as whether they use criterion C5 considering online playable. However, the content cannot be obtained and there is no way to play this music. Thus, when they answered the question of "did you think the music could be played" (Set B), they were very likely to respond in the negative. Thus, the final answer regarding whether an expert used a specific criterion to access allegedly infringing Web resources was derived from answers to both question Set A and Set B.

5.3 Analysis of Validation Result

The data collected from the experts' review was analysed to investigate 1) whether experts agreed with the usage of the criteria developed in the CLC Model when accessing different webpages; and 2) whether experts agreed with the output of the CLC Model, which is an infringement score.

5.3.1 Analysis for level of agreement on usage of each criterion

Every expert was given a certain number of webpages to view and examine. The frequency of usage of each criterion by all the experts was calculated, and was then compared to the usage by the CLC Model. Cohen's kappa coefficient (k) (Kraemer 1982) was used here to indicate the level of agreement between the two patterns of criterion usage. The SPSS (IBM) tool was used for statistical analysis. Different classifications have been suggested for assessing how good the strength of agreement is when based on the value of Cohen's kappa coefficient. The guidelines below are from Altman (Altman 1990), and adapted from Landis and Koch (Landis and Koch 1977):

Table 3. Classification of Cohen's kappa (k)

Value of κ	Strength of agreement
< 0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very good

As explained in Section 5.2.3, C1 and C2 were not presented, and C8 and C9 were combined into one. Cohen's k was run to determine the level of agreement between the two usages of each criterion. The analysis for each criterion is discussed in the following paragraphs.

- C3: Work (Audio) comparison

Table 4. System_use * Experts_use Crosstabulation of C3

			Experts_use		Total
			No	Yes	
System_use	No	Count	21	4	25
		% within System_use	84.0%	16.0%	100.0%
		% within Experts_use	75.0%	7.0%	29.4%
	Yes	Count	7	53	60
		% within System_use	11.7%	88.3%	100.0%
		% within Experts_use	25.0%	93.0%	70.6%
Total		Count	28	57	85
		% within System_use	32.9%	67.1%	100.0%
		% within Experts_use	100.0%	100.0%	100.0%

Table 5. Measurement of agreement on C3

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.70	.084	6.47	.000
N of Valid Cases		85			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 4 shows that, of the 57 times that the experts used the C3, on 53 occasions the system also used it; this equates to 93.0%. Of the 28 times that experts did not use C3, there were 21 occasions on which the system did not use it either; this equates to 75.0%.

Table 5 shows that the kappa coefficient k = .70, p < .05, thus suggesting that there is significantly good agreement between experts and the CLC Model on the usage of criterion C3 for assessing allegedly infringing Web resources.

- C4: Online access

Table 6. System_use * Experts_use Crosstabulation of C4

			Experts_use		Total
			No	Yes	
System_use	No	Count	3	3	6
		% within System_use	50.0%	50.0%	100.0%
		% within Experts_use	42.9%	3.8%	7.0%
	Yes	Count	4	76	80
		% within System_use	5.0%	95.0%	100.0%
		% within Experts_use	57.1%	96.2%	93.0%
Total		Count	7	79	86
		% within System_use	8.1%	91.9%	100.0%
		% within Experts_use	100.0%	100.0%	100.0%

Table 7. Measurement of agreement on C4

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.42	.18	3.89	.000
N of Valid Cases		86			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 6 shows that, of the 79 times that experts used the C4, there were 76 occasions when the system also used it; this equates to 96.2%. Of the 7 times that experts did not use C4, there were 3 times when the system did not use it either; this equates to 42.9%.

Table 7 shows that the kappa coefficient k = .42, p << .05, thus suggesting that there is significantly moderate agreement between experts and the CLC Model on the usage of criterion C4 for accessing allegedly infringing Web resources.

- C5: Online playable

Table 8 shows that, of the 63 times experts used the C5, there were 58 occasions on which the system also used it; this equates to 92.1%. Of the 21 times experts did not use C5, there were 16 occasions on which the system did not use it either; this equates to 76.2%.

Table 9 shows that the kappa coefficient k = .68, p << .05, thus suggesting that there is significantly good agreement between experts and the CLC Model on the usage of criterion C5 for accessing allegedly infringing Web resources.

Table 8. System_use * Experts_use Crosstabulation of C5

			Experts_use		Total
			No	Yes	
System_use	No	Count	16	5	21
		% within System_use	76.2%	23.8%	100.0%
		% within Experts_use	76.2%	7.9%	25.0%
		% of Total	19.0%	6.0%	25.0%
	Yes	Count	5	58	63
		% within System_use	7.9%	92.1%	100.0%
		% within Experts_use	23.8%	92.1%	75.0%
		% of Total	6.0%	69.0%	75.0%
Total		Count	21	63	84
		% within System_use	25.0%	75.0%	100.0%
		% within Experts_use	100.0%	100.0%	100.0%
		% of Total	25.0%	75.0%	100.0%

Table 9. Measurement of agreement on C5

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.68	.093	6.26	.000
N of Valid Cases		84			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

- C6: Download access

Table 10. System_use * Expert_use Crosstabulation of C6

			Expert_use_kappa		Total
			No	Yes	
System_use	No	Count	2	3	5
		% within System_use	40.0%	60.0%	100.0%
		% within Expert_use_kappa	100.0%	3.6%	5.9%
	Yes	Count	0	80	80
		% within System_use	0.0%	100.0%	100.0%
		% within Expert_use_kappa	0.0%	96.4%	94.1%
Total		Count	2	83	85
		% within System_use	2.4%	97.6%	100.0%
		% within Expert_use_kappa	100.0%	100.0%	100.0%

Table 11. Measurement of agreement on C6

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.56	.225	5.73	.000
N of Valid Cases		85			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 10 shows that, of the 83 times experts used the C6, there were 80 occasions on which the system also used it; this equates to 96.4%. Of the 2 times experts did not use C6, on 2 occasions the system did not use it either; this equates to 100.0%.

Table 11 shows that the kappa coefficient k = .56, p << .05, thus suggesting that there is significantly moderate agreement between experts and the CLC Model on the usage of criterion C6 for accessing allegedly infringing Web resources.

- C7: Downloadable

Table 12. System_use * Experts_use Crosstabulation of C7

			Experts_use		
			No	Yes	Total
System_use	No	Count	25	4	29
		% within System_use	86.2%	13.8%	100.0%
		% within Experts_use	62.5%	8.9%	34.1%
		% of Total	29.4%	4.7%	34.1%
	Yes	Count	15	41	56
		% within System_use	26.8%	73.2%	100.0%
		% within Experts_use	37.5%	91.1%	65.9%
		% of Total	17.6%	48.2%	65.9%
Total		Count	40	45	85
		% within System_use	47.1%	52.9%	100.0%
		% within Experts_use	100.0%	100.0%	100.0%
		% of Total	47.1%	52.9%	100.0%

Table 13. Measurement of agreement on C7

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.54	.089	5.20	.000
N of Valid Cases		85			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 12 shows that, of the 45 times experts used the C7, the system also used it on 41 occasions; this equates to 91.1%. Of the 40 times experts did not use C7, there were 25 occasions on which the system did not use it either; this equates to 62.5%.

Table 13 shows that the kappa coefficient k = .54, p << .05, thus suggesting that there is significantly moderate agreement between experts and the CLC Model regarding the usage of criterion C7 for accessing allegedly infringing Web resources.

- C8 – C9: Link type of online accessing resources/downloadable resources

Table 14. System_use * Experts_use Crosstabulation of C8-C9

			Experts_use		
			No	Yes	Total
System_use	No	Count	14	3	17
		% within System_use	82.4%	17.6%	100.0%
		% within Experts_use	93.3%	4.4%	20.5%
		% of Total	16.9%	3.6%	20.5%
	Yes	Count	1	65	66
		% within System_use	1.5%	98.5%	100.0%
		% within Experts_use	6.7%	95.6%	79.5%
		% of Total	1.2%	78.3%	79.5%
Total		Count	15	68	83
		% within System_use	18.1%	81.9%	100.0%
		% within Experts_use	100.0%	100.0%	100.0%
		% of Total	18.1%	81.9%	100.0%

Table 15. Measurement of agreement on C8-C9

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.85	.075	7.72	.000
N of Valid Cases		83			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 14 shows that, of the 74 times experts used the C8-C9, there were 68 occasions on which the system also used it; this equates to 95.6%. Of the 15 times experts did not use C8-C9, there were 14 occasions on which the system did not use it either; this equates to 93.3%.

Table 15 shows that the kappa coefficient k = .85, p << .05, thus suggesting that there is significantly moderate agreement between experts and the CLC Model regarding the usage of criteria C8-C9 for accessing allegedly infringing Web resources.

- C10: Title of copyright work

Experts stated that they always use this criterion to locate the allegedly infringing content on a webpage. In the CLC Model workflow, this criterion is also always used as the first step to check the existence of the content. While a kappa coefficient value k cannot be obtained in this case, however, it can be concluded that the usage of C10 in the CLC Model was met with a high level of agreement among the experts.

- C11: Performer of copyright work

Similar to C10, while no kappa coefficient value can be calculated, the usage of C11 in the CLC Model was met with a high level of agreement among experts.

- C12: URL suspicion

Experts stated that they always use this criterion to establish whether or not the current webpage looks suspicious. Identical to C10 and C11, while no kappa coefficient value can be calculated, the usage of C12 in the CLC Model was met with a high level of agreement among experts.

5.3.2 Analysis for correlation of infringement scores

To investigate whether the scores generated by the application of the CLC Model (it is called CLC score) agreed with the expert ratings, the Pearson correlation between the score and the experts' rating was calculated (Field 2013). The magnitude of the Pearson correlation coefficient determines the strength of the correlation. Although there are no hard rules for assigning strength of association to particular values, some general guidelines are provided by Cohen (Cohen 1988):

Table 16. Classification of correlation coefficient

Coefficient Value	Strength of Association
0.1 < | r | < 0.3	small correlation
0.3 < | r | < 0.5	medium/moderate correlation
| r | > 0.5	large/strong correlation

where | r | means the absolute value or r (e.g., | r | > .5 means r > .5 and r < -.5).

Table 17 shows the CLC score and the experts' rating for each URL that points to a webpage. Table 18 displays the correlation analysis result. It indicates that the correlation coefficient r = .537, p = 0.003, thus suggesting that when the experts give higher ratings on infringement, the CLC Model similarly gives higher scores. Figure 15 displays a scatter graph of correlation between the CLC score and the experts' rating.

Table 17. CLC score and experts rating for each URL

URL No.	CLC score	Experts rating	URL No.	CLC score	Experts rating	URL No.	CLC score	Experts rating
1	.976	4.7	11	.017	4.7	21	.000	2.3
2	.000	4.7	12	.000	1.7	22	.444	2.7
3	.994	4.0	13	.915	5.0	23	.986	4.3
4	.912	5.0	14	.017	1.3	24	.842	5.0
5	.000	3.7	15	.966	4.3	25	.946	4.7
6	.875	4.0	16	.000	3.7	26	.444	3.3
7	.001	2.7	17	.945	4.3	27	.000	2.3
8	.972	4.3	18	.964	4.3	28	.444	3.0
9	.000	5.0	19	.000	4.0	29	.120	4.3
10	.930	4.7	20	.001	3.3			

Table 18. Correlation between experts' rating and CLC score

		Experts_rating	CLC_score
Experts_rating	Pearson Correlation	1	.537**
	Sig. (2-tailed)		.003
	Sum of Squares and Cross-products	30.128	7.005
	Covariance	1.076	.250
	N	29	29
CLC_score	Pearson Correlation	.537**	1
	Sig. (2-tailed)	.003	
	Sum of Squares and Cross-products	7.005	5.640
	Covariance	.250	.201
	N	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

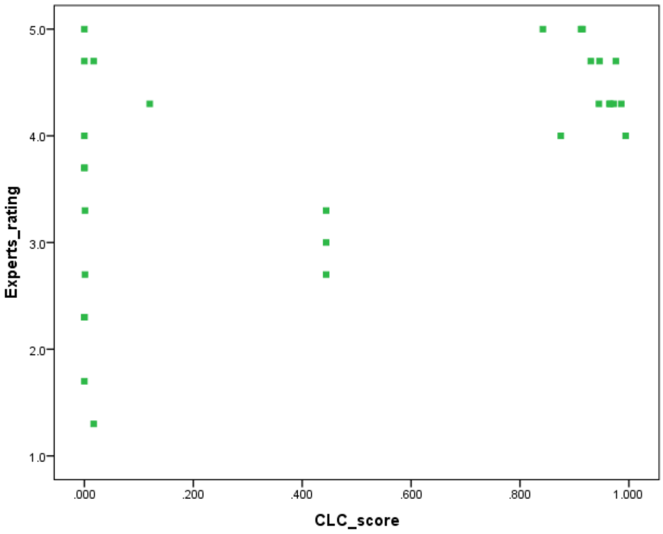


Figure 15. Scatter plot for correlation of infringement score

5.4 Discussion of Validation Result

5.4.1 Discussion of criteria usage

The previous section discussed the results of the analysis for criteria usage; it shows that experts and the CLC Model have a very good agreement on criteria C8-C9 (k=0.85), C10, C11 and C12, as well as good agreement on criteria C3 (k=0.70) and C5 (k=0.68). As explained earlier, experts agreed with the CLC Model on using criteria C10, C11 and C12 every time when accessing allegedly infringing content. With regard to the usage of criteria C3 and C5, it is fairly easy to understand the good agreement; this is because, when a claim is made that a work on a webpage infringes copyright, the assessment process used to investigate this claim is reasonably thorough. The process looks for the work, gains access to the work, and compares it with the original copyright work. Similarly, with regard to the usage of criteria C8-C9, considering where and how the work can be accessible was met with a high level of agreement among experts.

For criteria C4 (k=0.42) and C6 (k=0.56), the result shows that experts and the CLC Model have moderate agreement. This is because some experts thought the decision on whether or not copyright is infringed does not always seem relevant to whether a user interface/function is supplied for accessing the allegedly infringing content. Eventually it is related to the actual content. However, to determine if there is an infringement, it is important to clarify whether the content is de facto accessible; this is because different answers to these criteria will lead to different routes in the score generation algorithm in the CLC Model. The CLC Model aims to figure out the more accurate output of infringement score. For example, the content cannot be accessible in both of the two situations: 1) there is no user interface/function to facilitate the access; 2) there is a user interface/function, but the content cannot be accessed because of broken links. In situation 1), infringement assessment can be terminated immediately and a relatively lower score is supposed to be generated. However, in situation 2), the assessment process will continue going further to check on other criteria in order to generate the final score. These two situations are indicated as R3 and R6 in Figure 12, Chapter 4.

For criterion C7 (k=0.54), experts and the CLC Model have moderate agreement, which is slightly lower than expectation. The agreement expected on the usage of C7 would be very similar to C5 because both of the criteria were related to the content accessibility. The reason for this disparity was the fact that one expert answering the questionnaire claimed to have never used this criterion. She explained that she used other people's computer when she was viewing and examining the webpages, and because she did not want to download anything on another person's computer, she gave a negative answer regarding the usage of C7. However, she did check the download links. As such, it is believed that the level of agreement on the usage of C7 would be similar to C5, which showed good agreement.

In the questionnaire, open questions were asked to encourage experts to add any criteria that they felt should be used. They did not add any criteria. In general, experts have good agreement on the criteria developed in the CLC Model.

5.4.2 Discussion of infringement score

The correlation analysis result shows that the CLC score was significantly correlated with the experts' rating (r = 0.537, df = 27, p= 0.003), and the correlation is strong. A further question is whether this correlation is different from the correlation between experts. In order to answer this question, a further analysis on the correlation between the experts was conducted. Table 19 indicates that the correlation coefficient between Expert 1 and 2 is 0.701, between Expert 1 and 3 it is 0.456, between Expert 1 and 4 it is 0.629, between Expert 2 and 3 it is 0.537, and between Expert 2 and 4 it is 0.559. The average correlation coefficient is 0.576 (r_H). Thus, the question must be asked, is the correlation significantly different from 0.537 (r_O)? The null hypothesis is $r_O = r_H$, which means that the two correlations are not significantly different.

By checking the r to z table, the values of r_O and r_H are converted to normal deviates, z'_O = 0.600 and z'_H = 0.657. The online calculator (Lowry) was used to calculate the standard error of z' for N that is 29:

$$\sigma_{z'} = \sqrt{\frac{1}{\text{N}-3}} = \sqrt{\frac{1}{29-3}} = 0.196$$

Then a normal deviate is calculated according to the following formula:

$$\text{z}= \frac{Z'_O - Z'_H}{\sigma_{z'}} = \frac{0.600 - 0.657}{0.196} = -0.291$$

From the z value, the p can be got p= 0.77 >> 0.05, which is not significant. As such, the null hypothesis is accepted that the two correlations are not significantly different.

Table 19. Correlation between experts

		Expert1_rating	Expert2_rating	Expert3_rating	Expert4_rating
Expert1_rating	Pearson Correlation	1	.701**	.456	.629*
	Sig. (2-tailed)		.000	.101	.016
	N	29	29	14	14
Expert2_rating	Pearson Correlation	.701**	1	.537*	.559*
	Sig. (2-tailed)	.000		.048	.038
	N	29	29	14	14
Expert3_rating	Pearson Correlation	.456	.537*	1	.c
	Sig. (2-tailed)	.101	.048		.
	N	14	14	14	0
Expert4_rating	Pearson Correlation	.629*	.559*	.c	1
	Sig. (2-tailed)	.016	.038	.	
	N	14	14	0	14

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

c. Cannot be computed because at least one of the variables is constant.

The scatter plot in Figure 15 indicates that when the CLC scores are relatively low (0~0.2), the ratings from experts are very dispersive, ranging from 1 to 5. The URLs which had lower CLC scores were further checked, and found that for certain URLs, the scores between experts and the CLC Model are quite different. As presented in Table 17, these URLs are numbers 2, 9, 11, 16, 19, 20 and 29.

For URL 2, the CLC score is 0, much lower than the rating from experts, which is 4.7. For URL 16, the CLC score is 0, still much lower than the rating of 3.7 from the experts. This is because, for URL 2 and 16, the allegedly infringing work cannot be found on the webpage at all. In addition to this, there is no context information such as title and performer of the work. Following the CLC Model workflow and algorithm, the assessment process terminates immediately and the CLC score is 0 because the claimed content does not exist on the webpage. This situation is indicated as R2 in Figure 12. However, experts gave a higher rating because they found many other copyright music work displayed on these webpages. As such, their decision on infringement rating is mostly based on URL suspicion; indeed, as mentioned by an expert in the questionnaire, "the website contains many copyright music which looks very suspicious, should take down". The decision regarding whether or not such music should be taken down can be made by the judgement of URL suspicion. However, as discussed in Chapter 2 regarding Google's practice, this may result in over taken-down. By assessing Web resources, the CLC Model aims to indicate the likelihood of infringement for a specific protected work identified in a notice. As such, no change is made to the CLC Model when assessing URLs such as URL 2 and 16.

For URL 9, the CLC score is 0, much lower compared with the rating from experts, which is 5.0. For URL 19, the CLC score is also 0, still much lower compared with the rating of 4.0 from the experts. This is because, for URL 9 and 19, the music works hosted under the current webpage are in fact cover versions sung by music fans. The performer of the work is therefore not the same. This situation is indicated as R7 in Figure 12. In Chapter 6, details of how to employ an external fingerprint music library to compare music work will be introduced. It is very likely that the music work sung by general fans cannot be found in the fingerprint library, and thus the CLC score will be very low.

For URL 11, the CLC score is 0.017, which is very low compared with the experts' rating. This is because the claimed music work on webpage number 11 is a sample music file which is embedded from SoudCloud. In this case, the final CLC score is the smaller value between audio similarity score and the suspicion value of SoundCloud. This situation is explained as R4 in Figure 12. The smaller value is the level of suspicion for SoundCloud which is 0.017. For URL 20, the situation is exactly the same as URL 11. The difference is that the sample music file is embedded from beatport.com, and the suspicion of beatport is 0.01. In terms of why the experts gave a high rating, this might be because they did not realise that the music work was embedded from legal websites which were very likely to be authorised to host the content.

For URL 29, the music can be downloaded from the current webpage, but the music is not hosted under the current webpage. This situation is indicated as R8 in Figure 12. In the case of URL 29, the music work is hosted on the domain yt-downloader.org. The level of suspicion of this domain according to the Google transparency report is 0.12. The situation in URL 29 is different from URL 11 and 20. For URL 11 and 20, the CLC scores are more accurate than the experts' rating. However, for URL 29, the experts' rating is more accurate because further investigation of yt-downloader.org shows that it is very likely to illegally host music work without the copyright owner's consent, even though its level of suspicion is 0.12 according to Google's data. Thus, the question here is how to guarantee the accuracy of the level of URL suspicion. In this research, this issue is not explored in detail but Google's data is used to reflect it, which works effectively in most cases, such as URL 11 and 20.

5.5 Summary

To validate the CLC Model, four experienced experts working in the area of copyright were invited to participate in this study. Mixed qualitative and quantitative methods were used for expert validation and review. Certain webpages were selected and sent to them for viewing and examination. During this process, a questionnaire was completed by these experts.

Throughout the validation process, two validity elements needed to be confirmed. First, the criteria used in the CLC Model had to be validated. Quantitative data from experts was analysed to investigate the level of agreement on the usage of criteria between experts and the model when different allegedly infringing webpages are assessed. At the same time, qualitative data was obtained to identify any new criteria suggested by experts which were not presented in the CLC Model. Second, the infringement score generated by the CLC Model also had to be validated. Experts' ratings of infringement on webpages were compared with the CLC scores to identify the relationship between them.

The analysis results on level of agreement show that experts have good agreement on the usage of criteria developed in the CLC Model. They felt that the criteria in the CLC Model were comprehensive and they did not suggest new criteria. The correlation analysis result shows that the CLC score was significantly correlated with the experts' rating, and the correlation is strong. For a few webpages, when experts have different opinions on infringement rating from the CLC Model, the CLC Model will be adapted to these situations.

In the next chapter, how the CLC Model can be automated will be explored, and an automation system will be implemented on the basis of the CLC Model.

Chapter 6 CLC Automation System

As discussed in Chapter 4, the CLC Model aims to verify allegedly copyright infringing content on webpages in an automatic manner. In this chapter, the implementation of the CLC Model in an automated system is investigated in order to help assess Web resources in notice-and-take-down procedures. Firstly, the Web technologies implementation related to the CLC Model are analysed. Secondly, how the assessment for each criterion is automated is discussed, following which discussion switches to an automation system which connects each criterion to produce analytic results of allegedly infringing Web resources. Finally, concerns about the automation system are summarised.

6.1 Web Technologies Related to the CLC Model

6.1.1 Webpage delivery under Web architecture

Investigating the implementation of Web technology is very important when it comes to understanding the degree to which each criterion can be automated. All webpages which were flagged up for potentially containing copyright infringement content must follow the Web architecture; as such, it is critical to understand the workflow of requesting, presenting (or rendering) and interacting with webpages. As stated in Chapter 4, most of the criteria (C1, C4, C5, C6, C7, C8, C9 and C12) can only be analysed within the context of Web architecture.

The following Figure 16 and Figure 17, which is adapted from https://github.com/mcrayne/MeteorCookbookGitbook/blob/master/Cookbook%20Conventions.md explain how, in a modern Web browser, the webpage is delivered and rendered following a request in a browser. When an allegedly infringing URL is examined by typing in or copying & pasting the URL to the browser address bar, the browser sends a HTTP GET request to the Web server. According to the HTTP 1.1 specification ("Hypertext Transfer Protocol -- HTTP/1.1: 5 Request"), the HTTP request contains:

1. Request line: including the method (GET, POST, PUT, etc.) to be applied to a resource, the identifier of the resource (URI) and the protocol version in use.
2. Request headers: indicating, for example, the host of the server, the user-agent (browser), what format or language of the resource is requested, etc.
3. Request body: the body could include request parameters in some format, plain text content or binary content.

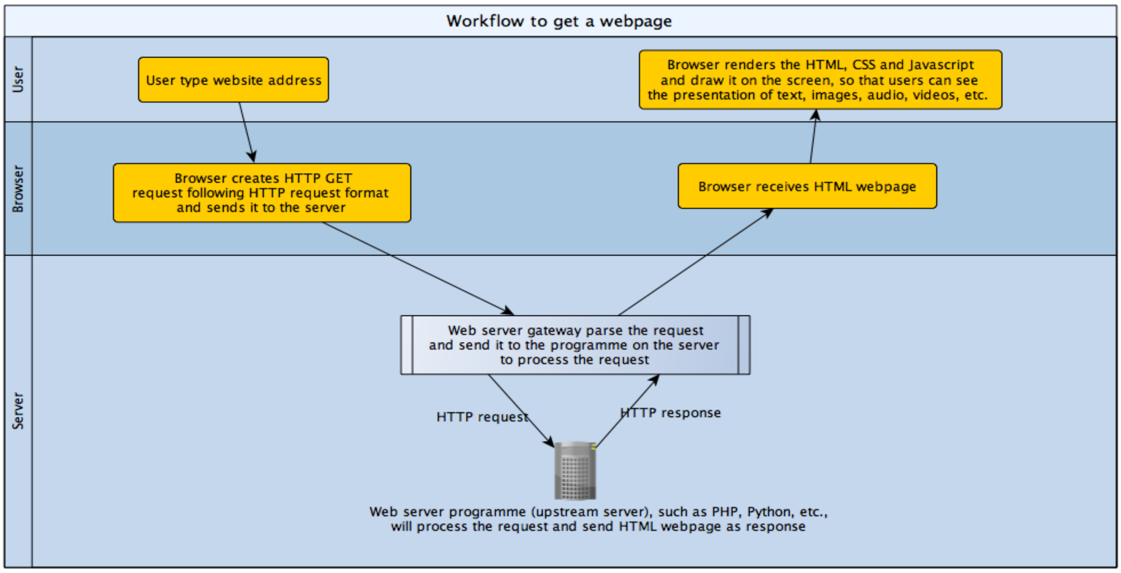

Workflow to get a webpage
User
Browser
Server
User type website address
Browser renders the HTML, CSS and Javascript and draw it on the screen, so that users can see the presentation of text, images, audio, videos, etc.
Browser creates HTTP GET request following HTTP request format and sends it to the server
Browser receives HTML webpage
Web server gateway parse the request and send it to the programme on the server to process the request
HTTP request
HTTP response
Web server programme (upstream server), such as PHP, Python, etc., will process the request and send HTML webpage as response

Figure 16. Webpage delivery

The information included in the HTTP request will be received by the server, following which the server can decide how to process the request and return the result. In the case of this research, when a webpage is asked, which is a document or resource stored on the server, the server will deliver the correct webpage according to the information contained in the HTTP GET request. During this process, the Web server can be divided into two layers: the HTTP gateway server and the upstream server. The HTTP gateway server receives the HTTP request and passes it over to the upstream server to process the request. Major open source HTTP gateway servers include the Apache HTTP Server and Nginx. The upstream server is usually implemented by means of specific technologies such as Java, Python, .Net, Nodejs, etc. Different from the HTTP gateway server, the upstream server will implement the business login to process the request and send the request. In the case of this research, the request usually asks for a webpage, so the server will send the HTML document which corresponds to the requested URL.

At this point a fact that must be emphasised is the gateway and upstream servers are not necessarily separated. The Apache HTTP server and Nginx can serve static webpages, which are stored in some directory of the server. Likewise, Tomcat can also be configured to serve the Java Servlet website directly without the gateway server. However, it is a common practice in modern Web applications for a HTTP gateway server to be configured to handle HTTP redirection, load balancing and other jobs that are not directly related to the business logic, while the upstream server will focus on the implementation of the business logic.

6.1.2 Webpage rendering and user interaction

Some of the criteria in the CLC Model describe how the infringement assesement process is affected by the UI and interaction on the webpage. As such, it is necessary to briefly introduce how the webpage is rendered through a HTML document and how a user can interact with it. Technically, all the elements on the webpage can be manipulated by JavaScript, and the modern webpage has been developed to such a stage that the implementation of the interactions can be very dynamic and flexible. Thus, it is critical to have an overview of the possible technology implementations of each criterion so that such technology implementations can be linked back to the legal cases mentioned in Chapter 2.

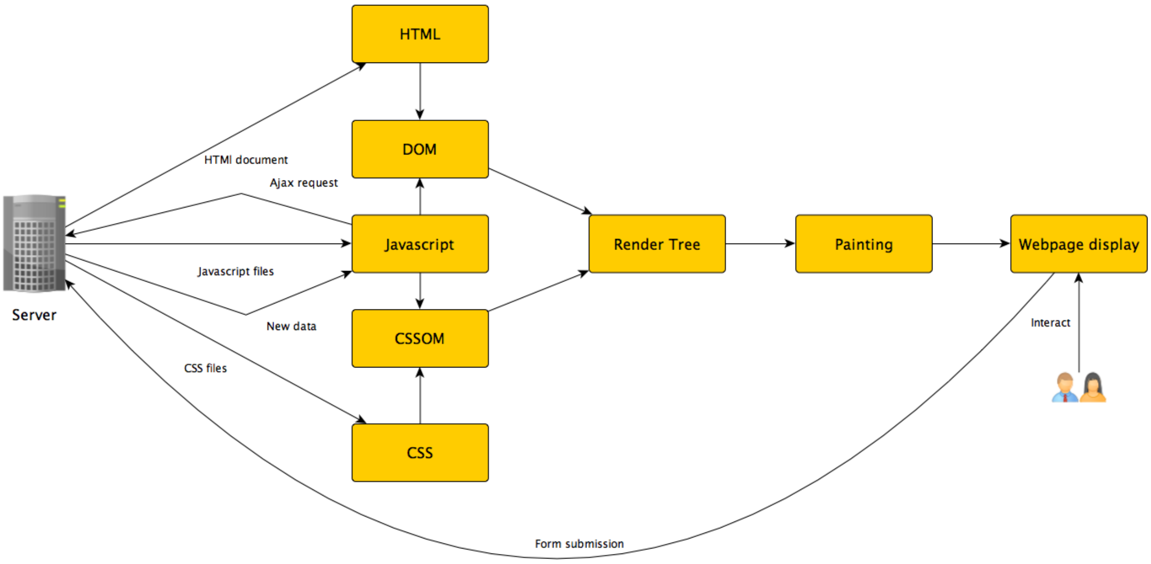


Figure 17. Webpage rendering and user interaction

When the HTML document is sent back to the browser, the browser will start to render the webpage. During this process, in addition to the HTML markups included in the HTML document, the browser will also send more requests to the Web server asking for CSS and JavaScript files included in the HTML document, which are necessary to render the webpage (see the following figure).

During the rendering process, the browser will construct the DOM (Document Object Model) structure according to the HTML markups. At the same time, the CSS will also be parsed into the CSS Object

Model (CSSOM). The DOM and CSSOM together decide the layout of the webpage and how each element will look. Finally, the browser will paint the rendering result on a webpage to be displayed to users. JavaScript can programmatically change both DOM and CSSOM in this process, but it will not affect the Render Tree and Painting directly. This process might be slightly different from browser to browser, but the general steps are the same.

After the rendering finishes, users will be able to view and interact with the webpage through the browser. Typical interactions include a user clicking a certain element on the webpage, or using the keyboard to fill in a form, etc.

Some of the interactions will trigger JavaScript on the client side to modify the webpage structure without contacting the server or making any HTTP calls to handle the interaction such as, for example, popping up an alert window on the webpage, displaying a dropdown menu, or changing the font-size by clicking a button, etc. On the other hand, in the take-down notices examined, there are several kinds of interactions that will trigger another HTTP request to the servers as a direct or indirect result:

1. Form submission: user fills in a form in <form> HTML document and submits to the server.
2. Ajax HTTP calls: user clicks a button or some other interactions that will send an asynchroised JavaScript HTTP call to the server and the server will return new data for JavaScript to re-render the page.
3. Webpage redirection: user clicks an anchor (<a> tag in HTML document) or some other interactions that will trigger the relocation of the current URL address. This will usually result in the opening of a completely new page.
4. Multimedia streaming: user interacts with the video or audio player embedded in the webpage and triggers the downloading of multimedia files through the browser's native <video> or <audio> player, or certain plugins that can play multimedia resources on the browser (for example, Flash Player).

6.2 Development of the CLC Automation System

6.2.1 Criteria automation

In this section, how Web technologies can be implemented to automate the 12 criteria in the CLC Model is discussed.

- **C1: URL accessibility**.

From a user's point of view in Figure 16, a URL is not accessible means that when the user tries to open the URL from a browser, the webpage will not be displayed. If the URL is not accessible the user sees an error page which may be provided by the server. Technically, the accessibility of the URL can be decided by checking HTTP code status ("Hypertext Transfer Protocol -- HTTP/1.1: 10 Status Code Definitions") when making an HTTP Get call to the URL.

The following table indicates the relationship between HTTP code that reflect C1 according to the study in Chapter 3. Not all the related 4xx and 5xx HTTP codes are listed in the table because some of them have not appeared in the experiment. Some technical explanation are also included on common causes of each HTTP code in the context of copyright infringement claim.

The following table lists that there are a number of reasons for an inaccessible URL, and generally people cannot further use the content of the Web page containing the inaccessible URL to decide whether there may be content infringement on the requested webpage.

To automatically detect the result of C1, regular HTTP client can be used to perform a request to the URL and check the response code. For example, curl is a command line tool in Linux-based operation system to make HTTP call to any server. There are also many UI based tools to help users make HTTP request, such as Postman.

Table 20. HTTP error code

HTTP Code	Code text	Explanation
400	Bad request	The URL in the claim does not form a valid HTTP request to the server.
401	Unauthorized	It indicates the Authentication header is missing in the HTTP request, or if the Authentication credential has been included, the credential has been refused by the server. In the experiment, if the claimed resources are restricted to login users only and the claimer(s) simply copy the URL, when other people try to access the webpage, they will get this HTTP Code. This is simply because the new and fresh HTTP request does not contain any token in the WWW-Authenticate header field.
403	Forbidden	The request is received and identified by the server. However the server refuses response based on the current identification. Different from 401, 403 tells the browser that the server has identified the requester, but the resource is not forbidden to be accessed by this user. In the experiment, the reason for the appearance of 403 is similar to 401, which is the users only provide a URL in the claim without any login or authentication information.
404	Not Found	This is the most common case for a URL not accessible according to the experiment in Chapter 3. The website of the claimed URL does not exist anymore. It might be temporal server crash, or the domain name doesn't exist or hasn't been pointed to any server. Many URLs in the experiment have been revised after a while, and many of them have fallen into this code because the website has been shut down or blocked by the OSP.
500	Internal server error	The server cannot fulfil the request and returns an error page.
502	Bad Gateway	The backend server, which should provide a response to the request is not responding correctly, while the frontend server which acts as a gateway cannot provide a valid response. The cause of this response may be similar to 404 in some cases in that the server programme cannot deliver the webpage.
503	Service unavailable	The server is unable to handle the request at the moment. It might be the server is overloading, or it is under maintenance.
504	Gateway timeout	On passing the request URL to the upstream server, the gateway or proxy cannot get a prompt response from it.

- **C2: Content existence.**

Criterion C2 co-works with C10 and C11 and, strictly speaking, the automation of this criterion involves the automation of C4, C5, C6 and C7. To decide whether the title or another content exists, text search and comparison functions similar to the Find function (Ctrl+F as short cut key combination) in the browser can be mainly used. The first step of search is to extract text information from the webpage. Then an algorithm is used to match the title or performer text specified in the take down notice with the text extracted from the HTML document.

The text usually can be extracted from the HTML markups directly after the webpage is rendered within the browser. There are many text extraction tools that can be used for the first step, such as Scraper, x-ray and IBM Watson's Document Conversion service. For some websites, this is sufficient because the HTML document renders completely including the necessary information. However, if the webpage contains Ajax programme, then people may need to wait a couple of seconds until the Ajax call has finished and the new DOM and CSSOM has been fully rendered. There are several HTML markups which are particularly used to present text information, and they may be not visible in the browser:

1. **<title> tag in the <head>.** The title tag is usually required in the HTML document. The page title will not be visible as the body content of the webpage, but it will be displayed as the title in the browser toolbar when you open a webpage, and it will be displayed as the title in the search result from search engine.
2. **<meta> tag in the <head> with keywords, name, author and description information**. The metadata provides important information about the webpage, and it has been widely used for Search Engine optimisation (SEO). Many Content Management System (CMS), like WordPress and Drupal, have automatically included that information when a webpage is published. So if a webpage contains copyright infringement information about a music record, the title or performer of the record might be included in the meta text.

3. **Heading tags such as <h1>, <h2>, etc**. The heading tags are similar to the chapters or sections of a book, which list the structure of the book.
4. **HTML text formatting tags, such as <b>, <i>, <em>, etc**. The text format tags in HTML usually means the text has a special meaning, and it should be displayed differently with other text. However, the tags only indicate the styles of the text instead of the semantic meanings.
5. **Text for HTML anchor <a>.** The text in the anchor usually indicates the behaviour or the main content if you click the anchor and follow the link. In some cases of the experiment mentioned in Chapter 3, the text of anchors contains the title or performer of the music.
6. **Form elements.** In HTML, the forms are used to collect user input or ask the user to select an option from existing content. The elements that could contain title and performer text information mainly include the label of radio button input (<input type="radio">) and the <select> and <option> tags, which will be displayed as a dropdown list to users.
7. **alt and title attributes.** These two attributes are originally designed to improve accessibility for <img> tags and other tags which needs further explanation of the content. Since it contains useful alternative information to describe some visual elements on the webpage, it could include the music title and performer's information.
8. **Other elements can contain text directly, such as <p>, <div> and <span>.**

One thing that needs to be emphasised is that all the tags are not necessarily individual to each other on a webpage. It is quite likely that many tags are embedded in each other. For example, the <form> tag may also contain <p> and <div> for its form elements, vice visa.

A webpage scrapper can be developed to extract the text information based on the analysis above. This work can be developed in two steps. Firstly, the scrapper requests the HTML document from the URL and analyse the document. For some websites, this step is enough as the HTML document is a fully rendered webpage, which includes all the necessary information. However, some websites use Ajax technology to request additional data after the HTML document has been delivered, which will result in the delay of re-rendering the webpage. In this case, people need to wait until the Ajax calls are finished and the webpage is fully rendered before people can start analysing the content.

Unfortunately, neither curl nor Postman can utilise this function easily. So Selenium based Web Browser Automation technology will be applied to analyse the webpage with Ajax calls. Selenium is a fake browser that can mockup the behaviours of a Web browser without actually opening the webpage in a browser. Then Developers could interact with the webpage through programmes to pretend a button is clicked or a form has been filled. In this way, people can programmatically get the page information without manually opening the webpage and find out where is the information.

- **C3: Work (Audio) similarity.**

There are many open source and commercial music repositories that can be used to compare the audio file hosted on the infringement URL to an 'official' repository of music. MusicBrainz's open source Fingerprinting service called AcoustID is used. The basic idea of such service is to extract the identical features of music from its original record and save the fingerprints into a database for comparison. Instead of comparing the audio file byte by byte, the fingerprint comparison is much quicker. Developers just need to download the file from the claimed URL and upload it to AcoustID, which will give developers feedback within a couple of seconds whether a match has been found and what is the similarity in percentage.

Even though the service is free and has been maintained by a reliable community, there are some shortcomings when using the service in the CLC automation system.

1. Many music records are still missing in AcoustID, especially in a language other than English and also the records from less famous singers. This mean if a high possibility match cannot be found from AcoustID, people can't say there is no copyright infringement, it may simply because the original record is missing in the repository, or the downloaded file is recorded from another source, for example, music radio.

2. In some take-down notices, the claimed infringement is a sample of the original copyright work. In this case, the fingerprinting is incomplete and the sample cannot adequately be compared with the full original.

It's very difficult to detect 'partial copy'. For example, if claimer(s) claim that their music has been misused within a longer or short record, the fingerprinting will not work in this case.

- **C4: Online access.**

This criterion is about whether the webpage provides access that can lead users to play an audio file online. The technology implementation of actually playing an audio file will be described in C5. If users see a "play" button on the webpage, usually it is one of the following implementations:

1. The play button of the HTML native audio and video player. The standard player includes buttons, such as play, volume control, etc., to control the play of the audio or video. There is usually a progress bar to indicate the duration and the remaining time of the audio or video. The play button is usually an icon with a triangle shape arrowhead or similar. The look and feel of the play button can be customised by JavaScript and CSS.
2. The play button of video or audio player implemented by a specific technology, such as Flash, Silverlight or other plugin for browsers. Before HTML5, browsers usually needed to install plugins to handle the play of video or audio files. A player can be embedded into the webpage and displays player controls.
3. An HTML button or link that will trigger JavaScript to play audio or video. This button is different from the play button provided by HTML's native audio or video tag, where the browser handles the default behaviour of the button. Here, the button or link triggers JavaScript functions to control the video or audio using HTML5 Media Element API. In some cases, the JavaScript can also control play by video/audio plugins.

This section only discusses the visual clues of the "Online access" function. The technology implementation of replaying an audio will be described in C5.

- **C5: Online playable.**

Following C4 about the visual clues of audio playing, this section explains the technologies that are commonly used on a webpage to actually play video or audio. It must be emphasised that, even though video or audio can be played, it does necessarily mean the webpage 'owns' the audio or video file. This is explained in more detail in C8.

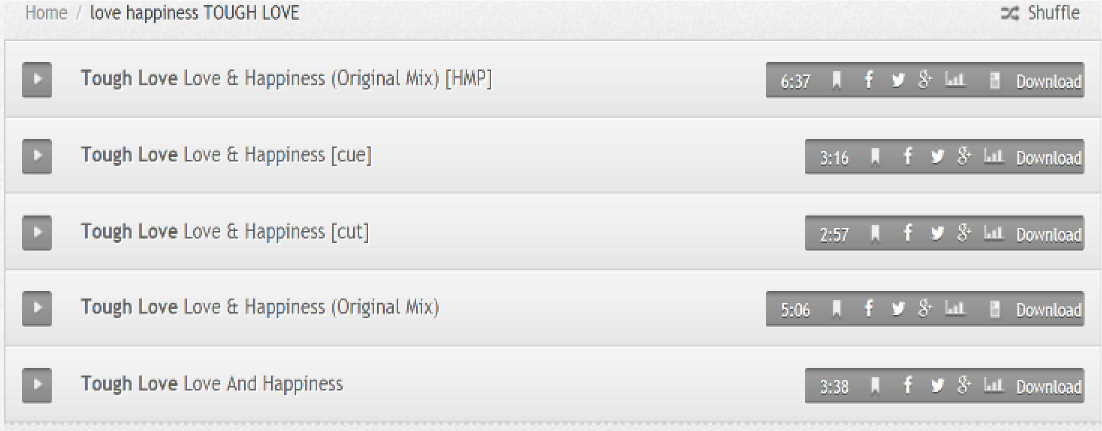
A webpage that can play audio or video implements one or more of the following techniques.

1. Native HTML <audio> or <video> tags: to specify which file to be replayed, the webpage owner needs to specify the 'src' attribute of <source> tag within <audio> or <video> like followings.

```
<video width="320" height="240" controls>
        <source src="movie.mp4" type="video/mp4">
                <source src="movie.ogg" type="video/ogg">
</video>
```

2. Video and audio playing from plugins such as Flash and Silverlight. This type of implementation has been largely discarded in modern webpages because it brings compatibility problem across browsers. However, from the previous analysis, some copy infringement websites still use legacy plugins to play audio. Technically, the plugin is invoked in an HTML document as an <embed> or <object> tag.
3. Video or audio play triggered by JavaScript. In this case, there are no <video>, <audio> or <embed> tags, and the JavaScript directly provides audio or video play. This may be automatic when opening the webpage, or the user clicks a button to trigger the play of a certain file.

The last technique makes the automated detection of C5 very difficult. Unlike the previous two techniques where developers can automatically locate the actual file corresponding to a player, the play control of JavaScript is totally dependent on how the webpage developer programs this function. Figure 18 can explain this situation. Visually, a play button can be related to the playback of the mp3 file in the same row, but an automated process is not able to "see" such visual clues.

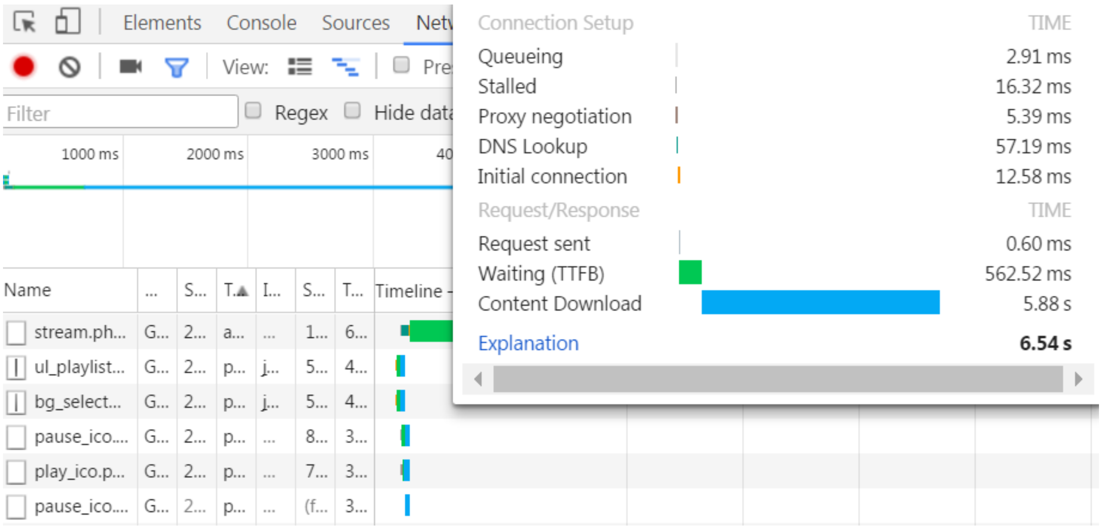


Home / love happiness TOUGH LOVE

Shuffle

▶	Tough Love Love & Happiness (Original Mix) [HMP]	6:37 Download
▶	Tough Love Love & Happiness [cue]	3:16 Download
▶	Tough Love Love & Happiness [cut]	2:57 Download
▶	Tough Love Love & Happiness (Original Mix)	5:06 Download
▶	Tough Love Love And Happiness	3:38 Download

Figure 18. My Free MP3 example of JavaScript controlled audio playback



Name	...	S...	T.▲	I...	S...	T...	Timeline -
stream.ph...	G...	2...	a...	...	1...	6...	
ul_playlist...	G...	2...	p...	j...	5...	4...	
bg_select...	G...	2...	p...	j...	5...	4...	
pause_ico....	G...	2...	p...	...	8...	3...	
play_ico.p...	G...	2...	p...	...	7...	3...	
pause_ico....	G...	2...	p...	...	(f...	3...	

Connection Setup	TIME
Queueing	2.91 ms
Stalled	16.32 ms
Proxy negotiation	5.39 ms
DNS Lookup	57.19 ms
Initial connection	12.58 ms
Request/Response	**TIME**
Request sent	0.60 ms
Waiting (TTFB)	562.52 ms
Content Download	5.88 s
Explanation	**6.54 s**

Figure 19. Network traffic when playback a streamed audio

To determine if audio or video is played after being triggered, BrowserProxy is used to monitor the network traffic. This shows whether an audio or video file is actually sent from a remote server. Using the example of Figure 18, clicking on the first song shows (Figure 19) an audio file with type 'audio/mpeg' being requested from 'stream.php.' Monitoring the network after the play button is clicked shows whether or not the file was actually streamed. The implementation of using BrowserProxy to monitor the network traffic is similar as the networking monitoring functions in Google Chrome's debug mode, where the traffic will be classified to XHR, JS, CSS, Media, etc. Whether there are downloaded packages belong to Media category will be monitored when a suspect play button is clicked.

- **C6: Download access.**

A recording can be made available on a webpage in two ways: (1) the file is played online with an audio or video player, but it is not downloadable or downloaded (i.e. it is streamed); or (2) the file can be downloaded and played offline. Technically, the two ways use completely different Web technologies implementation, so it is very important to look at them separately.

C6 is the starting point for the file downloading criteria group, and it mainly describes whether a file download function is available or visible on the webpage, which may lead to the actual download of an audio file. Different from how the file is downloaded (C7 and C9), C6 deals with the visual interface on the website that gives users access to a download through one or more steps (or user interactions, such as clicks). A download access function is commonly implemented in one of three ways:

1. An HTML button <button> with "Download" a text. Clicking the button submits an HTTP request to the server to initiate file download.
2. An HTML anchor or link. This points to a new URL that initiates file download.
3. Making an HTML element clickable and triggering JavaScript to download the file. The JavaScript code either relocates the current window to the file or submits a request to the server to download the file.

While the HTML button and HTML anchor are completely different components, modern websites sometimes use CSS and JavaScript to make a button seem like a link, and vice visa. For example, Bootstrap, a widely used CSS and JavaScript framework, defines a 'btn-link' to visually change a button to look like a link. An image on the HTML page can also be an anchor by adding <img> tag within <a> tag. So without inspecting the HTML code directly sometimes it can be difficult to decide whether the element corresponding to a download request is a button or a link. However, the code contains text or graphic information indicating that, by clicking the element, a download will be requested, so similar technology to C2 implements the automatic detection of C6.

- **C7: Downloadable.**

If the download access function is available on the webpage, a user may follow the download instructions or indications, and this will either lead to an actual download of the music file, or failure to download. There might be a possibility that the file cannot be downloaded directly and will become downloadable after a few steps, such as viewing ads, or be redirected to some external website. But the technics to enable the download are roughly the same, which is through specifying HTTP Content-Disposition Header as an attachment. The HTTP Content-Disposition Header indicates how the downloaded content should be treated. Possible values are inline, as an attachment, or as a named attachment.

The HTTP Content-Disposition Header indicates how the response content should be treated. Possible values are:

1. Content-Disposition: inline
2. Content-Disposition: attachment
3. Content-Disposition: attachment; filename="somefile.mp3 "

'inline' means the content should be rendered within the current webpage, while 'attachment' means the files needs to be downloaded. This header is given by the server and the default action in major browsers are quite similar.

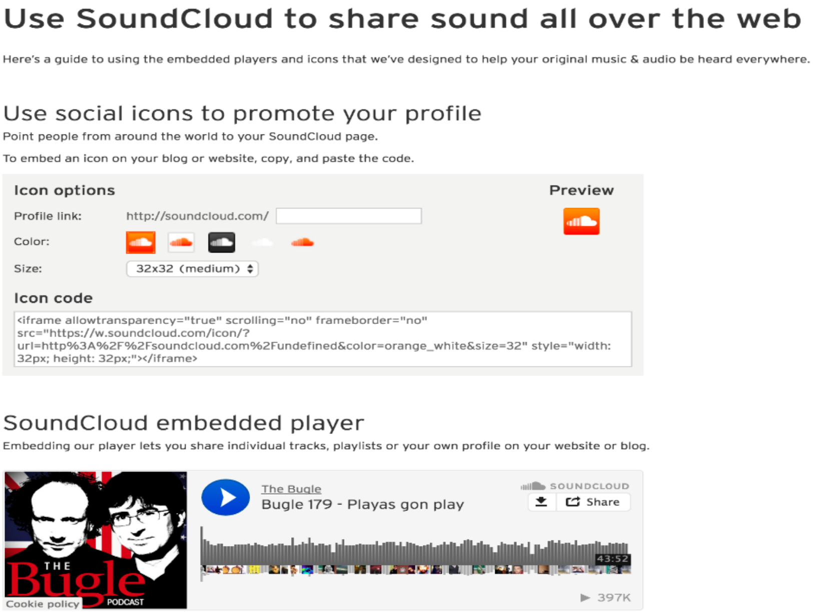
Postman and curl can be used to detect if a file download will happen for a given URL. However, it will be very difficult to automatically and correctly detect every step that leads to a download. Usually, there are many buttons on the webpage, as well as the page that directly follows the download link or button. So developers need to try recursively which button or link can lead to the final download. Again, this theoretically can be implemented by Selenium, but the accuracy may not be satisfactory. Developers need to define a maximum depth of following the button or links and a programme to detect which button or link has been clicked is also needed.

- **C8: Link type of online accessing resources.**

As discussed in C5, there are many technical ways to play audio and video online. From the visual information on the webpage, however, it is usually very difficult to tell which techniques have been used for playback and to tell where the file comes from. A website owner could source the streaming file from a local host or from an external link. As for C6, the network traffic needs to be monitored to decide the type of the link, that is, whether the file is hosted on the current domain, or is embedded from another domain.

Except for the type of player discussed in C4, a popular video or audio embedding technology is to use the <iframe> HTML tag, which displays information from another website inside the current website. The URL of the other website is given as the value for the 'src' attribute. These two websites can be in different locations and managed by different owners. It is very difficult, and usually impossible, to tell if any component on the website is delivered through an iframe without inspecting the code of the HTML document.

An <iframe> is widely used to provide social features on a website. For example, the Facebook "Like" button on many websites uses an iframe to deliver the button content (image, look and feel, and the id of the liked resource) from Facebook, so the button is not managed by the owner of the website, and it is only a reference. Another example closely related to music copyright infringement is an iframe embedded player from a multimedia sharing website such as YouTube, Vimeo, or SoundCloud (Figure 20). Encountering a player within an iframe makes it quite likely the online access is given by an embedded link.



Use SoundCloud to share sound all over the web

Here's a guide to using the embedded players and icons that we've designed to help your original music & audio be heard everywhere.

Use social icons to promote your profile

Point people from around the world to your SoundCloud page.

To embed an icon on your blog or website, copy, and paste the code.

Icon options

Profile link: http://soundcloud.com/

Color:

Size: 32x32 (medium)

Icon code

```
<iframe allowtransparency="true" scrolling="no" frameborder="no"
src="https://w.soundcloud.com/icon/?
url=http%3A%2F%2Fsoundcloud.com%2Fundefined&color=orange_white&size=32" style="width:
32px; height: 32px;"></iframe>
```

SoundCloud embedded player

Embedding our player lets you share individual tracks, playlists or your own profile on your website or blog.

The Bugle

Bugle 179 - Playas gon play

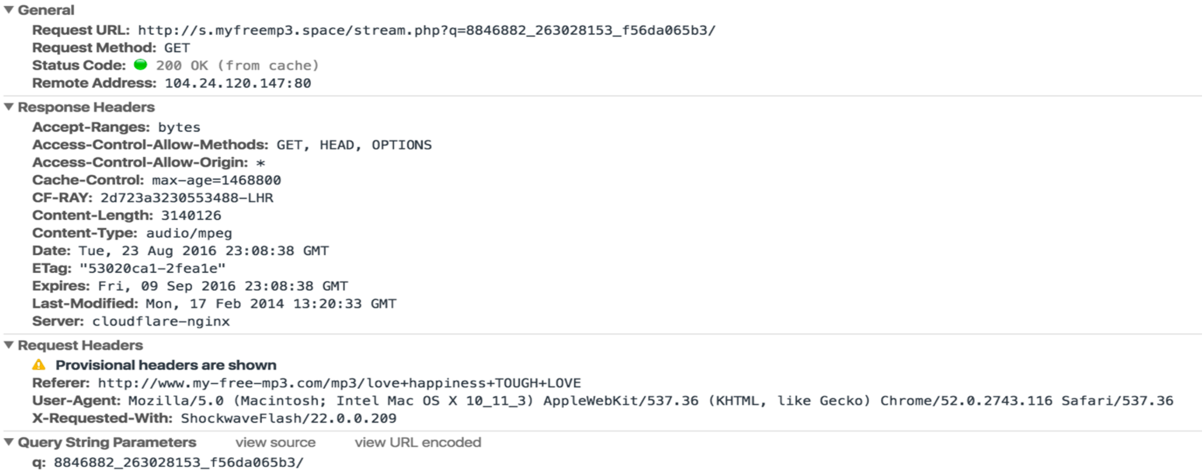
Figure 20. SoundCloud iframe embed player

- **C9: Link type of downloadable resources.**

Following C6 and C7, if an audio file can be downloaded, whether the file is hosted on the current domain or on another domain is needed to be examined. This is done by looking at the HTTP response of the file download, especially the request URL and the remote IP address for the request. Usually, there are two major categories:

1. The file is hosted on the same domain of the claimed URL. In this situation, it is clear that the website service provider should be responsible for the content of the file.
2. The file is a simple link or streaming address and the content is actually hosted on another domain. In this situation, it is not clear whether the file is within the control of the current website provider. Figure 21 shows an example of this case. The MP3 file is streamed from http://s.myfreemp3.space with IP address 104.24.120.147. Even though myfreemp3.space seems similar to the current website, my-free-mp3.com, whether they are managed by the same provider requires further investigation.

Further investigation is also required if content is delivered through the Content Delivery Network, which is a globally distributed server proxy to deliver files faster in different regions, especially for large multimedia files. In this case, the IP address could be masked deliberately by the publisher in order to hide the real IP address.



```
▼ General
    Request URL: http://s.myfreemp3.space/stream.php?q=8846882_263028153_f56da065b3/
    Request Method: GET
    Status Code: ● 200 OK (from cache)
    Remote Address: 104.24.120.147:80
▼ Response Headers
    Accept-Ranges: bytes
    Access-Control-Allow-Methods: GET, HEAD, OPTIONS
    Access-Control-Allow-Origin: *
    Cache-Control: max-age=1468800
    CF-RAY: 2d723a3230553488-LHR
    Content-Length: 3140126
    Content-Type: audio/mpeg
    Date: Tue, 23 Aug 2016 23:08:38 GMT
    ETag: "53020ca1-2fea1e"
    Expires: Fri, 09 Sep 2016 23:08:38 GMT
    Last-Modified: Mon, 17 Feb 2014 13:20:33 GMT
    Server: cloudflare-nginx
▼ Request Headers
    ⚠ Provisional headers are shown
    Referer: http://www.my-free-mp3.com/mp3/love+happiness+TOUGH+LOVE
    User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/52.0.2743.116 Safari/537.36
    X-Requested-With: ShockwaveFlash/22.0.0.209
▼ Query String Parameters      view source      view URL encoded
    q: 8846882_263028153_f56da065b3/
```

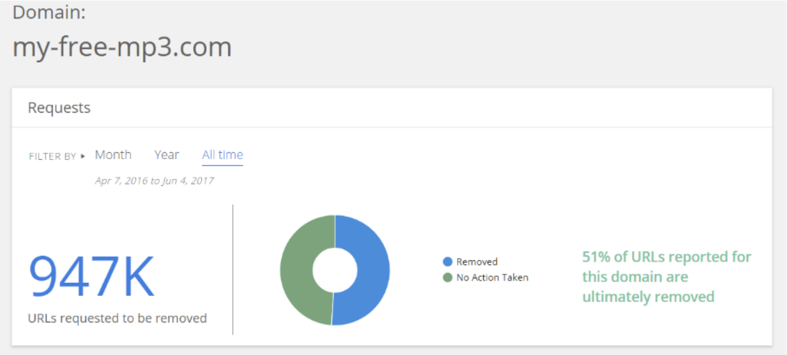
Figure 21. MP3 file hosted from an external address

- **C10 and C11: Matching of the title of copyright work and matching of the performer of copyright work**.

The technology analysis of checking these two criteria is discussed in C2.

- **C12: URL suspicion**.

The 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 the percentage of URLs for that domain which are ultimately removed. The higher the percentage, the higher the URL suspicion value to reflect the likelihood of infringement. Figure 22 shows an example from Google's data. Here it was found that 51% of URLs reported for this domain were ultimately removed by Google. Therefore, 51% is used as an indication of the level of suspicion for domain my-free-mp3.com in the CLC Model.



Domain:

my-free-mp3.com

Requests

51% of URLs reported for this domain are ultimately removed

Figure 22. Example of indication for level of URL suspicion

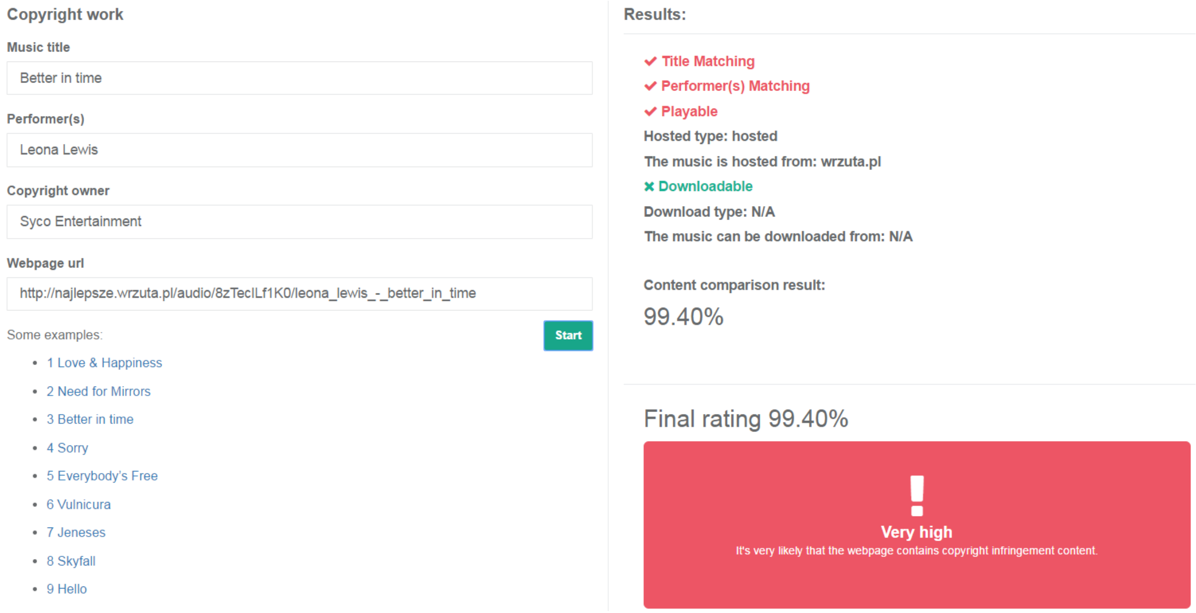
While the correlation between experts' ratings and the system's score is significant and reasonably substantial, it may be that a useful next step is gathering labels of more URLs from experts and developing machine learning algorithms to categorise the Web pages; indeed, this could improve system and expert agreement. However, obtaining labelled data from experts and using machine learning methods are not the main concerns of this research. Firstly, there are major difficulties in gathering more labelled data because most of the suspected infringement webpages are not stable, and can be taken offline quite quickly. Secondly, as mentioned in Chapter 2 and Chapter 3, requests are made for hundreds of thousands of URLs to be viewed and examined for take-down every day by Google. Even if machine learning techniques are going to be used, the real technical challenge is how to efficiently extract feature values from such a large number of webpages. This section proposes crucial steps and methods for automatic feature value extraction which will provide valid data that can be used later as training data in machine learning.

6.2.2 Implementation of automation system

The previous section investigated the extent to which each criterion in the CLC Model can be automated and potential technologies that can be used to implement the automation. Table 21 summarises the automation technologies used in the implementation process.

Table 21. Techniques used in the implementation process

Criterion	Degree of automation	Techniques/Tools used
C1	Fully	HTTP error code
C2, C10, C11	Fully	Web scrapper, Selenium based Web Browser Automation
C3	Fully	AcoustID
C4	Partly	Human intervention to locate the claimed element in HTML, particularly in the cases of JavaScript-triggered music play.
C5	Partly	BrowserProxy. Human intervention in the cases of JavaScript-triggered music play.
C6	Partly	Human intervention to locate the claimed element in HTML, particularly in the cases of JavaScript-triggered music download.
C7	Partly	curl, Postman Human intervention in the cases of JavaScript-triggered music download.
C8	Partly	Human intervention to locate the claimed element in HTML.
C9	Partly	Human intervention to locate the claimed element in HTML.
C12	Fully	Google transparency data.



Copyright work

Music title

Better in time

Performer(s)

Leona Lewis

Copyright owner

Syco Entertainment

Webpage url

http://najlepsze.wrzuta.pl/audio/8zTecILf1K0/leona_lewis_-_better_in_time

Some examples:

- 1 Love & Happiness
- 2 Need for Mirrors
- 3 Better in time
- 4 Sorry
- 5 Everybody's Free
- 6 Vulnicura
- 7 Jeneses
- 8 Skyfall
- 9 Hello

Results:

✓ Title Matching
✓ Performer(s) Matching
✓ Playable
Hosted type: hosted
The music is hosted from: wrzuta.pl
✗ Downloadable
Download type: N/A
The music can be downloaded from: N/A

Content comparison result:

99.40%

Final rating 99.40%

Very high

It's very likely that the webpage contains copyright infringement content.

Figure 23. Example of output from the CLC automation system

Information regarding allegedly infringing work, such as title, performer of the work and a URL that locates the work will be given to the system as input. Following this, the system will automatically examine and analyse the webpage to which the URL points, according to the criteria and workflow defined in the CLC Model. Answers related to each criterion will be given by the system. For example, if a URL is not accessible because the system gets a 404 error from a HTTP request to the webpage, the output of the system will be that the URL cannot be accessed. If a URL can be accessed, further criteria will be checked following the order of workflow defined in the CLC Model. The output of the system will take the form of detailed analytic results of each criterion and eventually an infringement score, which will be presented to users through the Web user interface. Figure 23 presents an example of the output from the CLC automation system and the working prototype of the system is located through http://clc-model.demoapps.me/#/.

6.3 Summary

For each criterion in the CLC Model, the implementation of the background technical elements needed to automate each criterion is investigated; this is followed by the development of an automatic system to dynamically apply the CLC Model and thus assess Web resources. The output of the system is a series of facts regarding the allegedly infringing Web resources and a score to indicate the likelihood of infringement with a view to supporting the decision-making process. In the CLC Model, it is difficult to fully automate all of the criteria. Given the variety and fast development of Web technologies used to present a webpage, people need more automatic and evolving mechanisms to detect the content and different components on the webpages. In addition to using the webpage information extraction and monitoring technologies proposed in Section 6.2, computer vision and machine learning technologies can also be used in future work to recognise the existence of certain Web components in the CLC Model.

The CLC automation system could be used by online service providers, such as search engine providers and index service providers. In the future, it could also be adopted 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 automatic take-down notices.

Chapter 7 Evaluation of the CLC Automation System

Chapter 6 investigated the potential Web technologies that can be used to automate each criterion in the CLC Model and discussed the development of an automation system. Following on, this chapter describes a user evaluation experiment aimed at evaluating the CLC automation system. Section 7.1 discusses the methodology used to conduct the user evaluation. Section 7.2 explores the process of user evaluation. Section 7.3 presents the data analysis following the user evaluation. Section 7.4 further explains and discusses the evaluation results, and Section 7.5 summarises the findings of the user evaluation study.

7.1 Methodology

The evaluation of computer-based systems is an important step when it comes to ascertaining the effectiveness of these systems (Weiss 1972). Evaluation is recommended as the primary technique for establishing the worth of an information system (Boloix and Robillard 1995). Kumar (Kumar 1990) classified evaluation into two categories, namely formative and summative evaluation. Formative evaluation produces information that is fed back during development. Summative evaluation is conducted after the development is completed. Summative evaluation results in beneficial outcomes such as ensured compliance with user objectives, and improvements in the effectiveness and productivity of the design (Green and Keim 1983).

As suggested by Hamilton and Chervany (Hamilton and Chervany 1981), one primary purpose of the computer-based information systems is to enhance an organisation's ability to accomplish its objectives. The CLC automation system aims to support users' decision-making by offering them useful information about the allegedly infringing content. As such, the summative evaluation method was adopted to evaluate the effectiveness of users employing the system to accomplish their mission in relation to infringement decisions. The CLC Model and the automation system are supposed to be deployed by online service providers such as link providers and online anti-piracy service providers in the notice-and-take-down context; as such, besides the normal users, users with basic knowledge of copyright infringement and Web technologies are also participants in the evaluation process. The CLC automation system will be supplied to a certain number of users as a supportive tool for their decision-making process. The quantitative method was utilised in this study. Data generated by users was analysed and compared with experts' data to indicate the effectiveness of the system.

During the evaluation process, users were divided into different groups depending on whether they had been trained and given the necessary introduction to copyright and whether they were offered the automation system for support. These users were presented with the same webpages used in the expert validation experiment, and experts' ratings of these webpages were treated as a standard. An electronic document questionnaire was administered to users, thus giving them adequate time to consider their answers. For each webpage they viewed and examined, they gave a rating on the likelihood of infringement on the webpage with or without the system support. The rating scores were analysed using statistical methods to gauge the system's impact on users and their decisions.

7.2 Evaluation Process

7.2.1 Selection of users

The system is designed to be used by online service providers in order to support their decision-making process; an example of such a user group would be Google employees working in the area of assessing allegedly infringing work and dealing with take-down requests (Google 2013). These people are not legal experts and do not need to have a professional legal background. However, they may have the copyright and Web knowledge necessary to understand the infringement activity in the context of the Web. As such, the participants used to evaluate the system should have a similar level of background knowledge to those people mentioned above. Students completing a PhD in Web Science were selected as participants in the study. All of them were researchers working in the interdisciplinary area of examining the Web and understanding its impact on contemporary society. Some of them also had Web design and development experience.

To evaluate the system's impact on users and their decisions regarding infringement ratings, users were divided into four groups in order to compare the rating results among them. One group of users were given training on the basic principles of copyright and introduced to the issues of online copyright infringement, following which they viewed and examined webpages. One group of users were offered the automation system as a support tool when they viewed and examined webpages. One group of users were given both training and system support, while the last group of users were just normal users who were not provided with training or system support. Table 22 shows the different groups of users.

Table 22. Categorisation of users

	Number of users	Number of webpages viewed per user	Training	System Support
Group 1	6	9~10	No	No
Group 2	6	9~10	Yes	No
Group 3	6	9~10	No	Yes
Group 4	6	9~10	Yes	Yes

Each user was allocated 9 to 10 webpages, while there were 29 webpages in total. Thus, in each group, at least 3 users were needed to view and examine all the webpages. In this study, 6 users in each group were selected, while the total number of respondents in the evaluation process was 24.

7.2.2 Conducting the evaluation

Because the evaluation result was based on the comparison between experts' ratings and users' ratings of infringement on webpages, the webpages given to users were the same as those given to the experts. These webpages covered all five types of URL listed in Section 3.1.3 and all five scenarios listed in Section 4.2. Participants were contacted by email and provided with a detailed explanation regarding the purpose and process of the study; they were also sent an overview of the questionnaire, and instructions on how to complete said questionnaire. For each webpage a user viewed and examined, only one question was asked in the questionnaire; this question was related to the user's rating of the likelihood of infringement on the webpage.

As stated earlier, in order to investigate the system's effect on the different groups of users' ratings, two groups of users were trained. This training was delivered via a seminar which lasted approximately forty minutes. The seminar included a presentation which introduced the principles of copyright, copyright concerns in the context of the Web, notice and take-down procedures, and a number of debatable legal cases relating to linking issues. Because the participants were PhD students, the purpose of the training was to make their roles closer to those of the real users of the system. Two groups of users who would use the system as a support tool were given a short introduction to the system face to face. This introduction described how to use the system and explained the meaning of system's outputs.

A questionnaire and related documents including allegedly infringing webpages were sent to all the participant users. For users in Group 2, training was provided before they started the evaluation. For users in Group 3, a short introduction to the system was delivered. Moreover, users in Group 4 received both training and a system introduction. The analysis of the quantitative data resulting from the questionnaire will be presented in the next section.

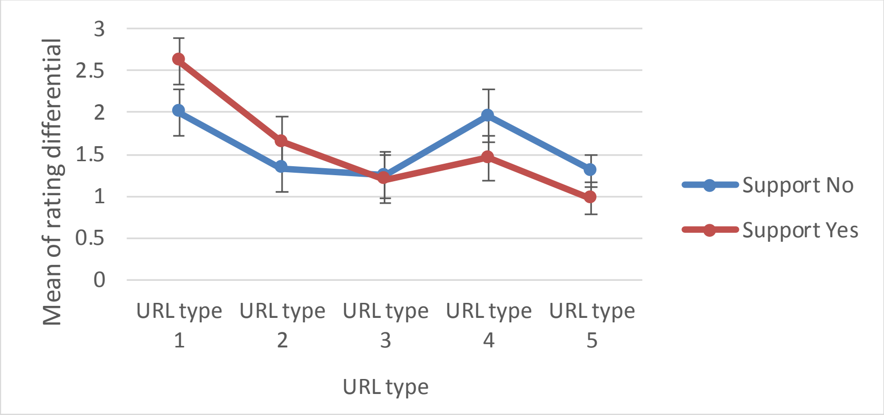
7.3 Analysis of Evaluation Result

As discussed in previous sections, four groups of users participated in the study, and the webpages presented to them belonged to five different categories which were configured in the preliminary study. The target is to explore the effect of training and system support on users' rating discrepancy in comparison to the ratings of experts. In addition, it is suspected that the difference between users' ratings and those of the experts would also depend on the types of webpages they viewed and examined. Here the Type 2 webpages can be taken as an example. On these webpages, only context information regarding the allegedly infringing content was found, e.g. title and performer; moreover, there was no interface at all to enable users to get access to the content. In this situation, users' rating differential from experts without training or system support might be quite close to those with training or system support. As such, the degree to which users' ratings differ from those of the experts might depend on training, system support and URL type.

In order to investigate whether training and system support help to bring users' ratings closer to those of the experts for specific type of URL, a three-way analysis of variance (ANOVA) was carried out.

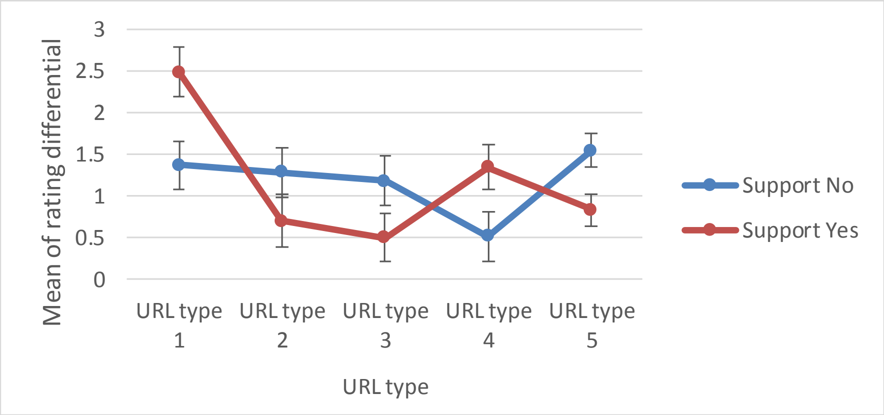
7.3.1 Users' rating differential from experts

A three-way ANOVA test determines whether there is a three-way interaction between three variables, in this case Training, Support and URL_type. Table 23 shows, for each type of URL, the mean of rating differential between users and experts in the different situations of training and support. The results are displayed in Figure 24 and Figure 25. Judging by the two line graphs, it is assumed that there is a three-way interaction because the lines do not appear to be parallel. This assumption can be confirmed by the three-way ANOVA test, which is shown in Table 24.



Error bars are ±1 standard error

Figure 24. Profile graph of simple two-way Support*URL_type interaction effect on rating differential without Training



Error bars are ±1 standard error

Figure 25. Profile graph of simple two-way Support*URL_type interaction test on rating differential with Training

Looking at the results in Table 24, it is confirmed that there is a statistically significant three-way interaction between training, system support and URL_type (Training * Support * URL_type), F(4, 271) =2.62, p=0.035. This means that, for different types of URL, training and system support have a significant interaction effect on users' rating differential from experts.

Table 23. Mean of rating differential on URL_type*Training*Support

URL_type	Training	Support	Mean	Std. Error
URL type 1	No	No	2.00	.29
		Yes	2.62	.29
	Yes	No	1.37	.29
		Yes	2.48	.30
URL type 2	No	No	1.33	.29
		Yes	1.65	.30
	Yes	No	1.28	.30
		Yes	.70	.31
URL type 3	No	No	1.25	.29
		Yes	1.20	.29
	Yes	No	1.18	.30
		Yes	.50	.29
URL type 4	No	No	1.96	.31
		Yes	1.46	.26
	Yes	No	.52	.30
		Yes	1.34	.26
URL type 5	No	No	1.30	.19
		Yes	.97	.19
	Yes	No	1.55	.19
		Yes	.83	.19

Table 24. Three-way ANOVA test of between-subjects effects for Training*Support*URL_type

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Training	10.33	1	10.33	10.60	.001
Support	.00	1	.00	.00	.996
URL_type	36.47	4	9.12	9.36	.000
Training * Support	.005	1	.005	.005	.945
Training * URL_type	6.59	4	1.65	1.69	.153
Support * URL_type	17.26	4	4.32	4.43	.002
Training * Support * URL_type	10.23	4	2.56	2.62	.035
Error	264.02	271	.97		
Total	857.64	291			

In light of the detailed effects resulting from interaction between the three factors of Training, Support and URL_type, three separate analyses of simple two-way interaction were carried out to find out the 1) effect of Training*Support for each type of URL; 2) effect of Training*URL_type at each level of support; 3) effect of Support*URL_type at each level of training.

7.3.2 Interaction effects of Training and Support for different types of URL

In order to explore for which type of URL training and system support have significant interaction effects on users' rating differential from experts, a simple two-way interaction analysis was carried out. The following syntax was used to test two-way interaction effects.

*/Test=Training*Support VS 264.02 DF(271)*

Here the number 264.02 is the sum of squares obtained from Table 24. Moreover, the number of degrees of freedom for the error term is also declared, which is 271. Table 25 shows the test result from the simple two-way interaction analysis. It indicates that there is a statistically significant simple two-way interaction between training and support for URL type 4 (p=0.021). The interaction effects between training and support for other types of URL are not significant.

Table 25. Test of simple two-way interactions Training*Support at each level of URL_type

URL_type	Source	Sum of Squares	df	Mean Square	F	Sig.
URL type 1	Training*Support	.73	1	.73	.75	.388
	Error	264.02	271	.97		
URL type 2	Training*Support	2.19	1	2.19	2.25	.135
	Error	264.02	271	.97		
URL type 3	Training*Support	1.17	1	1.17	1.20	.274
	Error	264.02	271	.97		
URL type 4	Training*Support	5.28	1	5.28	5.42	.021
	Error	264.02	271	.97		
URL type 5	Training*Support	.96	1	.96	.99	.321
	Error	264.02	271	.97		

For URL type 4, a simple simple main effects test was conducted to gauge the effect of training at each level of support and the effect of support at each level of training. For other URL types, a simple main effects analysis was carried out.

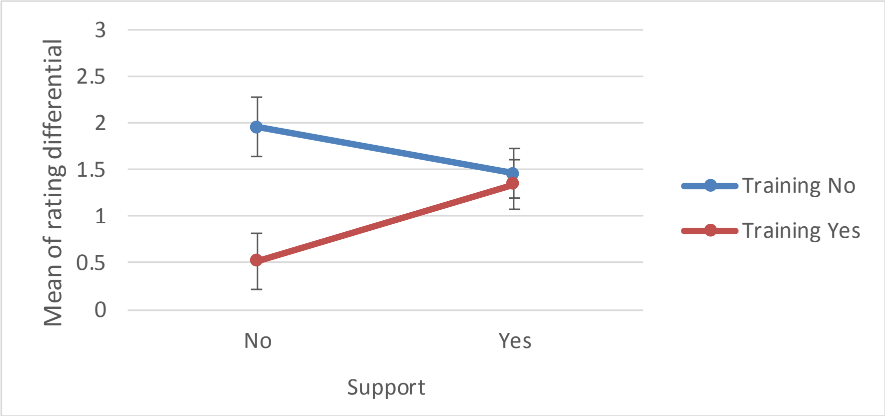
7.3.2.1 Simple simple main effects analysis for URL type 4

a) Simple simple main effect of Training at each level of Support

Table 26 shows that, for URL type 4, the simple simple main effect of training for users' rating differential is significant when the users are not given system support. Without system support and without training, users' rating differential from experts is 1.96±0.31 (Table 23); and with training, their rating differential is 0.52±0.30 (Table 23), the difference of 1.44 is statistically significant (95% confidence interval, p<0.05). Figure 26 indicates that for URL type 4, when users are not given system support, users with training have significantly lower rating differential from experts comparing to users without training. In other words, training results in significantly lower rating differential from experts when there is no system support.

Table 26. Test of simple simple main effect of Training at each level of Support

URL_type	Support	Source	Sum of Squares	df	Mean Square	F	Sig.
URL type 4	No	Training	10.89	1	10.89	11.18	.001
			264.02	271	.974		
	Yes	Training	.091	1	.091	.094	.760
			264.02	271	.974		



Error bars are ±1 standard error

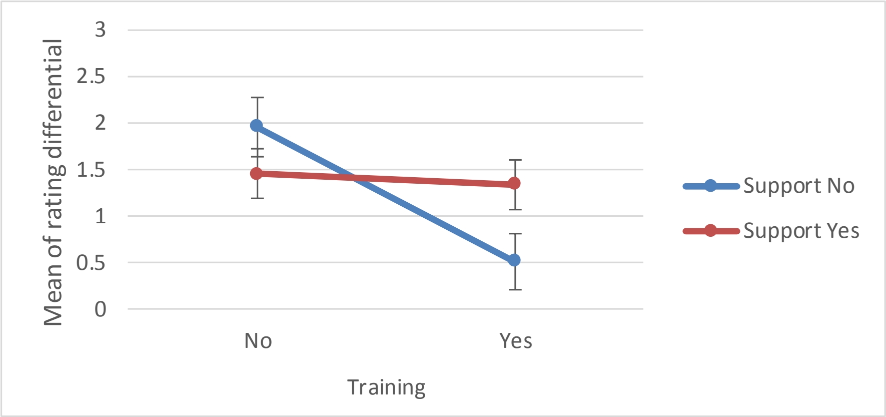
Figure 26. Profile graph of simple simple main effect of Training on rating differential at each level of Support for URL type 4

b) Simple simple main effect of Support at each level of Training

Table 27 shows that, for URL type 4, the simple simple main effect of support for users' rating differential is significant when the users are given training. With training and without support, users' rating differential from experts is 0.52±0.30 (Table 23), and with support, their rating differential is 1.34±0.26 (Table 23); the difference of 0.82 is statistically significant (95% confidence interval, $p<0.05$). Figure 26 indicates that, for URL type 4, when users are given training, users without system support have a significantly lower rating differential from experts compared to users with support. In other words, system support results in a significantly higher rating differential from experts when users are trained.

Table 27. Test of simple simple main effect of Support at each level of Training

URL_type	Training	Source	Sum of Squares	df	Mean Square	F	Sig.
URL type 4	No	Support	1.48	1	1.48	1.51	.220
		Error	264.02	271	.97		
	Yes	Support	4.19	1	4.19	4.30	.039
		Error	264.02	271	.97		



Error bars are ±1 standard error

Figure 27. Profile graph of simple simple main effect of Support on rating differential at each level of Training for URL type 4

7.3.2.2 Simple main effects analysis for URL type 1

For URL type 1, training and support have no interaction effect on rating differential. Tests of simple main effect are carried out to explore respectively the effect of training and the effect of support on users' rating differential from experts.

a) Simple main effect of Training

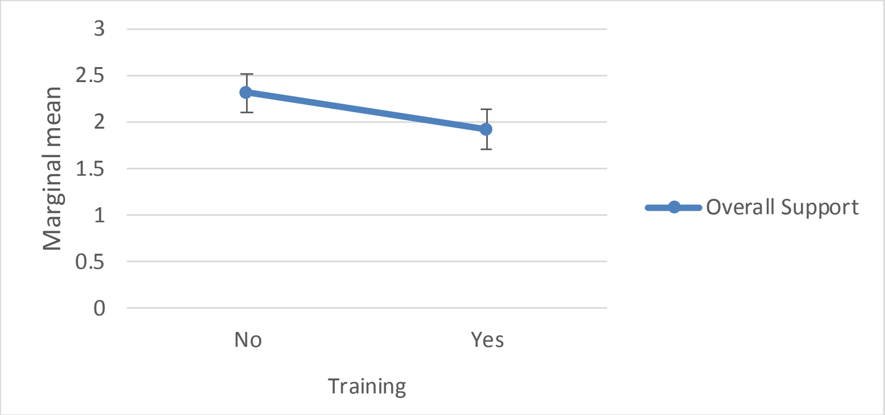
Table 28 and Figure 28 show, regardless of support, the mean of users' rating differential from experts in the situation of without and with training. Without training, the mean is 2.31±0.20, and with training, the mean is 1.92±0.21; the difference is not significant, as shown in Table 29 (p=0.196). In conclusion, for URL type 1, regardless of support, training has no significant effect on users' rating differential from experts.

Table 28. Mean of rating differential for Training on URL type 1

Training	Mean	Std. Error
No	2.31	.20
Yes	1.92	.21

Table 29. Test of simple main effect of Training on URL type 1

	Sum of Squares	df	Mean Square	F	Sig.
Training	1.73	1	1.73	1.73	.196
Error	43.10	43	1.00		



Error bars are ±1 standard error

Figure 28. Profile graph of simple main effect of Training overall Support for URL type 1

b) Simple main effect of Support

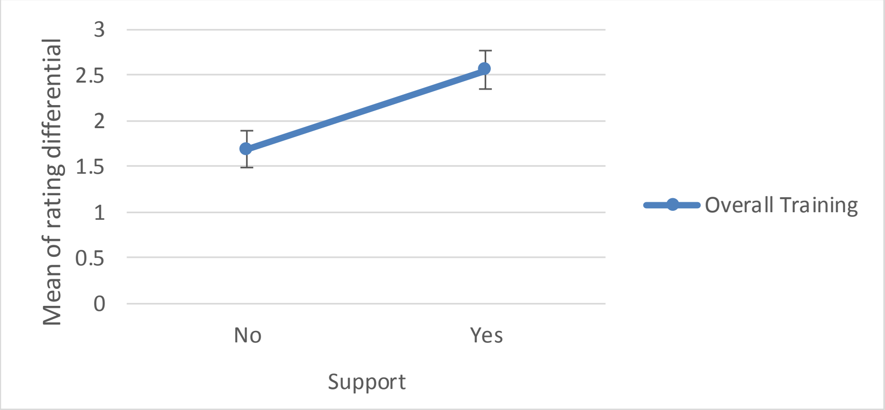
Table 30 and Figure 29 show, regardless of training, the mean of users' rating differential from experts in the situation of without and with system support. Without support, the mean is 1.68±0.20, and with training, the mean is 2.55±0.21; the difference is significant, as shown in Table 31 (p=0.005). In conclusion, for URL type 1, regardless of training, support results in a significantly higher rating differential.

Table 30. Mean of rating differential for Support on URL type 1

Support	Mean	Std. Error
No	1.68	.20
Yes	2.55	.21

Table 31. Test of simple main effect of Support on URL type 1

	Sum of Squares	df	Mean Square	F	Sig.
Support	8.80	1	8.80	8.78	.005
Error	43.10	43	1.00		



Error bars are ±1 standard error

Figure 29. Profile graph of simple main effect of Support overall Training for URL type 1

7.3.2.3 Simple main effects analysis for URL type 2

For URL type 2, training and support have no interaction effect on rating differential. Tests of simple main effect are carried out to explore respectively the effect of training and the effect of support on users' rating differential from experts.

a) Simple main effect of Training

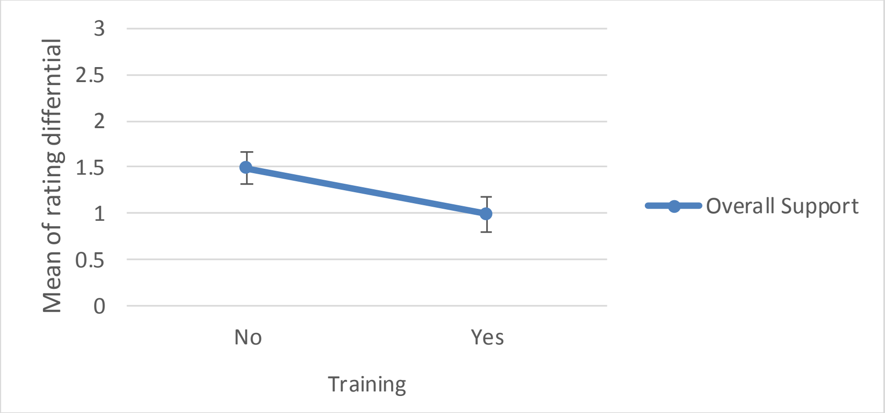
Table 32 and Figure 30 show, regardless of support, the mean of users' rating differential from experts in the situation of without and with training. Without training, the mean is 1.49±0.18, and with training, the mean is 0.99±0.19; the difference is not significant, as shown in Table 33 (p=0.065). In conclusion, for URL type 2, regardless of support, training has no significant effect on users' rating differential from experts.

Table 32. Mean of rating differential for Training on URL type 2

Training	Mean	Std. Error
No	1.49	.18
Yes	.99	.19

Table 33. Test of simple main effect of Training on URL type 2

	Sum of Squares	df	Mean Square	F	Sig.
Training	2.72	1	2.72	3.60	.065
Error	30.27	40	.76		



Error bars are ±1 standard error

Figure 30. Profile graph of simple main effect of Training overall Support for URL type 2

b) Simple main effect of Support

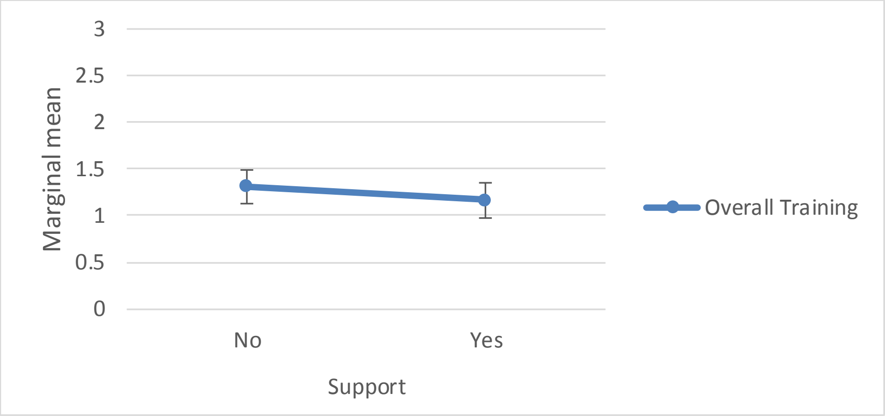
Table 34 and Figure 31 show, regardless of training, the mean of users' rating differential from experts in the situation of without and with system support. Without support, the mean is 1.31±0.18, and with training, the mean is 1.17±0.19; the difference is not significant, as shown in Table 35 (p=0.611). In conclusion, for URL type 2, regardless of training, support has no significant effect on users' rating from experts.

Table 34. Mean of rating differential for Support on URL type 2

Support	Mean	Std. Error
No	1.31	.18
Yes	1.17	.19

Table 35. Test of simple main effect of Support on URL type 2

	Sum of Squares	df	Mean Square	F	Sig.
Support	.20	1	.20	.26	.611
Error	30.27	40	.76		



Error bars are ±1 standard error

Figure 31. Profile graph of simple main effect of Support overall Training for URL type 2

7.3.2.4 Simple main effects analysis for URL type 3

For URL type 3, training and support have no interaction effect of on rating differential. Tests of simple main effect are carried out to explore respectively the effect of training and the effect of support on users' rating differential from experts.

a) Simple main effect of Training

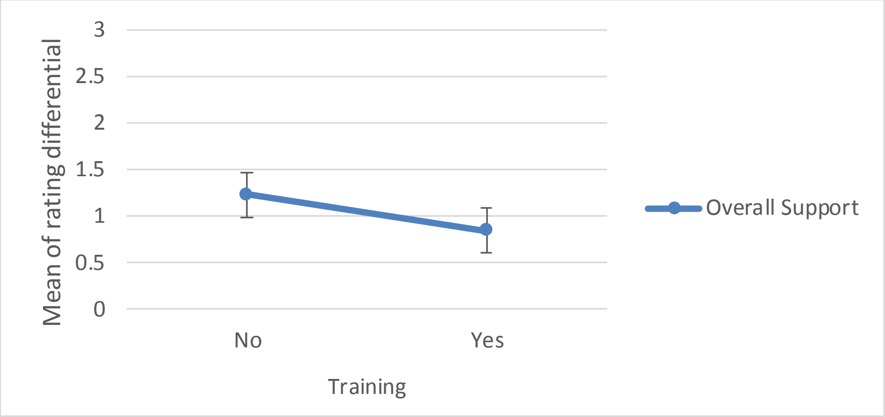
Table 36 and Figure 32 show, regardless of support, the mean of users' rating differential from experts in the situation of without and with training. Without training, the mean is 1.23±0.24, and with training, the mean is 0.84±0.24; the difference is not significant, as shown in Table 37 (p=0.265). In conclusion, for URL type 3, regardless of support, training has no significant effect on users' rating differential from experts.

Table 36. Mean of rating differential for Training on URL type 3

Training	Mean	Std. Error
No	1.23	.24
Yes	.84	.24

Table 37. Test of simple main effect of Training on URL type 3

	Sum of Squares	df	Mean Square	F	Sig.
Training	1.73	1	1.73	1.28	.265
Error	58.25	43	1.36		



Error bars are ±1 standard error

Figure 32. Profile graph of simple main effect of Training overall Support for URL type 3

b) Simple main effect of Support

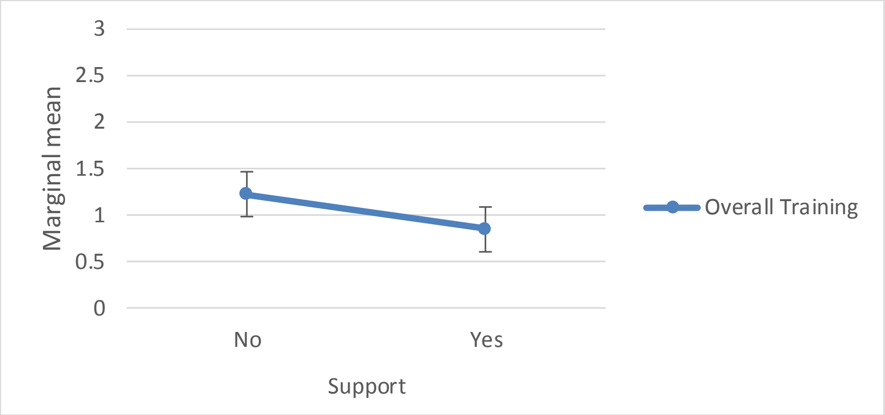
Table 38 and Figure 31 show, regardless of training, the mean of users' rating differential from experts in the situation of without and with system support. Without support, the mean is 1.22±0.24, and with training, the mean is 0.85±0.24; the difference is not significant, as shown in Table 39 (p=0.288). In conclusion, for URL type 3, regardless of training, support has no significant effect on users' rating differential from experts.

Table 38. Mean of rating differential for Support on URL type 3

Support	Mean	Std. Error
No	1.22	.24
Yes	.85	.24

Table 39. Test of simple main effect of Support on URL type 3

	Sum of Squares	df	Mean Square	F	Sig.
Support	1.57	1	1.57	1.16	.288
Error	58.25	43	1.36		



Error bars are ±1 standard error

Figure 33. Profile graph of simple main effect of Support overall Training for URL type 3

7.3.2.5 Simple main effects analysis for URL type 5

For URL type 5, training and support have no interaction effect on rating differential. Tests of simple main effect are carried out to explore respectively the effect of training and the effect of support on users' rating differential from experts.

a) Simple main effect of Training

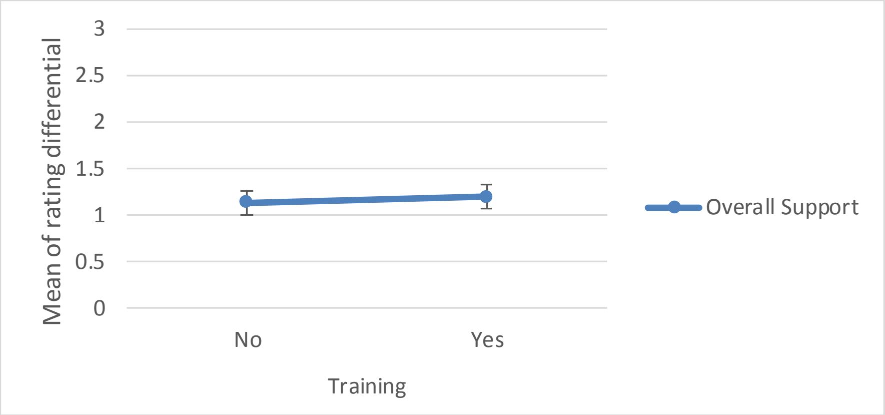
Table 40 and Figure 34 show, regardless of support, the mean of users' rating differential from experts in the situation of without and with training. Without training, the mean is 1.14±0.13, and with training, the mean is 1.19±0.13; the difference is not significant, as shown in Table 41Table 37 (p=0.761). In conclusion, for URL type 5, regardless of support, training has no significant effect on users' rating differential from experts.

Table 40. Mean of rating differential for Training on URL type 5

Training	Mean	Std. Error
No	1.14	.13
Yes	1.19	.13

Table 41. Test of simple main effect of Training on URL type 5

	Sum of Squares	df	Mean Square	F	Sig.
Training	.075	1	.075	.093	.761
Error	81.18	100	.812		



Error bars are ±1 standard error

Figure 34. Profile graph of simple main effect of Training overall Support for URL type 5

b) Simple main effect of Support

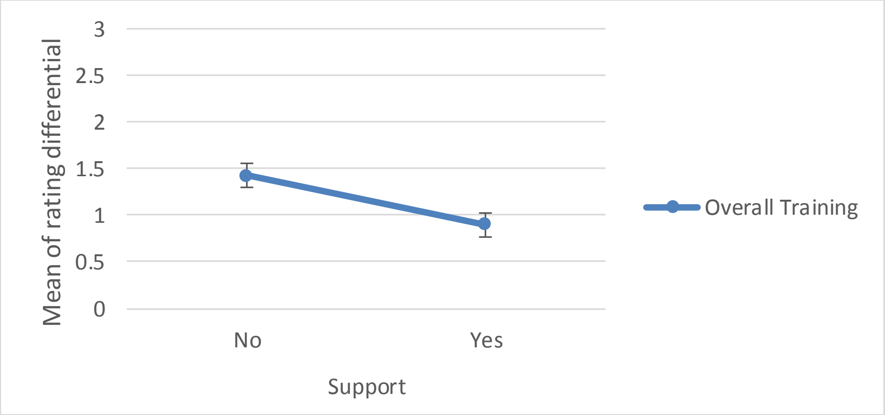
Table 42 and Figure 35 show, regardless of training, the mean of users' rating differential from experts in the situation of without and with system support. Without support, the mean is 1.42±0.13, and with training, the mean is 0.90±0.13; the difference is significant, as shown in Table 43 (p=0.004). In conclusion, for URL type 5, regardless of training, support results in a significantly lower rating differential.

Table 42. Mean of rating differential for Support on URL type 5

Support	Mean	Std. Error
No	1.42	.13
Yes	.90	.13

Table 43. Test of simple main effect of Support on URL type 5

	Sum of Squares	df	Mean Square	F	Sig.
Support	7.11	1	7.11	8.76	.004
Error	81.18	100	.81		



Error bars are ±1 standard error

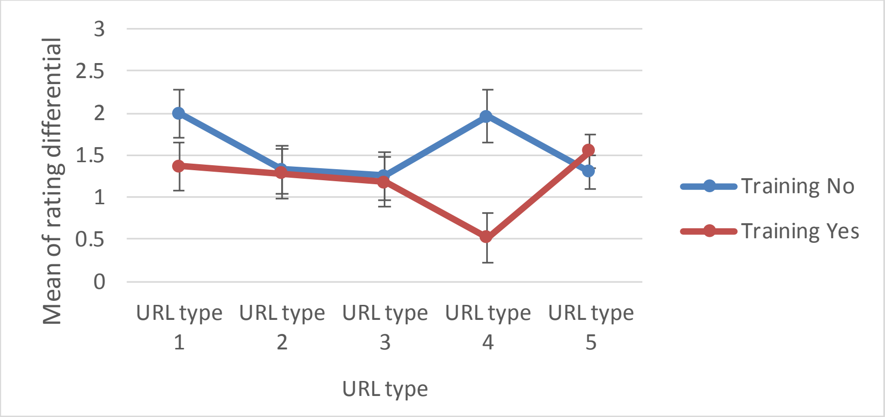
Figure 35. Profile graph of simple main effect of Support overall Training for URL type 5

7.3.3 Interaction effects of Training and URL type at each level of Support

In order to investigate whether training and URL type had interaction effects on rating differential when users were given system support and not given support, it was appropriate to conduct a test of simple two-way interaction Training*URL_type at each level of support. The results are displayed in Figure 36 and Figure 37.

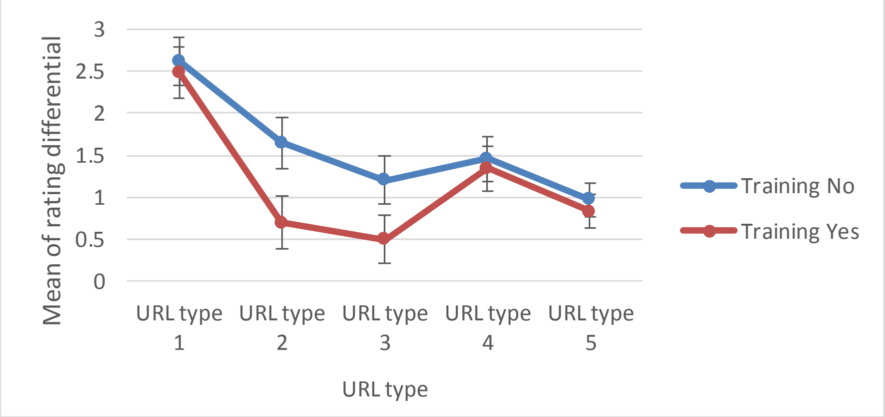
Table 44 shows the test result from the simple two-way interaction analysis. There is a statistically significant simple two-way interaction between training and URL_type in the situation of no support (p=0.017). In other words, for the users without system support, training has a significant effect on their rating differential for specific types of URL. In order to establish the type of URL and whether the ratings are closer to those of the experts, a simple simple main effect test was carried out to find out the detailed effect of training on different types of URL and the effect of URL type at each level of training.

When there is support, the interaction effect between training and URL_type is not significant. A test of simple main effect of training and a test of simple main effect of URL_type were carried out.



Error bars are ±1 standard error

Figure 36. Profile graph of simple two-way Training*URL_type interaction effect on rating differential without Support



Error bars are ±1 standard error

Figure 37. Profile graph of simple two-way Training*URL_type interaction effect on rating differential with Support

Table 44. Test of simple two-way interactions Training*URL_type at each level of Support

Support	Source	Sum of Squares	df	Mean Square	F	Sig.
No	Training*URL_type	11.94	4	2.98	3.06	.017
	Error	264.02	271	.97		
Yes	Training*URL_type	3.82	4	.96	.98	.418
	Error	264.02	271	.97		

7.3.3.1 Simple simple main effects analysis in the situation of no support

a) Simple simple main effect of Training at each level of URL_type

Table 45. Test of simple simple main effect of Training at each level of URL_type

Support	URL_type	Source	Sum of Squares	df	Mean Square	F	Sig.
No	URL type 1	Training	2.41	1	2.41	2.47	.117
		Error	264.02	271	.97		
	URL type 2	Training	.015	1	.015	.016	.901
		Error	264.02	271	.974		
	URL type 3	Training	.027	1	.027	.027	.869
		Error	264.02	271	.974		
	URL type 4	Training	10.90	1	10.90	11.18	.001
		Error	264.02	271	.974		
	URL type 5	Training	.788	1	.788	.809	.369
		Error	264.02	271	.974		

Table 46. Mean of rating differential for Training without Support

Support	URL_type	Training	Mean	Std. Error
No	URL type 1	No	2.00	.31
		Yes	1.37	.31
	URL type 2	No	1.33	.27
		Yes	1.28	.28
	URL type 3	No	1.25	.40
		Yes	1.18	.42
	URL type 4	No	1.96	.29
		Yes	.52	.27
	URL type 5	No	1.30	.19
		Yes	1.55	.19

Table 45 shows that for URL type 4, the simple simple main effect of training for users' rating differential is significant when the users are not given system support (p=0.001). Without system support and without training, users' rating differential from experts is 1.96±0.29 (Table 46), and with training, the rating differential is 0.52±0.27; the difference of 1.44 is statistically significant. In other words, for URL type 4, training results in a significantly lower rating differential from experts when there is no system support. When there is no support, training has no significant effect on users' rating differential for other types of URL. This result is the same as the previous analysis result discussed in Section 7.3.2.1.

b) Simple simple main effect of URL_type at each level of Training

Table 47 indicates that when there is no system support, the simple simple main effect of URL_type on rating differential is not significant for those users who are given training. Moreover, the effect is not significant for the users who are not given training either. Table 48 further explains this situation through a pairwise comparison of users' rating differential across different types of URL. Figure 38 visually indicates the simple simple main effect of URL_type on rating differential at each level of training. The graph shows that, for the users who are not given training, their rating differentials for all the five types of URL are not significantly different. For those users who are given training, their rating differentials for all five types of URL are not significant either. Without training, the five points which represent users' rating differential on the five types of URL are quite close, and with training, the five points are still quite close. While the point representing URL type 4 is a little far from the other four points, it is not significantly different.

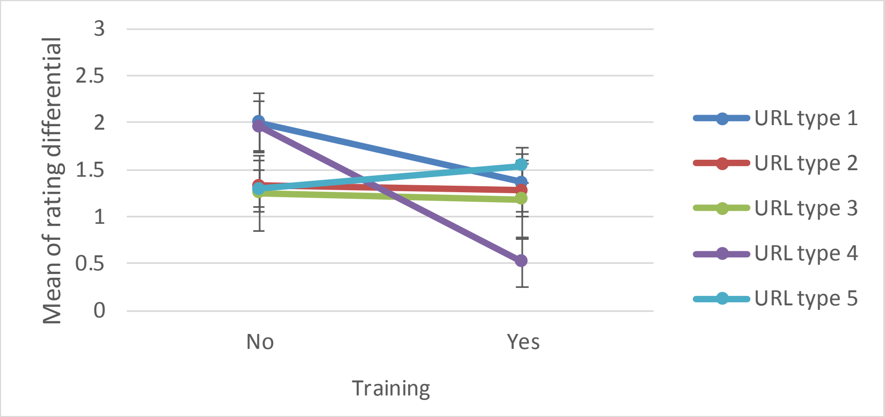
Table 47. Test of simple simple main effect of URL_type at each level of Training

Support	Training	Source	Sum of Squares	df	Mean Square	F	Sig.
No	No	URL_type	7.24	4	1.81	1.86	.118
		Error	264.02	271	.97		
	Yes	URL_type	8.39	4	2.10	2.15	.075
		Error	264.02	271	.97		

Table 48. Pairwise comparison among URL types without support

							95% Confidence Interval for Difference	
Support	Training	(I) URL_type	(J) URL_type	Mean Difference (I-J)	Std. Error	Sig	Lower Bound	Upper Bound
No	No	URL type 1	URL type 2	.667	.403	.992	-.474	1.807
			URL type 3	.750	.403	.638	-.390	1.890
			URL type 4	.040	.423	1.000	-1.156	1.236
			URL type 5	.700	.344	.431	-.275	1.675
		URL type 2	URL type 1	-.667	.403	.992	-1.807	.474
			URL type 3	.083	.403	1.000	-1.057	1.224
			URL type 4	-.627	.423	1.000	-1.823	.569
			URL type 5	.033	.344	1.000	-.942	1.008
		URL type 3	URL type 1	-.750	.403	.638	-1.890	.390

		URL type 2	-.083	.403	1.000	-1.224	1.057
		URL type 4	-.710	.423	.941	-1.906	.486
		URL type 5	-.050	.344	1.000	-1.025	.925
	URL type 4	URL type 1	-.040	.423	1.000	-1.236	1.156
		URL type 2	.627	.423	1.000	-.569	1.823
		URL type 3	.710	.423	.941	-.486	1.906
		URL type 5	.660	.367	.735	-.379	1.699
	URL type 5	URL type 1	-.700	.344	.431	-1.675	.275
		URL type 2	-.033	.344	1.000	-1.008	.942
		URL type 3	.050	.344	1.000	-.925	1.025
		URL type 4	-.660	.367	.735	-1.699	.379
Yes	URL type 1	URL type 2	.085	.412	1.000	-1.081	1.251
		URL type 3	.185	.412	1.000	-.981	1.351
		URL type 4	.848	.412	.404	-.318	2.015
		URL type 5	-.179	.344	1.000	-1.154	.795
	URL type 2	URL type 1	-.085	.412	1.000	-1.251	1.081
		URL type 3	.100	.421	1.000	-1.091	1.291
		URL type 4	.764	.421	.707	-.428	1.955
		URL type 5	-.264	.355	1.000	-1.269	.740
	URL type 3	URL type 1	-.185	.412	1.000	-1.351	.981
		URL type 2	-.100	.421	1.000	-1.291	1.091
		URL type 4	.664	.421	1.000	-.528	1.855
		URL type 5	-.364	.355	1.000	-1.369	.640
	URL type 4	URL type 1	-.848	.412	.404	-2.015	.318
		URL type 2	-.764	.421	.707	-1.955	.428
		URL type 3	-.664	.421	1.000	-1.855	.528
		URL type 5	-1.028	.355	.041	-2.033	-.023
	URL type 5	URL type 1	.179	.344	1.000	-.795	1.154
		URL type 2	.264	.355	1.000	-.740	1.269
		URL type 3	.364	.355	1.000	-.640	1.369
		URL type 4	1.028	.355	.041	.023	2.033



Error bars are ±1 standard error

Figure 38. Profile graph of simple simple main effect of URL_type on rating differential at each level of Training

7.3.3.2 Simple main effects analysis in the situation of support

When there is support, Training*URL_type has no interaction effect on rating differential. Tests of simple main effect are carried out respectively to explore the effect of training on users' rating differential and the difference among the five types of URL.

a) Simple main effect of Training

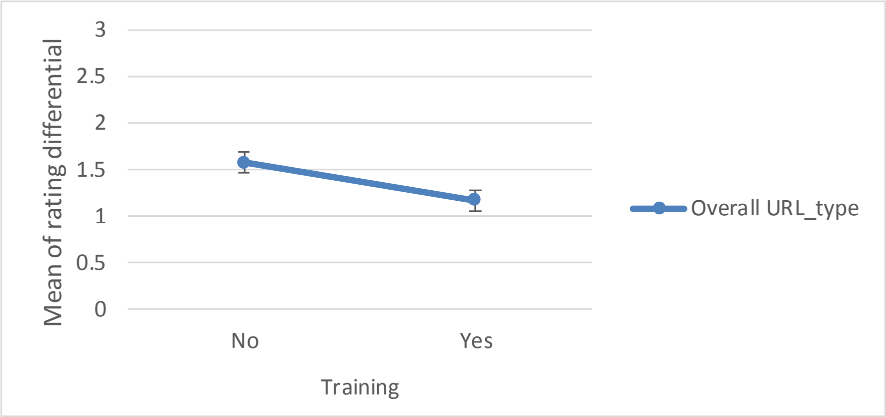
Table 49 and Figure 39 show, regardless of types of URL, the mean of users' rating differential from experts in the situation of without and with training. Without training, the mean is 1.58±0.11, and with training, the mean is 1.17±0.11; the difference is significant, as shown in Table 50 (p=0.012). In conclusion, when users are given system support, regardless of the type of URL, training results in a significantly lower rating differential from experts.

Table 49. Mean of rating differential for Training overall URL_type

Training	Mean	Std. Error
No	1.58	.11
Yes	1.17	.11

Table 50. Test of simple main effect of Training

	Sum of Squares	df	Mean Square	F	Sig.
Training	5.50	1	5.50	6.46	.012
Error	117.50	138	.85		



Error bars are ±1 standard error

Figure 39. Profile graph of simple main effect of Training overall URL_type

b) Simple main effect of URL_type

Table 51 shows, regardless of training, the mean of users' rating differential for every type of URL when those users are given system support. The means of rating differential among these five types of URL are significantly different, as shown in Table 52 (p=0.000). Table 53 further explains this situation through a pairwise comparison of users' rating differential across the five types of URL. Figure 40 visually indicates the simple main effect of URL_type overall training. The results shows that the users' rating differential for URL type 1 is significantly higher than for the other four types of URL.

Table 51. Mean of rating differential for URL_type overall Training

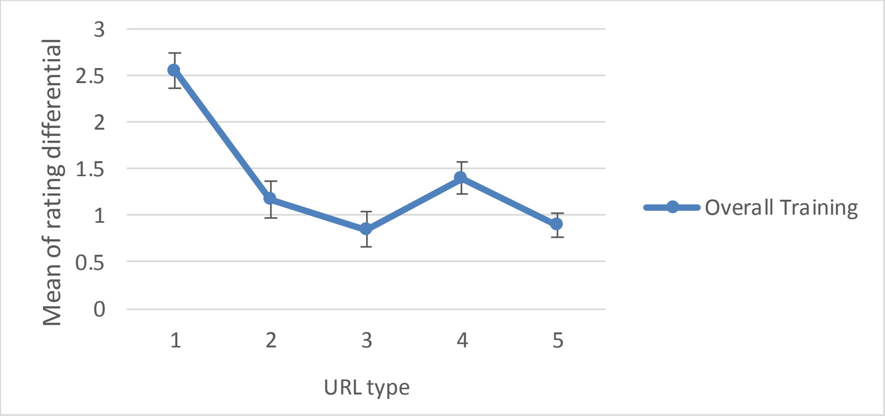
URL_type	Mean	Std. Error
URL type 1	2.55	.19
URL type 2	1.17	.20
URL type 3	.85	.19
URL type 4	1.40	.17
URL type 5	.90	.13

Table 52. Test of simple main effect of URL_type overall Training

	Sum of Squares	df	Mean Square	F	Sig.
URL_type	49.57	4	12.39	14.56	.000
Error	117.50	138	.85		

Table 53. Pairwise comparison among URL types with support

(I) URL_type	(J) URL_type	Mean Difference (I-J)	Std. Error	Sig	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
URL type 1	URL type 2	1.377	.279	.000	.581	2.172
	URL type 3	1.699	.269	.000	.931	2.468
	URL type 4	1.149	.260	.000	.408	1.890
	URL type 5	1.649	.231	.000	.990	2.309
URL type 2	URL type 1	-1.377	.279	.000	-2.172	-.581
	URL type 3	.323	.276	1.000	-.464	1.110
	URL type 4	-.227	.267	1.000	-.988	.533
	URL type 5	.273	.239	1.000	-.408	.954
URL type 3	URL type 1	-1.699	.269	.000	-2.468	-.931
	URL type 2	-.323	.276	1.000	-1.110	.464
	URL type 4	-.550	.257	.339	-1.282	.182
	URL type 5	-.050	.228	1.000	-.700	.600
URL type 4	URL type 1	-1.149	.260	.000	-1.890	-.408
	URL type 2	.227	.267	1.000	-.533	.988
	URL type 3	.550	.257	.339	-.182	1.282
	URL type 5	.500	.216	.223	-.117	1.117
URL type 5	URL type 1	-1.649	.231	.000	-2.309	-.990
	URL type 2	-.273	.239	1.000	-.954	.408
	URL type 3	.050	.228	1.000	-.600	.700
	URL type 4	-.500	.216	.223	-1.117	.117



Error bars are ±1 standard error

Figure 40. Profile graph of simple main effect of URL_type overall Training

7.3.4 Interaction effects of Support and URL type at each level of Training

In order to establish whether support and URL_type had interaction effects on rating differential when users were given training and not given training; in order to achieve this, a test of simple two-way interaction Support*URL_type at each level of training was carried out. The result graphs are not presented here as they are exactly the same as the graphs displayed in Figure 24 and Figure 25.

Table 54 shows the test result from the simple two-way interaction analysis. There is a statistically significant simple two-way interaction between support and URL_type in the situation of no training (p=0.000). In other words, for the users with training, support has a significant effect on their rating differential for specific types of URL. In order to establish the type of URL and whether the ratings are closer to those of the experts, a simple simple main effect test was carried out to find out the detailed effect of support at different types of URL and the effect of URL type at each level of support.

When there is training, the interaction effect between support and URL_type is not significant. A test of simple main effect of support and a test of simple main effect of URL_type was carried out.

Table 54. Test of simple two-way interactions Support*URL_type at each level of Training

Training	Source	Sum of Squares	df	Mean Square	F	Sig.
No	Support*URL_type	5.64	4	1.41	1.45	.219
	Error	264.02	271	.97		
Yes	Support*URL_type	21.78	4	5.45	5.59	.000
	Error	264.02	271	.97		

7.3.4.1 Simple simple main effects analysis in the situation of training

a) Simple simple main effect of Support at each level of URL_type

Table 55 shows that, for URL type 1 (p=0.007), URL type 4 (p=0.039) and URL type 5 (p=0.009), the simple simple main effect of support for users' rating differential is significant when the users are given training. With training and without system support, users' rating differential from experts for URL type 1 is 1.37±0.29 (Table 56), and with system support, the rating differential is 2.48±0.30; the difference of 1.11 is statistically significant. In other words, for URL type 1, support results in a significantly higher rating differential from experts when they are trained.

With training and without system support, users' rating differential from experts for URL type 4 is 0.52±0.30 (Table 56), and with system support, the rating differential is 1.34±0.26; the difference of 0.82 is statistically significant. In other words, for URL type 4, support results in a significantly higher rating differential from experts when they are trained.

With training and without system support, users' rating differential from experts for URL type 5 is 1.55±0.19 (Table 56), and with system support, the rating differential is 0.83±0.19; the difference of 0.72 is statistically significant. In other words, for URL type 5, support results in a significantly lower rating differential from experts when they are trained.

Table 55. Test of simple simple main effect of Support at each level of URL_type

Training	URL_type	Source	Sum of Squares	df	Mean Square	F	Sig.
Yes	URL type 1	Support	7.14	1	7.14	7.33	.007
		Error	264.02	271	.97		
	URL type 2	Support	1.77	1	1.77	1.82	.178
		Error	264.02	271	.97		
	URL type 3	Support	2.67	1	2.67	2.74	.099
		Error	264.02	271	.97		
	URL type 4	Support	4.19	1	4.19	4.30	.039
		Error	264.02	271	.97		
	URL type 5	Support	6.65	1	6.65	6.83	.009
		Error	264.02	271	.97		

Table 56. Mean of rating differential for Support with Training

Training	URL_type	Support	Mean	Std. Error
Yes	URL type 1	No	1.37	.29
		Yes	2.48	.30
	URL type 2	No	1.28	.30
		Yes	.70	.31
	URL type 3	No	1.18	.30
		Yes	.50	.29
	URL type 4	No	.52	.30
		Yes	1.34	.26
	URL type 5	No	1.55	.19
		Yes	.83	.19

b) Simple simple main effect of URL_type at each level of Support

Table 57 indicates that when users are trained, the simple simple main effect of URL_type on rating differential is not significant for those users who are not given system support. Moreover, the effect is significant for those users who are given system support.

Table 58 further explains this situation through a pairwise comparison of users' rating differential across different types of URL. Figure 41 visually indicates the simple simple main effect of URL_type on rating differential at each level of support. The graph shows that, for the users who are not given system support, their rating differentials for all five types of URL are not significantly different. For those users who are given system support, their rating differential on URL type 1 is significantly higher than for other types of URL.

Table 57. Test of simple simple main effect of URL_type at each level of Support

Training	Support	Source	Sum of Squares	df	Mean Square	F	Sig.
Yes	No	URL_type	8.39	4	2.10	2.15	.075
		Error	264.02	271	.97		
	Yes	URL_type	29.63	4	7.41	7.60	.000
		Error	264.02	271	.97		

Table 58. Pairwise comparison among URL types with training

				Mean	Std.		95% Confidence Interval for Difference	
Training	Support	(I) URL_type	(J) URL_type	Difference (I-J)	Error	Sig.	Lower Bound	Upper Bound
Yes	No	URL type 1	URL type 2	.085	.412	1.000	-1.081	1.251
			URL type 3	.185	.412	1.000	-.981	1.351
			URL type 4	.848	.412	.404	-.318	2.015
			URL type 5	-.179	.344	1.000	-1.154	.795
		URL type 2	URL type 1	-.085	.412	1.000	-1.251	1.081
			URL type 3	.100	.421	1.000	-1.091	1.291
			URL type 4	.764	.421	.707	-.428	1.955
			URL type 5	-.264	.355	1.000	-1.269	.740
		URL type 3	URL type 1	-.185	.412	1.000	-1.351	.981
			URL type 2	-.100	.421	1.000	-1.291	1.091
			URL type 4	.664	.421	1.000	-.528	1.855
			URL type 5	-.364	.355	1.000	-1.369	.640
		URL type 4	URL type 1	-.848	.412	.404	-2.015	.318
			URL type 2	-.764	.421	.707	-1.955	.428
			URL type 3	-.664	.421	1.000	-1.855	.528
			URL type 5	-1.028	.355	.041	-2.033	-.023
		URL type 5	URL type 1	.179	.344	1.000	-.795	1.154
			URL type 2	.264	.355	1.000	-.740	1.269
			URL type 3	.364	.355	1.000	-.640	1.369
			URL type 4	1.028	.355	.041	.023	2.033
	Yes	URL type 1	URL type 2	1.782	.431	.000	.561	3.002
			URL type 3	1.982	.412	.000	.816	3.148
			URL type 4	1.139	.398	.045	.013	2.265
			URL type 5	1.651	.355	.000	.646	2.656
		URL type 2	URL type 1	-1.782	.431	.000	-3.002	-.561
			URL type 3	.200	.423	1.000	-.996	1.396

	URL type 4	-.643	.409	1.000	-1.799	.514
	URL type 5	-.131	.367	1.000	-1.170	.909
URL type 3	URL type 1	-1.982	.412	.000	-3.148	-.816
	URL type 2	-.200	.423	1.000	-1.396	.996
	URL type 4	-.843	.388	.308	-1.942	.256
	URL type 5	-.331	.344	1.000	-1.306	.644
URL type 4	URL type 1	-1.139	.398	.045	-2.265	-.013
	URL type 2	.643	.409	1.000	-.514	1.799
	URL type 3	.843	.388	.308	-.256	1.942
	URL type 5	.512	.327	1.000	-.414	1.438
URL type 5	URL type 1	-1.651	.355	.000	-2.656	-.646
	URL type 2	.131	.367	1.000	-.909	1.170
	URL type 3	.331	.344	1.000	-.644	1.306
	URL type 4	-.512	.327	1.000	-1.438	.414

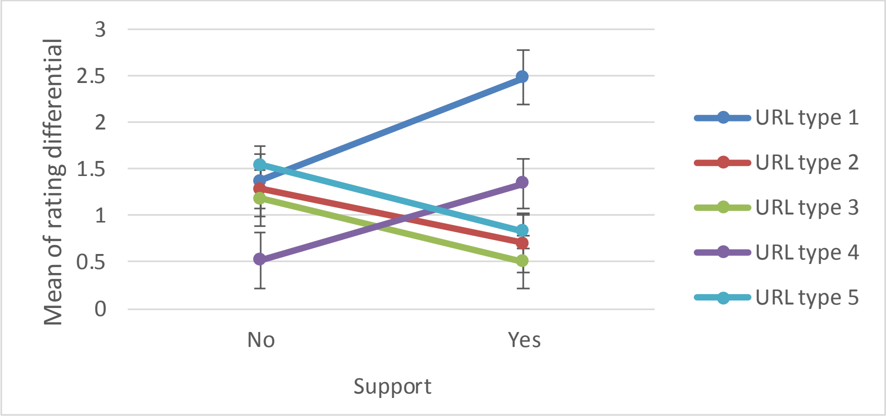


Figure 41. Profile graph of simple simple main effect of URL_type on rating differential at each level of Support

7.3.4.2 Simple main effects analysis in the situation of no training

When users are not trained, Support*URL_type has no interaction effect on rating differential. A test of simple main effect is carried out respectively to explore the effect of system support on users' rating differential and the difference among five types of URL.

a) Simple main effect of Support

Table 59 and Figure 42 show, regardless of types of URL, the mean of users' rating differential from experts in the situation of without and with system support. Without support, the mean is 1.57±0.13, and with support, the mean is 1.58±0.13; the difference is not significant, as shown in Table 60. (p=0.961). In conclusion, when users are not given training, regardless of types of URL, support has no significant effect on users' rating differential from experts.

Table 59. Mean of rating differential for Support overall URL_type

Support	Mean	Std. Error
No	1.57	.13
Yes	1.58	.13

Table 60. Test of simple main effect of Support

	Sum of Squares	df	Mean Square	F	Sig.
Support	.003	1	.003	.002	.961
Error	150.39	137	1.10		

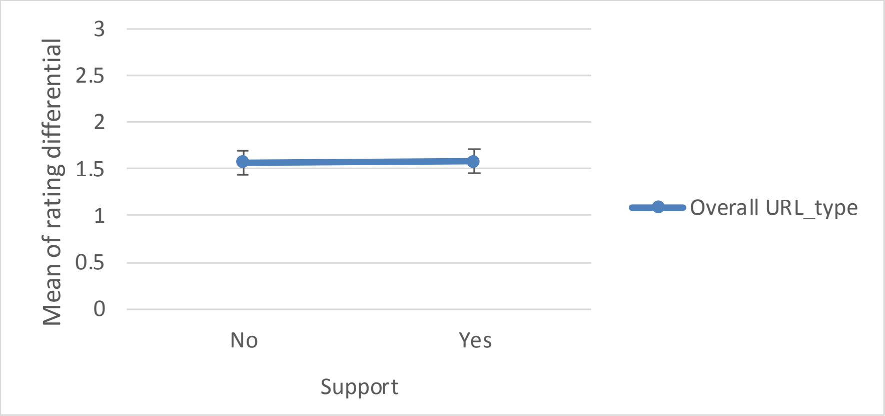


Figure 42. Profile graph of simple main effect of Support overall URL_type

b) Simple main effect of URL_type

Table 61 shows, regardless of system support, the mean of users' rating differential for every type of URL when those users are not given training. The means of rating differential among these five types of URL are significantly different as shown in Table 62 (p=0.000).

Table 63 further explains this situation through a pairwise comparison of users' rating differential among the five types of URL. Figure 43 visually indicates the simple main effect of URL_type overall support. The results indicates that the users' rating differential for URL type 1 is significantly higher than for the other four types of URL.

Table 61. Mean of rating differential for URL_type overall Support

URL_type	Mean	Std. Error
URL type 1	2.31	.21
URL type 2	1.49	.22
URL type 3	1.23	.21
URL type 4	1.71	.22
URL type 5	1.14	.15

Table 62. Test of simple main effect of URL_type overall Support

	Sum of Squares	df	Mean Square	F	Sig.
URL_type	25.44	4	6.36	5.79	.000
Error	150.39	137	1.10		

Table 63. Pairwise comparison among URL types without training

(I) URL_type	(J) URL_type	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
URL type 1	URL type 2	.819	.306	.083	-.054	1.692
	URL type 3	1.083	.302	.005	.220	1.946
	URL type 4	.600	.305	.510	-.269	1.469
	URL type 5	1.174	.259	.000	.436	1.911
URL type 2	URL type 1	-.819	.306	.083	-1.692	.054
	URL type 3	.264	.306	1.000	-.608	1.137
	URL type 4	-.219	.308	1.000	-1.098	.660
	URL type 5	.355	.263	1.000	-.394	1.104
URL type 3	URL type 1	-1.083	.302	.005	-1.946	-.220
	URL type 2	-.264	.306	1.000	-1.137	.608
	URL type 4	-.484	.305	1.000	-1.353	.386
	URL type 5	.090	.259	1.000	-.647	.828
URL type 4	URL type 1	-.600	.305	.510	-1.469	.269
	URL type 2	.219	.308	1.000	-.660	1.098
	URL type 3	.484	.305	1.000	-.386	1.353
	URL type 5	.574	.261	.296	-.171	1.319
URL type 5	URL type 1	-1.174	.259	.000	-1.911	-.436
	URL type 2	-.355	.263	1.000	-1.104	.394
	URL type 3	-.090	.259	1.000	-.828	.647
	URL type 4	-.574	.261	.296	-1.319	.171

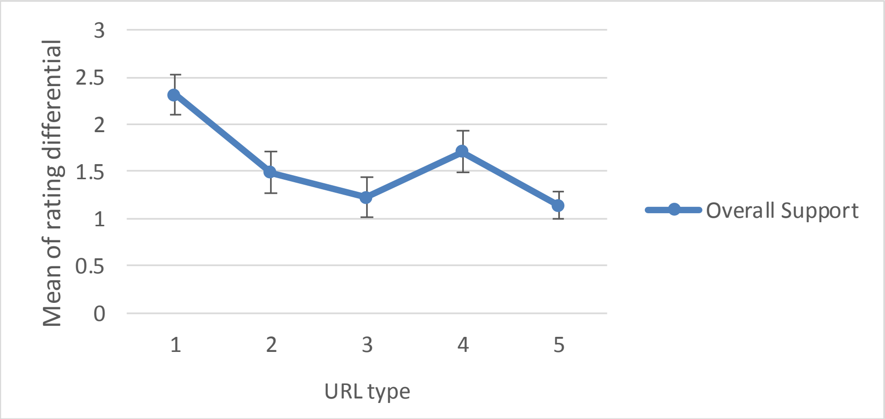


Figure 43. Profile graph of simple main effect of URL_type overall Support

7.3.5 Findings from evaluation results

From the evaluation results, the findings are summarised as below.

- For URL type 1, regardless of training, system support results in a significantly higher rating differential between users and experts. Regardless of support, training has no significant effects.
- For URL type 2, regardless of training, system support has no significant effect on users' rating differential from experts. Regardless of system support, training has no significant effect on users' rating differential.
- For URL type 3, regardless of training, system support has no significant effect on users' rating differential from experts. Regardless of system support, training has no significant effect on users' rating differential.
- For URL type 4, when users are trained, system support results in a significantly higher rating differential. When they are not trained, support has no significant effect. When users are not given support, training significantly reduces users' rating differential from experts. When they are given support, training has no significant effect.
- For URL type 5, regardless of training, support results in a significantly lower rating differential from experts. Regardless of support, training has no significant effect on rating differential.

7.4 Discussion of Evaluation Results

Findings from the evaluation experiment are listed in the last section. Three questions are put forward based on the findings.

1. It is expected that system support will close the gap between users' ratings and experts' ratings; in other words, there will be a lower rating differential between users and experts. However, for URL type 1, why does system support result in a higher rating differential between users and experts?
2. For URL type 4, training and system support have an interaction effect on the rating differential, but does this interaction effect occur for every URL in this type?
3. Why, for URL type 2 and URL type 3, does system support have no significant effect on users' rating differential from experts, but for URL type 5 system support results in a lower rating differential?

In the following sections, answers to these three questions are given and discussed.

7.4.1 Discussion of question 1

For URL type 1, system support results in a higher rating differential. The reason for this is that the system generated significantly different ratings from experts for URL type 1 compared with the other four types of URL. When users were given the system support, their decisions regarding URL type 1 were largely affected by the system. A one-way ANOVA test was conducted to investigate whether the rating differential between the system and experts for URL type 1 was significantly different from other types of URL. The results are displayed in Table 64. and Table 65. Table 65 shows that the system's rating is significantly different from experts' rating for specific types of URL (p=0.000). Figure 44 displays exactly this type of URL. Indeed, the graph confirms that, for URL type 1, system support has a significantly different rating from experts compared with the other four types of URL.

As stated in Chapter 5, system support yielded a much lower infringement rating than that given by experts for URL number 2 and URL number 16, which belong to URL type 1. For URL type 1, experts gave a higher rating because they found other infringement content on the webpage even if the content claimed by copyright owners did indeed not exist. Therefore, the CLC automation system generated a very low infringement score because neither the context information nor the allegedly infringing content could be found on the webpages. When the system was used by users, information such as title not matching, performer not matching and content not found was supplied to them. This information very likely led them to decide that there was a low probability of infringement for URL type 1.

Table 64. Mean of rating differential between system and experts for five types of URL

	N	Mean	Std. Deviation	Std. Error
URL type 1	24	1.50	.51	.10
URL type 2	24	.70	.31	.06
URL type 3	24	.40	.30	.06
URL type 4	28	.87	.72	.14
URL type 5	52	.79	.52	.07
Total	152	.84	.60	.05

Table 65. One-way ANOVA test for rating differential between five types of URL

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	15.72	4	3.93	15.16	.000
Within Groups	38.09	147	.26		
Total	53.81	151			

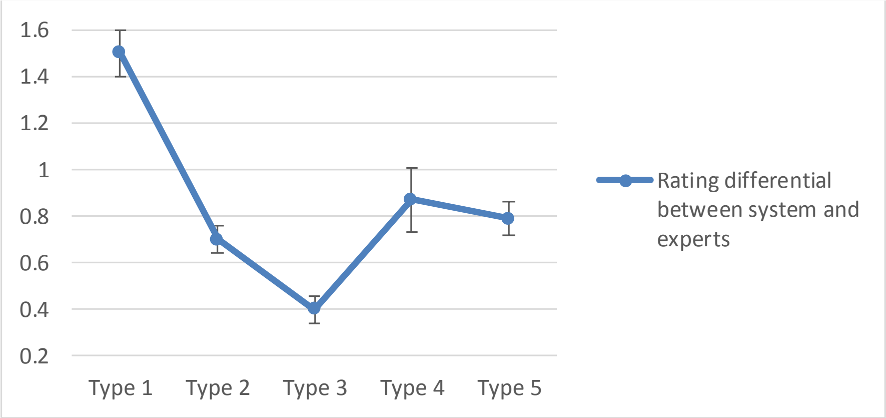


Figure 44. Line graph of rating differential between system and experts for five types of URL

7.4.2 Discussion of question 2

For URL type 4, training and system support have interaction effect on users' rating differential from experts. In order to find out whether this interaction effect happens to every URL in type 4, a three-way ANOVA analysis Training*Support*URL_Number was carried out. Table 66 shows the test results. The table shows that for URL type 4, the three-way interaction between training, support and URL_Number is not significant, F(4, 24) = 1.54, p= .222. In other words, the interaction effect of training and support works exactly the same for every URL in type 4.

Table 66. Three-way ANOVA test of between-subjects effects for Training*Support*URL_Number

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	73.82	1	73.82	120.86	.000
Training	6.57	1	6.57	10.76	.003
Support	1.34	1	1.34	2.20	.151
URL_Number	10.64	6	1.77	2.90	.028
Training * Support	5.08	1	5.08	8.31	.008
Training * URL_Number	4.79	6	.80	1.31	.292
Support * URL_Number	15.85	5	3.17	5.19	.002
Training * Support * URL_Number	3.76	4	.94	1.54	.222
Error	14.66	24	.61		
Total	147.57	49			

7.4.3 Discussion of question 3

For URL type 5, regardless of training, system support makes the users' ratings significantly closer to those of the experts. However, for URL type 2 and URL type 3, this is not the case, firstly because, for URL type 2 and URL type 3, the content cannot be accessed; moreover, the decisions made by users, systems and experts are mainly based on website suspicion. The effect of the system is not very strong in this case. In addition, for URL type 5, the system clearly tells users where the content is sourced from, which helps students to clarify whether the content is hosted or linked. Secondly, for URL type 2 and URL type 3, there is only one piece of allegedly infringing content showing on the webpage; for students, it may be more straightforward to make their decisions even without support. For URL type 5, some webpages contain only one piece of allegedly infringing content, while some contain a great deal of content, including the allegedly infringing content. In this situation, students may not have a clear idea of whether there is an infringement. As such, the system clearly tells them whether the content already exists and where the content is from. Therefore, for those webpages that contain a great deal of content, the system offers effective support which is capable of helping bring users' ratings significantly closer to those of the experts. The following data analysis is carried out to confirm the above explanation.

The URLs in type 5 were divided into two groups, with URLs in Group 1 only containing one piece of allegedly infringing content on the webpage, and URLs in Group 2 containing multiple pieces of content, including the allegedly infringing content.

- For users without system support, a one-way ANOVA test was carried out to compare their rating differential between these two URL groups. Table 67 and Table 68 show the results. Figure 45 visually displays the results by means of a line graph. Without support, users' rating differential from experts is not significantly different (p=0.177) between URL type 5.1 (one piece of content) and type 5.2 (multiple pieces of content).

Table 67. Mean of rating differential between users and experts for two sub types of URL type 5 without support

	N	Mean	Std. Deviation	Std. Error
Type 5.1	24	1.64	.98	.20
Type 5.2	27	1.27	.95	.18
Total	51	1.45	.97	.14

Table 68. One-way ANOVA test for rating differential between two sub types of URL without support

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.75	1	1.75	1.88	.177
Within Groups	45.78	49	.93		
Total	47.53	50			

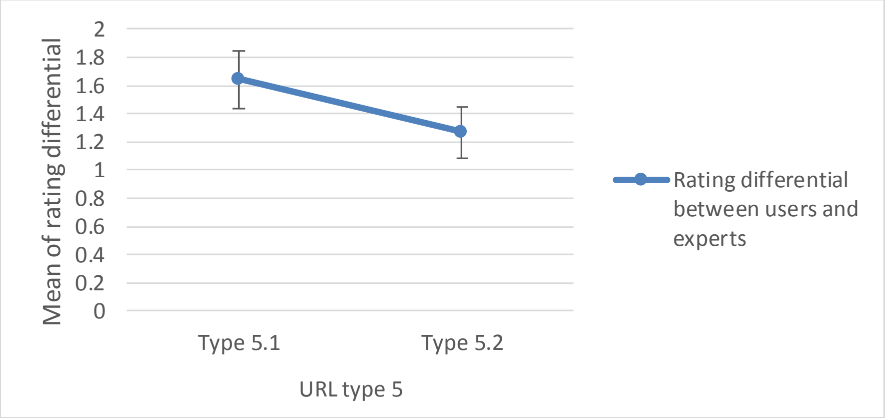


Figure 45. Line graph of rating differential between users and experts for two sub types of URL without support

- For the users with system support, another one-way ANOVA test was conducted to compare their rating differential between these two URL groups. Table 69 and Table 70 show the results. Figure 46 visually displays the results by means of a line graph. With system support, users' rating differential from experts is significantly different (p=0.022) between URL type 5.1 (one piece of content) and type 5.2 (multiple pieces of content). Moreover, users' ratings are significantly closer to the experts' ratings.

Table 69. Mean of rating differential between users and experts for two sub types of URL type 5 with support

	N	Mean	Std. Deviation	Std. Error
Type 5.1	24	1.27	1.01	.21
Type 5.2	28	.59	.39	.07
Total	52	.90	.81	.11

Table 70. One-way ANOVA test for rating differential between two sub types of URL with support

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5.99	1	5.99	10.93	.002
Within Groups	27.41	50	.55		
Total	33.40	51			

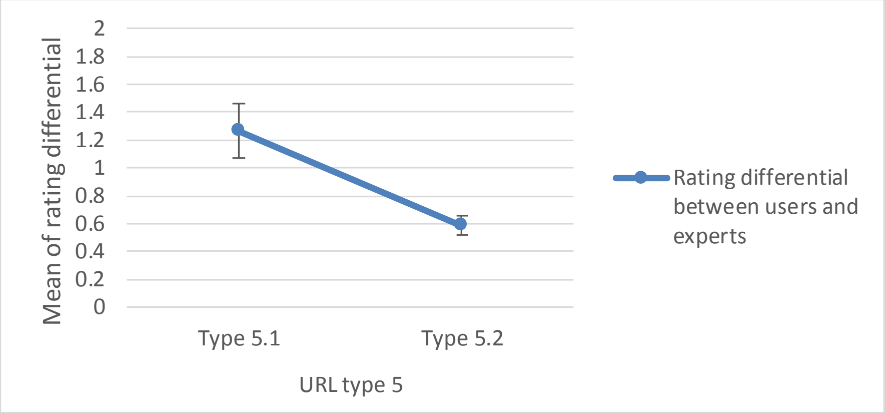


Figure 46. Line graph of rating differential between users and experts for two sub types of URL with support

7.5 Summary

In order to evaluate the CLC automation system, 24 PhD students majoring in Web Science were invited as users to participate in the study. They were divided into four groups depending on whether they were trained and had basic knowledge related to copyright in the context of the Web, and whether they were offered the system as a support tool. The same webpages that were given to experts in the validation process were presented to these users for viewing and examination.

During the evaluation process, quantitative data rating the differential between users and experts was analysed. The four groups were compared in terms of users' rating data and the effect of training and system support. The purpose of this was to find out whether the system helps to bring users' ratings closer to those of the experts. The evaluation results showed that the users' rating differential from experts depended on training, system support and types of URL. For URL type 1, system support significantly increased users' rating differential from experts. For URL type 4, system support and training had significant interaction effects on users' rating differential. For URL type 5, system support significantly reduced the users' rating differential.

The next chapter will discuss the research questions in the context of the main results, summarise the research and list future work.

Chapter 8 Discussion, Conclusion and Future Work

The issue of how to reform the notice-and-take-down procedure is the subject of intense discussion among legal professionals. Applying proper criteria to assess Web resources in removal requests in order to support notice receivers' decision-making process is essential to improve the procedure. A CLC Model was designed to represent 12 criteria and how these criteria operate for the analysis of allegedly infringing Web resources was indicated. A system to apply the CLC Model was also developed in order to automatically assess Web resources, and generate analytic results and, eventually, a score to indicate the likelihood of infringement with a view to supporting the copyright verification in the notice-and-take-down procedure. This chapter discusses and summarises the research work that has been carried out and outline the future work. Section 8.1 examines the research questions in the context of the main results and findings. Section 8.2 explores the limitations of the research, and Section 8.3 discusses the research implications. Section 8.3 concludes the research, and Section 8.4 lists the future work.

8.1 Answering the Research Questions

8.1.1 First research question

Question 1: What is an appropriate model that can be used to assess allegedly infringing content on webpages in the notice-and-take-down procedure?

To answer the question, the following three sub-questions are addressed:

Sub-question 1.1: What is the current state of allegedly infringing Web content and notice-and-take-down practice?

Sub-question 1.2: How can the model be developed?

Sub-question 1.2.1: What criteria should be considered in the model to assess whether a webpage contains copyright infringement content?

Sub-question 1.2.2: What is the workflow of the criteria in the model?

Sub-question 1.3: Is the model valid?

The first sub-question 1.1, was answered in Chapter 2 and Chapter 3. Google's practice and transparency report in the notice-and-take-down procedure were used as a benchmark to investigate the patterns of infringement through webpages. According to the literature review, the copyright infringement in this study was limited to a smaller scope. Four infringment scenarios and one non-infringement scenario were defined. According to the study on Google's transparency report pertaining to the state of allegely infringment URLs, five types of URL were summarised and listed. Research on sub-question 1.1 supplied a theoretical basis for the building of the CLC Model.

The second sub-question 1.2, was answered in Chapter 4 with reference to the research carried out in Chapter 2 and Chapter 3. A Content-Linking-Context Model for copyright related criteria, which will be applied to analyse the alleged infringing webpages, was developed. A total of 12 criteria were defined in the CLC Model. Criteria C1 to C3 considered whether the content existed and whether the content was a substantial copy of the original copyright work. Criteria C4 to C9 considered the linking issues of the content, such as whether the content was hosted by the current webpage or linked from external websites. Criteria C10 to C12 indicated the level of suspicion that a webpage contained infringing content. These criteria were connected together to assess Web resources step by step and to eventually generate a score to indicate the likelihood of infringement on webpages.

The third sub-question 1.3, was answered in Chapter 5. The usage of the 12 criteria developed in the CLC Model and the infringement score generated by the CLC Model were validated by 4 experienced experts. Quantitative and qualitative methods were used to analyse experts' data. The results showed that experts have a good agreement on using the 12 criteria developed in the model to assess Web resources. Moreover, the infringement score generated by the CLC Model was strongly correlated with the experts' ratings.

8.1.2 Second research question

Question 2: What is an appropriate automation system for applying the model to automatically assess allegedly infringing content on webpages in the notice-and-take-down procedure?

Sub-question 2.1: To what degree can the model be automatically implemented in the system?

Sub-question 2.2: How good is the automation system at supporting the assessment?

The first sub-question 2.1, was answered in Chapter 6. The investigation focused on implementing the CLC Model in an automated system to help assess Web resources. The main issue related to building the system was to implement the 12 criteria automatically. Related Web technologies and tools were explored in order to understand the degree to which each criterion can be automated. Exteranl reporitories, HTML analysis tools, HTTP client, and network traffic monitoring tools were proposed as solutions to automate each criterion. Because of the variety and fast development of Web technologies, not all the criteria were fully automated in the current version of the system.

The second sub-question 2.2, was answered in Chapter 7. The automation system is aimed at supporting online service providers in assessing infringement content on webpages. Thus, a user evaluation process was conducted to evaluate the effectiveness of the system. The sample of users who participated in the evaluation comprised PhD students majoring in Web Science. They were divided into four groups to give their infringement ratings on the webpages. Training and the system as a support tool were offered to different groups of URLs respectively. Quantitative data was collected from the four groups and compared in order to establish whether the system support made users' ratings closer to those of the experts. The evaluation results showed that, for specific types of URL, the system had a significant effect on helping bring users' ratings closer to those of the experts.

8.2 Challenges of the Research

Certain challenges were encountered even though the research fulfil its task of devising an appropriate model and an automation system to support online service providers in assessing allegedly infringing content on webpages in the notice-and-take-down procedure. These challenges are listed below:

- There is a lack of studies, publications and references in the interdisciplinary area of automatically assessing allegedly infringing Web resources in the context of the notice-and-take-down procedure. Indeed, only the Urban team and Seng have published related work. The Urban team (Urban and Quilter 2005; Karaganis and Urban 2015; Urban, Karaganis, and Schofield 2016) conducted an empirical study to outline the big picture of how online service providers and rights holders experience and practice notice-and-take-down on a day-to-day basis. The team also examined a random sample of notices to see who was sending notices, why, and whether they were valid take-down requests. Seng (Seng 2014) discussed, albeit fairly vaguely, what techniques were adopted by Google to deal with take-down requests. These studies are significant when it comes to helping understand the lifecycle of the notice-and-take-down procedure. However, there remains a lack of detailed information relating to take-down accuracy, how notice receivers check the lawfulness of allegedly infringing content, and what criteria and technologies they use for the action. Therefore, the challenge faced by this research is to start from the very beginning and to figure out the criteria and workflow for assessing the lawfulness of Web resources from the perspective of legal principles and requirements; indeed, this process must produce a technical solution.
- Given the variety and fast development of Web technologies used to present a webpage, it is difficult to fully automate certain criteria in the CLC Model. In other words, it is difficult to use one technical solution to cover all the Web presentation patterns identified by the criteria in the CLC Model. As stated in criteria C4 to C7 in Section 6.2.1, for the JavaScript-triggered cases, it is difficult to automatically locate the actual allegedly infringing file on the webpage. In this research, while the automation mechanism for these criteria were not fully implemented, the potential solutions were proposed such as using BrowserProxy to monitor the network traffic to recognise the existence of certain Web components in the CLC Model.

8.3 Implications of the Research

Linking issues have caused many discussions and debates for legal professionals. Where and how the allegedly infringing content is linked is also a major concern in the CLC Model. The user evaluation results indicates that the automation system significantly helps users' rating closer to experts on URL type 5 regarding linking issue. When users without any Web development experience view and examine webpages, they may not have correct understanding of where the content on the webpage is from and how the content is presented on the webpage. The system clarifies it and prompt users with clear information of the content sources, which largely helps them to make more accurate decision. So for a model or framework that used for assessing allegedly infringing content on webpages, it is believed linking is a necessary and important component.

In the CLC Model and automation system, the external database and repository are employed to indicate the answers to certain criteria. For example, AcoustID is integrated into the automation system to indicate the similarity between allegedly infringing music work and original copyright work. A more accurate result from AcoustID will directly lead to a more accurate output from the CLC automation system. Although the AcoustID service has been maintained by a community, many music records are still missing from its database, especially those in a language other than English, as well as records from less famous singers. As such, people cannot conclude that there is no copyright infringement if a high possibility match cannot be found from this database; it may simply be that the original record is missing from the repository. It is certainly worthwhile to have more databases similar to AcoustID, not only for the CLC automation system, but also for other copyright enforcement services. These databases can store the copyright-related information pertaining to a musical work, such as fingerprints, metadata etc. The data can be published in a structured format and linked together to form a reliable source of copyright work. Information on copyright work can be extracted for comparison and infringement detection.

As mentioned in Chapter 2, a review of notice-and-take-down system and how to reform it are currently being discussed, both in Europe and US. A research work undertaken by the Urban team has revealed that one of the most troubling problem in the current notice-and-take-down system is the high number of questionable notices sent by less accurate automatic systems or bad-faith senders. They suggest that better methods of preventing and remedying mistaken notices should be a high priority for reform. In fact, the CLC Model not only can be applied to verify allegedly copyright infringing content on webpages automatically, and thereby the validity of notices automatically. In addition, before notices are sent to online service providers, notice senders could use the criteria and the CLC system to help them filter out mistaken notices. The CLC workflow is relevant for both notice senders and notice receivers.

Google's transparency report was used as a benchmark in this research in order to understand the lifecycle of the notice-and-take-down procedure. Its domain data is also used to indicate the level of URL suspicion. The validation results from experts indicates that, in some cases, such as URL number 29, the accuracy of the level of URL suspicion is not as good as expectation. One potential solution is to employ and analyse more data from multiple online service providers. However, with the exception of Google, no similar transparency report was published by other online service providers (e.g. Facebook and Twitter), even if they had adopted the notice-and-take-down procedure. Another issue is the release of additional reports and information; indeed, this would promote transparency and be very valuable for legal professionals, researchers, online platforms and users when it comes to understanding the impact that copyright has on available content.

8.4 Conclusion

A huge amount of copyright-protected content, including books, pictures, music, and videos, is widely shared and distributed on the Internet. Internet intermediaries play an important role in facilitating the sharing and distribution of this content. The DMCA 1998 can be considered a milestone in regard to the way it imposes regulatory duties on Internet intermediaries by adopting a notice-and-take-down procedure and creating limitations in respect of their liability for copyright infringement. Although there exists no similar procedure in the EU, many online service providers follow this procedure and have developed technologies to apply it, e.g. an online anti-piracy system that can automatically issue DMCA take-down notices to infringing parties, and host providers or information location tools which receive notices, assess related Web resources and make decisions on take-down. During this process, it is important to understand the appropriate criteria used to examine the alleged infringing Web resources and, furthermore, to implement technologies which support the examination process automatically.

A preliminary study was conducted to understand the patterns of infringement activities through webpages. 730 URLs were manually examined, 528 of which were still available at the time of the experiment. Most of the claimed copyright works on the 528 webpages took the form of music, books, and video, and the majority of requests were sent from copyright detection agents rather than the owners of the copyright works. Most works were not hosted by the current website but instead linked from external sources. The preliminary result has shown that, of the 528 URLs Google decided to take down, 431 of these decisions were considered correct, i.e. the claimed copyright work was found on the webpages and so it was right to take them down; in contrast, there were only 7 URLs which should not have been taken down. With regard to the remaining 90 URLs, they were categorised as uncertian cases due to various reasons, such as the content being embedded from another source and a lack of clarity regarding whether the original source allows such embedment or whether the original source is legal. Based on the experiment results, the accuracy of take-down in the Google Transparency Report was calculated. Depending on whether different types of uncertainty were considered as correct decisions, the accuracy of the take-down actions for Google Search was between 81.6% and 98.7%. Consideration was first given to the features of the URLs, such as whether the allegedly infringing work exists on the webpage, whether the work can be accessed, and how the work can be accessed; based on this, the URLs reviewed were divided into five types, all of which are used in the later study.

Based on a literature review and analysis of the Google Transparency Report, which provides information regarding URLs and webpages with potentially copyright-infringing content, a Content-Linking-Context Model (CLC Model) was designed which contains 12 criteria (C1 to C12) to indicate different factors people should considered when verifying allegedly infringing Web resources in a notice. To build the CLC model, a conservative approach was taken and five conventional scenarios were defined. Following this, the criteria were organised into Content, Linking, and Context, and a workflow in the CLC Model was designed to connect each criterion for infringement assessment on webpages. In both the literature review and the experiment, the linking issues brought a lot of uncertainty to the assessment process, especially when the linking produced sources which crossed domains. Therefore, the CLC Model has considered linking as an important judging criterion, as it relates to the content presentation contained on the webpages and whether the owner of the webpage has the right to make the content available through the webpage. The purpose of the CLC Model is to support online service providers in their decision-making instead of replacing it; as such, the output of the CLC Model is a score used to indicate the likelihood of infringement.

The criteria developed in the CLC Model and the output of the model were validated through the experts' validation experiment. In total, 29 webpages which covered the 5 types of URL were given to 4 experienced experts for viewing and examination. Their answers were recorded and analysed. From the validation results, it is confirmed that the experts had a good level of agreement regarding the usage of criteria developed in the CLC Model. In addition to this, the infringement score generated by the CLC Model was strongly correlated with the experts' ratings.

An automation system which applies the CLC Model was developed. For each criterion in the CLC Model, the visual clues and the background technology implementation were analysed. Since all the URLs claimed in take-down notices are webpages developed using modern Web technologies, it is important to have a deep understanding of the major patterns according to which each criterion is implemented. Possible technical solutions to automate the assessment process were also proposed. External repositories were used to compare the similarity of audio content and calculate the value of URL suspicion. For other visual, content and context information presented on the webpage, HTML analysis tools, HTTP client, and network traffic monitoring tools to automate each criterion were used.

The automation system was evaluated by 24 users. The same webpages that were given to the experts were presented to the users for viewing and examination. Their answers were compared to those of the experts in order to establish whether the system helps bring their infringement decisions closer to those of the experts. From the evaluation results, the conclusion is that through clarifying linking issues, such as where and how the allegedly infringing content was presented on the webpage, the automation system significantly helps users to assess specific types of URL, such as URL type 5.

8.5 Future work

The work to be carried out in the future is listed below and broken down into three major parts:

- **Extend the CLC Model.**

The CLC Model currently works for music, but could be extended to videos or books. Existing criteria may be modified, and more criteria can be added to the model. Furthermore, each criterion can be further broken down into sub-workflow if necessary. For example, if books are considered for copyright infringement, criterion C5 can be changed to *Online readable*. This indicates whether the book can be read directly on the website. In the preliminary study, 16% of notices claimed were in relation to online books and most of their book pages (for example comic books) were images embedded from other websites or cloud services. In these cases, the linking issue is still a problem and the CLC Model works properly in this context. In the future work, the CLC Model could also be extended to consider exceptions such as fair use. For example, if an audio track matches the audio track of the same copyright work, more criteria such as detailed metadata including length of the audio and reliability of the source where the track is from can be adopted to further assess the probability of infringement.

- **Further research on the infringement score algorithm in the CLC Model**

As discussed in the sections related to validation of results and research implications, the external database and repository are adopted to generate the infringement score. A more accurate result from these external resources will directly lead to a more accurate output from the CLC Model and automation system. For content comparison, AcoustID was used in this research. The future work will investigate more libraries or similar databases in order to generate a more accurate content comparison score. In addition, Google transparency data was used in the research to identify the level of URL suspicion. In future work, more data sources and how to link them together to store comprehensive URL-related information will be explored. In addition, and as mentioned in the research implications section, gathering labels of more URLs from experts and using them to train a standard machine learning classifier could also be an effective method with which to generate a more accurate value of URL suspicion.

- **Further automatic implementation of the system**

In the CLC model, it is difficult to fully automate certain criteria, particularly for the JavaScript-triggered Web components, as stated in Section 6.2.1. Indeed, to supplement the use of the webpage information extraction and monitoring technologies proposed in Chapter 6, computer vision and machine learning technologies can be used in future work to recognise the existence of certain Web components in the CLC Model. A machine learning model will be developed and will take the screenshots of the different statuses of the webpage as input; following this, the machine will visually analyse whether components, such as the video player, play button, download button, login forms, etc., are likely to be presented to viewers.

References

Aikenhead, Michael. 1995. “Legal Knowledge Based Systems: Some Observations on the Future.” *Web Journal of Current Legal Issues* 2: 72.

Altman, Douglas G. 1990. *Practical Statistics for Medical Research*. CRC press.

“Architecture of the World Wide Web, Volume One.” http://www.w3.org/TR/webarch/.

Arezzo, Emanuela. 2014. “Hyperlinks and Making Available Right in the European Union--What Future for the Internet After Svensson?” *IIC-International Review of Intellectual Property and Competition Law* 45 (5). Springer: 524–55.

Ashley, Kevin D. 1992. “Case-Based Reasoning and Its Implications for Legal Expert Systems.” *Artificial Intelligence and Law* 1 (2-3). Springer: 113–208.

Australian Administrative Review Council. 2003. *Automated Assistance in Administrative Decision Making*.

Barlas, Yaman. 1994. “Model Validation in System Dynamics.” In *Proceedings of the 1994 International System Dynamics Conference*, 1–10.

Barlas, Yaman. 1996. “Formal Aspects of Model Validity and Validation in System Dynamics.” *System Dynamics Review* 12 (3): 183–210.

Barlett, James E, Joe W Kotrlik, and Chadwick C Higgins. 2001. “Organizational Research: Determining Appropriate Sample Size in Survey Research.” *Information Technology, Learning, and Performance Journal* 19 (1). Organizational Systems Research Association: 43.

Beebe, Barton. 2008. “An Empirical Study of US Copyright Fair Use Opinions, 1978-2005.” *University of Pennsylvania Law Review*. JSTOR, 549–624.

Bendrath, Ralf, and Milton Mueller. 2011. “The End of the Net as We Know It? Deep Packet Inspection and Internet Governance.” *New Media & Society* 13 (7). SAGE Publications: 1142–60.

Boloix, Germinal, and Pierre N Robillard. 1995. “A Software System Evaluation Framework.” *Computer* 28 (12). IEEE: 17–26.

Brace, Ian. 2008. *Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research*. Kogan Page Publishers.

Bridy, Annemarie. 2011. “Is Online Copyright Enforcement Scalable?” *Vanderbilt Journal of Entertainment & Technology Law* 13 (4): 695–737.

Carson, John S. 2002. “Model Verification and Validation.” In *Simulation Conference, 2002. Proceedings of the Winter*, 1:52–58.

Charrington, Dwayne. 2013. “Does Content ID Look for a Match of the Actual Code or a Match of the Audio Produced by the Code?” https://www.quora.com/Does-Content-ID-look-for-a-match-of-the-actual-code-or-a-match-of-the-audio-produced-by-the-code.

Clayton, Richard. 2005. “Anonymity and Traceability in Cyberspace.” University of Cambridge.

Cobia, Jeffrey. 2008. “Digital Millennium Copyright Act Takedown Notice Procedure: Misuses, Abuses, and Shortcomings of the Process, The.” *Minn. JL Sci. & Tech.* 10. HeinOnline: 387.

Cohen, Jacob. 1988. “Statistical Power Analysis for the Behavioral Sciences Lawrence Earlbaum Associates.” *Hillsdale, NJ*, 20–26.

Coombes, Hilary. 2001. *Research Using IT*. Palgrave Macmillan.

Cornish, William Rodolph, David Llewelyn, and Tanya Frances Aplin. 2010. *Intellectual Property : Patents , Copyright , Trade Marks and Allied Rights*. 7th ed. Sweet & Maxwell London.

Daniel, Wayne W, and Chad L Cross. 2013. *Biostatistics: A Foundation for Analysis in the Health Sciences*. John Wiley & Sons, Inc.

Dawes, John G. 2012. “Do Data Characteristics Change according to the Number of Scale Points Used?

An Experiment Using 5 Point, 7 Point and 10 Point Scales."

Deveci, H A. 2004. "Hyperlinks Oscillating at the Crossroads." *CTLR-OXFORD-* 10 (4). SWEET & MAXWELL: 82–94.

Digital Millennium Copyright Act, H.R. 2281, 105th Congress. 1998.

"Digital Rights Management (DRM): Media Companies' Next Flop." 2006. http://knowledge.wharton.upenn.edu/article/digital-rights-management-drm-media-companies-next-flop/.

Elkin-Koren, Niva. 2014. "After Twenty Years: Copyright Liability of Online Intermediaries." *The Evolution and Equilibrium of Copyright in the Digital Age (Susy Frankel & Daniel J Gervais eds.)(2014 Forthcoming)*.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2). Springer: 175–91.

Feiler, Lukas. 2012. *Website Blocking Injunctions under EU and US Copyright Law – Slow Death of the Global Internet or Emergence of the Rule of National Copyright Law?* 13. Transatlactic Technology Law Forum.

Felten, Edward W. 2003. "A Skeptical View of DRM and Fair Use." *Communications of the ACM* 46 (4): 56–59.

Field, Andy. 2013. *Discovering Statistics Using IBM SPSS Statistics*. 4th ed. Sage.

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.

Fox, Nick, and Nigel Mathers. 1997. "Empowering Research: Statistical Power in General Practice Research." *Family Practice* 14 (4). Oxford Univ Press: 324–29.

Frystyk, Henrik. 1994. "The Internet Protocol Stack." http://www.w3.org/People/Frystyk/thesis/TcpIp.html.

Google. 2013. *How Google Fights Piracy*.

Google, and PRS for Music. 2012. *The Six Business Models for Copyright Infringement - A Data-Driven Study of Websites Considered to Be Infringing Copyright*.

"Google Transparency Report - Domain." 2015. Accessed February 8. http://www.google.com/transparencyreport/removals/copyright/domains/vmusice.net/.

"Google Transparency Report - Explore the Data." 2016. Accessed December 1. http://www.google.com/transparencyreport/removals/copyright/domains/?r=all-time.

"Google Transparency Report - FAQ." 2016a. Accessed December 1. https://www.google.com/transparencyreport/removals/copyright/faq/#response_time.

"Google Transparency Report - FAQ." 2016b. Accessed December 1. http://www.google.com/transparencyreport/removals/copyright/faq/#abusive_copyright_requests.

Green, G I, and R T Keim. 1983. "After Implementation Whats Next-Evaluation." *Journal of Systems Management* 34 (9). ASSOC SYSTEMS MANAGEMENT 24587 BAGLEY RD, CLEVELAND, OH 44138: 10–15.

Hamilton, Scott, and Norman L Chervany. 1981. "Evaluating Information System Effectiveness-Part I: Comparing Evaluation Approaches." *MIS Quarterly*. JSTOR, 55–69.

Hargreaves, Ian. 2011. "Digital Opportunity: A Review of Intellectual Property and Growth: An Independent Report." Intellectual Property Office.

"HTML5- A Vocabulary and Associated APIs for HTML and XHTML." http://www.w3.org/TR/html5/links.html.

"Hypertext Transfer Protocol -- HTTP/1.1: 10 Status Code Definitions."

"Hypertext Transfer Protocol -- HTTP/1.1: 5 Request." https://www.w3.org/Protocols/rfc2616/rfc2616-sec5.html.

IBM. "IBM SPSS."

IPL. 2013. *A Report by IPL for Google - Modelling the Takedown Process*.

Jick, Todd D. 1979. "Mixing Qualitative and Quantitative Methods: Triangulation in Action." *Administrative Science Quarterly* 24 (4). JSTOR: 602–11.

Jørgensen, Magne. 2004. "A Review of Studies on Expert Estimation of Software Development Effort." *Journal of Systems and Software* 70 (1). Elsevier: 37–60.

Justice, I P, Electronic Frontier Finland, Vereniging Open Source Nederland, and Cory Doctorow. "Digital Rights Management: A Failure in the Developed World, a Danger to the Developing World."

Karaganis, Joe, and Jennifer Urban. 2015. "The Rise of the Robo Notice." *Communications of the ACM* 58 (9). ACM: 28–30.

Kesan, Jay P, and Rajiv C Shah. 2003. "Deconstructing Code." *Yale JL & Tech.* 6. HeinOnline: 277.

Kim, Jin. 2012. "The Institutionalization of YouTube: From User-Generated Content to Professionally Generated Content." *Media, Culture & Society* 34 (1). Sage Publications: 53–67.

Kraemer, Helena C. 1982. "Kappa Coefficient." *Wiley StatsRef: Statistics Reference Online*. Wiley Online Library.

Ku, William, and Chi-Hung Chi. 2004. "Survey on the Technological Aspects of Digital Rights Management." *Information Security*. Springer, 391–403.

Kuczerawy, Aleksandra. 2015. "Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative." *Computer Law & Security Review* 31 (1). Elsevier: 46–56.

Kumar, Kuldeep. 1990. "Post Implementation Evaluation of Computer-Based Information Systems: Current Practices." *Communications of the ACM* 33 (2). ACM: 203–12.

Landis, J Richard, and Gary G Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics*. JSTOR, 159–74.

Lannella, Renato. 2001. "Digital Rights Management(DRM) Architecture." http://www.dlib.org/dlib/june01/iannella/06iannella.html.

Latham, Robert P, Carol C Butzer, and Jeremy T Brown. 2008. "Legal Implications of User--Generated Content: YouTube, MySpace, Facebook." *Intellectual Property & Technology Law Journal* 20 (5): 1–11.

Lauinger, Tobias, Martin Szydlowski, Kaan Onarlioglu, Gilbert Wondracek, Engin Kirda, and Christopher Kruegel. 2013. "Clickonomics: Determining the Effect of Anti-Piracy Measures for One-Click Hosting." In *NDSS*.

Leiser, Mark. 2013. "The Copyright Issue and Censorship Threat Buried within Google's Transparency Report." http://www.thedrum.com/news/2013/12/23/copyright-issue-and-censorship-threat-buried-within-googles-transparency-report.

Lemley, Mark A, and R Anthony Reese. 2004. "Reducing Digital Copyright Infringement without Restricting Innovation." *Stanford Law Review*. JSTOR, 1345–1434.

Lessig, Lawrence. 2009. *Code: And Other Laws of Cyberspace*. ReadHowYouWant. com.

Lindsay, David. 2000. *Copyright Infringement via the Internet: The Liability of Intermediaries*.

Lowry, Richard. "VassarStats: Website for Statistical Computation." http://vassarstats.net/index.html.

MacCallum, Robert C, Keith F Widaman, Shaobo Zhang, and Sehee Hong. 1999. "Sample Size in Factor Analysis." *Psychological Methods* 4 (1). American Psychological Association: 84.

"Media Identification." 2017. Accessed April 30. http://www.audiblemagic.com/media-identification/.

Miller, Claire Cain. 2010. "YouTube Ads Turn Videos into Revenue." *New York Times* 2.

Murray, Andrew. 2010. *Information Technology Law: The Law and Society*. Oxford University Press.

Newman, Isadore, Janine Lim, and Fernanda Pineda. 2013. "Content Validity Using a Mixed Methods Approach: Its Application and Development through the Use of a Table of Specifications Methodology." *Journal of Mixed Methods Research* 7 (3). SAGE Publications Sage CA: Los Angeles, CA: 243–60.

Pal, Kamalendu, and John Campbell. 1998. "ASHSD-II: A Computational Model for Litigation Support." *Expert Systems* 15 (3). Wiley Online Library: 169–81.

Perel, Maayan, and Niva Elkin-Koren. 2016. "Accountability in Algorithmic Copyright Enforcement."

Popple, James. 1993. "SHYSTER: A Pragmatic Legal Expert System." Australian National University. doi:http://dx.doi.org/10.2139/ssrn.1335637.

Reichman, Jerome H, Graeme B Dinwoodie, and Pamela Samuelson. 2007. "Reverse Notice and Takedown Regime to Enable Pubic Interest Uses of Technically Protected Copyrighted Works, A." *Berkeley Tech. LJ* 22. HeinOnline: 981.

Reidenberg, Joel R. 1997. "Lex Informatica: The Formulation of Information Policy Rules through Technology." *Tex. L. Rev.* 76. HeinOnline: 553.

Richardson, T. 2004. "ISPA Seeks Anaysis of BT's 'CleanFeed' Stats: Web Filtering Figures 'Could Be Misleading.'" *The Register, July* 21: 2004.

Ricketson, Sam. 1987. *The Berne Convention for the Protection of Literary and Artistic Works: 1886-1986*. Centre for Commercial Law Studies, Queen Mary College.

Ricketson, Sam, and Jane C Ginsburg. 2005. *International Copyright and Neighbouring Rights, v 1-2*. Oxford Univ. Press.

Rosati, Eleonora. 2016. "Do Machines Work Better than Humans? You Can Find out by Helping Research!" http://the1709blog.blogspot.co.uk/2016/08/do-machines-work-better-than-humans-you.html.

Rosati, Eleonora. 2017. "The CJEU Pirate Bay Judgment and Its Impact on the Liability of Online Platforms." European Intellectual Property Review, Forthcoming. Available at SSRN: https://ssrn.com/abstract=3006591

Rosati, Eleonora, and Oliver Löffel. 2014. "That BestWater Order: It's up to the Rightholders to Monitor Online Use of Their Works." http://ipkitten.blogspot.fr/2014/10/that-bestwater-order-its-up-to.html.

Rosenblatt, Bill. 2008. *Content Identification Technologies - Business Benefits for Content Owners*. Las Vegas.

Rowland, Diane, Uta Kohl, and Andrew Charlesworth. 2010. "Information Technology Law 4/e."

Samuelson, Pamela. 1993. "Fair Use for Computer Programs and Other Copyrightable Works in Digital Form: The Implications of Sony, Galoob and Sega." *J. Intell. Prop. L.* 1. HeinOnline: 49.

Sargent, Robert G. 2005. "Verification and Validation of Simulation Models." In *Proceedings of the 37th Conference on Winter Simulation*, 130–43.

Schauer, Frederick. 1991. *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Clarendon Press.

Schlesinger, Stewart. 1979. "Terminology for Model Credibility." *Simulation* 32 (3): 103–4.

Schuman, Howard, and Stanley Presser. 1979. "The Open and Closed Question." *American Sociological Review*. JSTOR, 692–712.

Seltzer, Wendy. 2001. "Lumen Database."

Seng, Daniel. 2014. "The State of the Discordant Union: An Empirical Analysis of DMCA Takedown

Notices." *Virginia Journal of Law and Technology, Forthcoming*.

Singh, Prabhishek, and R S Chadha. 2013. "A Survey of Digital Watermarking Techniques, Applications and Attacks." *International Journal of Engineering and Innovative Technology (IJEIT)* 2 (9).

Stalla-Bourdillon, Sophie. 2012a. "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.

Stalla-Bourdillon, Sophie. 2012b. "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." *Journal of International Commercial Law and Technology* 7 (2).

Stalla-Bourdillon, Sophie. 2016. "Open Letter to the European Commission – On the Importance of Preserving the Consistency and Integrity of the EU Acquis Relating to Content Monitoring within the Information Society." *Peep Beep!* https://peepbeep.wordpress.com/2016/10/10/open-letter-to-the-european-commission-on-the-importance-of-preserving-the-consistency-and-integrity-of-the-eu-acquis-relating-to-content-monitoring-within-the-information-society/.

"Status Code." 2016. Accessed March 15. http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html.

Stone, Brad, and Miguel Helft. 2007. "New Weapon in Web War over Piracy." *The New York Times, Section C, Column* 6.

Subramanya, S R, and Byung K Yi. 2006. "Digital Rights Management." *IEEE Potentials* 25 (2). IEEE: 31–34.

Sung, Chunhsien, and Po-Hsian Huang. 2014. "Copyright Infringement and Users of P2P Networks in Multimedia Applications: The Case of the US Copyright Regime." *Peer-to-Peer Networking and Applications* 7 (1). Springer: 31–40.

Tushnet, Rebecca. 2013. "PTO/NTIA: Notice and Takedown- Improving the Operation of the Notice and Takedown System." http://tushnet.blogspot.co.uk/2013/12/ptontia-notice-and-takedown.html.

Twentieth Century Fox Film Corp v BT [2011] EWHC 1981 (Ch).

Urban, Jennifer M, Joe Karaganis, and Brianna L Schofield. 2016. "Notice and Takedown in Everyday Practice." *Available at SSRN 2755628*.

Urban, Jennifer M, and Laura Quilter. 2005. "Efficient Process or Chilling Effects-Takedown Notices under Section 512 of the Digital Millennium Copyright Act." *Santa Clara Computer & High Tech. LJ* 22. HeinOnline: 621.

Van der Sar, Ernesto. 2014. "Embedding Is Not Copyright Infringement, EU Court Rules." https://torrentfreak.com/embedding-copyright-infringement-eu-court-rules-141025/.

Wallach, Dan S. 2001. "Copy Protection Technology Is Doomed." *Computer* 34 (10). IEEE: 48–49.

Weiss, Carol H. 1972. "Methods for Assessing Program Effectiveness." *Englewood Cliffs*.

Zeleznikow, John, and Andrew Stranieri. 1995. "The Split-up System: Integrating Neural Networks and Rule-Based Reasoning in the Legal Domain." In *Proceedings of the 5th International Conference on Artificial Intelligence and Law*, 185–94.

References - Cases

A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004. 2001.

CJEU C-128/11 Usedsoft GmbH v Oracle International Corp, 3 July 2012 ECLI:EU:C:2012:407.

CJEU C-160/15 GS Media BV v Sanoma Media Netherlands BV and Others, 8 September 2016 ECLI:EU:C:2016:644.

CJEU C-348/13 BestWater International GmbH v Michael Mebes and Stefan Potsch of 21 October 2014 ECLI:EU:C:2014:2315.

CJEU C-360/10 Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV, 16 February 2012 ECLI:EU:C:2012:85.

CJEU C-466/12 Nils Svensson et al v Retriever Sverige AB, 13 February 2014 ECLI:EU:C:2014:76.

CJEU C-527/15 Stichting Brein v Jack Frederik Wullems, 26 April 2017 ECLI:EU:C:2017:300.

CJEU C-610/15 Stichting Brein v Ziggo BV and XS4All Internet BV, 14 June 2017 ECLI:EU:C:2017:456

CJEU C-70/10 Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM), 24 November 2011 ECLI:EU:C:2011:771.

Lenz v. Universal Music Corp., 801 F.3d 1126. 2015.

Appendix A: Questionnaire for Expert Validation

Study of copyright infringement in the notice-and-take-down procedures

Researcher: Pei Zhang

Ethics Reference Number: ERGO/FPSE/22074

The purpose of this research is to automatically help verify copyright infringing activity on webpages in the notice-and-take-down procedures. We designed a Content-Linking-Context (CLC) Model to represent a serial of criteria and indicate how these criteria operate for the analysis of allegedly infringing Web resources. We also investigated how a system can be implemented to apply this model in an automatic manner.

Your answers and opinions in the questionnaire will be used to refine my model. Your attendance is highly appreciated. Your responses will be treated as anonymous and used for this research purpose only. Many thanks for your time and help.

Introduction

Some online service providers such as Google, Twitter etc receive notifications from copyright owners, that a webpage contains copyright-infringing content. A URL, in the form of *www.sample.com/path/page.html*, is used to identify the webpage. The notification requests that the URL is removed from the service provider's server or from the search results they publish.

In this study, you will be given several URLs. For each URL, you will be given key information about a copyright work (music) that may be accessible through that URL. This information will include the title of the music, the name of the performer, and the name of the copyright owner. After you have viewed and examined the webpage which the given URL points to, you will be asked three questions. **Question 1** asks you to give your rating about how likely you think the webpage infringes the copyright that is claimed by copyright owner. In answering the questions, you might use some criteria to inform and support your decision. **Question 2** asks you to fill in a table. The first column *Criterion* lists candidate criteria; you need to indicate whether you used each criterion (Yes/No) in the third column. The fourth column *Question* list some questions, you need to give answers (Yes/Not sure/No) for each question in the following column. You can also add your own criteria at the end of the table.

Questionnaire

Firstly, please open the document ***URLs.docx,*** which lists the information relating to allegedly infringing copyright works (*URL reference number*, *URL*, *Title of copyright work*, *Performer* and *copyright owner*). Secondly please write down the *URL reference number* below, which indicates what URL you are about to view and examine. Thirdly, please click the related URL and open the webpage to start your viewing and examination. Finally, after your viewing and examination, please answer Questions 1 to 3.

URL reference number:____________________

Question 1

The URL points to a webpage. How likely do you think the page contains copyright infringing content? Please select **ONE** number from a 5-point scale that best describes your opinion of the likelihood of copyright infringement.

Very low	**1**	**2**	**3**	**4**	**5**	**Very high**
	●	●	●	●	●	

Question 2

While viewing and examining the webpage, you may have used criteria to arrive at your decision. Some criteria are listed in the following table. For each criterion, please answer the two sets of questions (**Question A** and **Question B**) about it.

You can also **add your own criteria** at the end of the table.

Criterion	**Question A**	**Answer A (Yes/No)**	**Question B**	**Answer B (Yes/No)**
Music title matching. Whether the music title in question matches any music title shown on the webpage.	Did you use the criterion?		Did you see the music title on the webpage?	
Music performer matching. Whether the name of the music performer in question matches any name of performers shown on the webpage.	Did you use the criterion?		Did you see the name of the performer on the webpage?	
User interface for downloading. **Whether the webpage provides a user interface to enable users to download the music. For example, a 'Download' button shown on the webpage.**	Did you use the criterion?		Did you see a button or a clickable text or an icon for downloading the music?	
Music downloadable.	Did you use the criterion?		i) Did you click the button or the clickable	

Whether the music can be downloaded.			text or the icon to download the music?	
			ii) If you click any button or text to download the music, did you eventually successfully download it?	
User interface for playing the music online. **Whether the webpage provides a user interface to enable users to play the music online. For example, a 'Play' button shown on the webpage.**	Did you use the criterion?		Did you see a button or a clickable text or icon for playing the music online?	
Online Playable. **Whether the music can be played online?**	Did you use the criterion?		**i)** Did you click the button or the clickable text or the icon to play the music?	
			ii) Did you think the music can be played?	
Music content matching. **Whether there is a substantial similarity between the music in question with the original copyright music.**	Did you use the criterion?		**i)** Did you listen to the music?	
			ii) Did you think the music on the webpage substantially matches the original copyright music? (You can go to YouTube to listen to the original music if you like.)	
Music file source. **Whether the music is hosted on the current website or hosted on other websites.**	Did you use the criterion?		**i)** Did you think the music in question is hosted by the current website?	
			ii) Did you think the music in question is hosted by other website, and the current website only supply links?	
			iii) Did you think the music in question is hosted on other website, but it is embedded into the current website and can be played on the current website?	

URL suspicion. **The likelihood that the current website contains allegedly infringing content. For example, "fileshare.com" is a file sharing website which contains a large amount of infringing content. When the URL points to this domain, we may suspect an infringement.**	Did you use the criterion?		Did you think the webpage you viewed is suspicious?	
Other criteria you considered:				
Other criteria you considered:				
Other criteria you considered:				

Other comments:

Appendix B: Questionnaire for User Evaluation

Study of copyright infringement in the notice-and-take-down procedures

Researcher: Pei Zhang

Ethics Reference Number: ERGO/FPSE/22074

The purpose of this research is to automatically help verify copyright infringing activity on webpages in the notice-and-take-down procedures. We designed a Content-Linking-Context (CLC) Model to represent a serial of criteria and indicate how these criteria operate for the analysis of allegedly infringing Web resources. We also investigated how a system can be implemented to apply this model in an automatic manner.

Your answers and opinions in the questionnaire will be used to refine my model. Your attendance is highly appreciated. Your responses will be treated as anonymous and used for this research purpose only. Many thanks for your time and help.

Introduction

Some online service providers such as Google, Twitter etc receive notifications from copyright owners, that a webpage contains copyright-infringing content. A URL, in the form of *www.sample.com/path/page.html*, is used to identify the webpage. The notification requests that the URL is removed from the service provider's server or from the search results they publish.

In this study, you will be given several URLs. For each URL, you will be given key information about a copyright work (music) that may be accessible through that URL. This information will include the title of the music, the name of the performer, and the name of the copyright owner. After you have viewed and examined the webpage which the given URL points to, you will be asked to give your rating about how likely you think the webpage infringes the copyright that is claimed by copyright owner.

Questionnaire

Firstly, please open the document ***URLs.docx,*** which lists the information relating to allegedly infringing copyright works (*URL reference number*, *URL*, *Title of copyright work*, *Performer* and *copyright owner*). Secondly please write down the *URL reference number* below, which indicates what URL you are about to view and examine. Thirdly, please click the related URL and open the webpage to start your viewing and examination. Finally, after your viewing and examination, please answer Questions 1.

URL reference number:____________________

Question 1

The URL points to a webpage. How likely do you think the page contains the infringing content that claimed by the copyright owner? Please select **ONE** number from a 5-point scale that best describes your opinion of the likelihood of copyright infringement.

Very low	**1**	**2**	**3**	**4**	**5**	**Very high**
	●	●	●	●	●	

Other comments: