volume 12 issue 23/2023

VIEW

JOURNAL OF EUROPEAN TELEVISION HISTORY & CULTURE

WE WANT YOUR TOOLS! OR DO WE?

ON DIGITIZED CULTURAL HERITAGE ARCHIVES AND COMMERCIAL CONTENT IDENTIFICATION TOOLS

Maria Eriksson
Basel University
maria.eriksson@unibas.ch

Abstract: This article reflects on the technical gap that exists between academic and corporate capacities to study how digitized cultural heritage is reused online. In the context of tracing how audiovisual archival content is remixed and reinserted into new cultural contexts, the article asks what it would mean for humanistic researchers—and cultural heritage institutions more broadly—to utilize content identification tools provided by actors such as Google. How could commercial techniques for policing copyrights and tracing the whereabouts of online content be re-purposed to assist in research concerning remix practices and transformed cultural memories? What technical and legal consequences would such partnerships yield? And would such collaborations be ethical and scientifically defendable in the first place? Ultimately, the article reflects on the legal and technical discrepancies that exist between academic and commercial actors when it comes to monitoring how cultural content moves online. It also asks questions about what it means to care for digitized heritage collections in the 21st century.

Keywords: cultural heritage, remix, digital methods, content ID, video reuse

1 Introduction

In 2018, six colleagues of mine received an EU Horizon 2020 grant for a research project entitled *European History Reloaded: Curation and Appropriation of Digital Audiovisual Heritage* (or CADEAH)—a project to which I would later become associated. The project set out to study the online circulation and re-use of digitized audiovisual content by combining digital tracking technologies, critical cultural analysis, and ethnographic fieldwork. During the past decade, a massive body of audiovisual heritage has been digitized and made openly accessible online through initiatives such as Europeana and EUscreen, which currently host an abundance of video content covering European and world history. However, surprisingly little research has explored what happens to audiovisual cultural heritage once it is made freely and publicly available online. We were curious to study how the digitization of European history fosters new forms of participatory engagements among the public. We also wanted to explore how the meaning of digitized audiovisual heritage changes when it is re-appropriated and re-injected into new cultural contexts online.

A key part of the research was planned to involve the use of digital methods for tracking *how, when,* and *where* digitized audiovisual cultural heritage collections are remixed and re-used. Using technologies such as audiovisual fingerprint techniques and machine vision tools we were, for instance, interested in exploring how digitized archival content is picked up by amateur historians on YouTube, or discovered by subcultures such as the Vaporwar scene, which specializes in remixing historical war footage using a VHS-video aesthetic[1]. How can the "cultural biographies"[2] and "social life"[3] of digitized audiovisual cultural heritage collections be traced and tracked in the online domain? In what myriad ways is cultural heritage currently being re-used on digital platforms? Or, perhaps most importantly, is it appreciated and re-used at all? The prospect of finding answers to such questions were not just of interest to us as scholars of history, media, and digital culture but also caught the attention of several online museums, as demonstrated by the fact that four different European digital archives and cultural heritage institutions—EUROCLIO, Europeana, the EUscreen Foundation, and the Netherlands Institute for Sound and Vision—joined the research project as partners.

As we quickly realized, however, following how digitized audiovisual content moves on the internet was far more difficult than expected. While commercial actors like Google have developed advanced systems for scanning and identifying the whereabouts cultural content online, open-source solutions for tracing and tracking how digital content is reused are scarce.[4] As a result, Google is capable of continuously scanning more than 400 hours of user-uploaded YouTube videos *per minute* in search of cultural reuse[5], yet scholars who want to monitor how digitized cultural heritage is re-purposed online quickly run into technical, legal, and ethical problems. Such problems include gaining access to the considerable hardware resources needed to analyze large amounts of web content, navigating thorny regulations regarding the scraping, collection, and storage of online data, and thinking through the wider dilemmas of using—or not using—corporate software in humanistic research.

Against this background, the following article reflects on the fundamental technical gap that exists between academic and corporate capacities to study how our digitized collective memory is evolving and being re-shaped online. Given the legal and technical difficulties of scanning the entire internet—or at the very least, studying selected parts of it—the question arises as to whether humanistic scholars wanting to study cultural re-use should strive to collaborate with commercial actors who have access to advanced content identification techniques. What would it mean for humanistic researchers—and cultural heritage institutions more broadly—to utilize the content identification tools provided by actors such as Google? What technical and legal consequences would such partnerships yield? And would such collaborations be ethical and scientifically defendable in the first place?

By considering these questions, I hope to shed light on a discrepancy that exists between academic and corporate access to advanced content identification techniques. At the moment, it is not far-fetched to speak of a "digital divide"[6] with regards to who has the possibility to view and study online cultural reuse at scale. While actors such as Google are identifying and tracking the circulation of mind-boggling amounts of digital content each and every day, round the clock, minute by minute, the possibility for archivists and researchers to do the same is seriously limited. This, I argue, has great consequences for who has the possibility to grasp how popular historic narratives and new ways of dealing with the past are unfolding in the online domain. It also awakens long-standing debates concerning the ethics of academic/corporate collaborations and the need to safeguard the independence of academic research.

Ultimately, I suggest that scholarly access to corporate content identification tools (including commercial techniques for machine vision, reading, and listening) must be seen as a double-edged sword. On the one hand, the possibility of accessing cutting-edge corporate content identification tools opens radically new possibilities for critical humanistic research and provides fascinating opportunities to study how our shared cultural memory transforms through time. On the other hand, the decision to utilize toolkits developed by actors such as Google raises fundamental ethical questions regarding how to best care for and protect our collective digitized cultural heritage.

2 On Corporate Tools for Monitoring Cultural Reuse

In the recent decades, automated technologies for identifying cultural content online have been rapidly developed and expanded into an industry on its own. The origins of this technical area of expertise can largely be found in efforts to safeguard copyrights in the online domain. Following early 21st century debates and controversies concerning online piracy, several online platforms and corporations began developing technical solutions for controlling and/or monitoring how users share, remix, and interact with copyright protected content online. Commonly captured under the headline Digital Rights Management tools, these techniques have come to involve the development of encryption technologies, digital watermarking and hashing methods, and software standards that shape users' ability to copy and circulate cultural content online[7]. With time, they have also come to involve the development of content identification techniques that automate the process of scanning user generated content in search for copyright abuse. While originally developed and applied to protect the interest of rightsowners, these tools are—in essence—also techniques that allow for studying cultural reuse at scale. In short, they keep track of how vast amounts of cultural content is reused and circulates online.

Since 2007, for instance, Google has invested more than $100 million in the development of **Content ID**, which is currently one of the world's most advanced and widely used technologies for automatic content identification[8]. Content ID scans all user-uploaded YouTube videos prior to their publication in search for instances of copyright abuse and is at the heart of YouTube's (and Google's) efforts to help content owners safeguard copyrights. The tool is offered as a service to those wanting to remove copyright protected content from the web, or to those who are interested in monetizing videos when someone uploads copies of their works online. In 2018, Google claimed that Content ID was responsible for handling more than 98 percent of the copyright disputes that took place on YouTube, meaning that the technology plays a key role in controlling and overseeing how content enters and circulates on one of the world's largest video websites[9].

Elsewhere, Facebook relies on the technological expertise of the company **Audible Magic** for the development and provision of its **Rights Manager System**, which allows cultural creators and publishers to administer *where* and *how* their content appears on the website. Audible Magic—which also sells their services to major platforms like Vimeo and Twitch—has developed and licensed automatic content recognition tools since 2000, and specializes in the real-time identification and classification of recorded speech, videos, and music. In 2020, Audible Magic claimed to have the capacity to identify over 25 million "media assets" stemming from 1000 video suppliers and 140,000 record labels worldwide[10]. Aside from overseeing content on Facebook, Audible Magic monitors where its "assets" (presumably individual video files or music recordings) appear on live television broadcasts, live streaming platforms, archived TV, and the motion pictures.

Google and Audible Magic's development and use of content identification techniques illustrate commercial capacities to monitor how large amounts of cultural content circulates online. At the same time, however, the possibility to trace the whereabouts of audiovisual cultural heritage remains firmly out of reach for most academics and archivists. While recent years has seen a growing interest in the use of digital methods to analyze audiovisual content[11], explorations of the use of computational techniques to study online video reuse and remix practices has remained largely understudied.

Why? To begin with, the possibility of scanning the content of individual websites such as YouTube—not to mention the entire internet—requires access to extensive computing power in ways that are inaccessible to most research teams in the humanities. In December 2021, for example, YouTube claimed that more than 500 hours of video were uploaded on the platform *every minute*[12]. This is equivalent to 30,000 hours of new content every hour, or 86 years of new content every day, and it is simply not possible to keep track of such vast amounts of information using a small-scale computational infrastructure. In addition, scanning and scraping content from commercial websites such as YouTube is commonly prohibited by the platforms and scholars who systematically violate such rules enter a controversial legal grey zone[13]. For instance, **YouTube's terms of service agreement** states that its users are not allowed to download *any* content that is found on the platform or access the service using automated means—two rules that prohibit scholars from scraping and downloading content if they would be followed to the point.

As a result, researchers looking to study the reuse and remixing of cultural heritage archives at scale (that is, beyond the manual and small-scale collection of empirical data) may be tempted to use commercial content identification techniques in their search for examples of video reuse. YouTube's and Audible Magic's content identification systems would undoubtedly offer powerful tools for tracing and tracking how archival content circulates online and could provide groundbreaking opportunities for humanistic research *if*—and this is a key point—they were available for academic investigations. To the best of my knowledge, no publicly funded audiovisual archive has currently been granted access to Content ID—or purchased the services of content identification specialist companies such as Audible Magic. There are, however, several examples of private archives making use of commercial content identification infrastructures to monitor how their collections are reused online.

3 Archives on YouTube and the Financial Logic of Tracking Online Content

In 2014, the private archive British Pathé took the world with surprise when it decided to upload its entire collection of more than 85,000 historical newsreels and documentaries on YouTube[14]. Spanning the years from 1896 to 1978, the collection includes footage of world changing events such as the first and second World War, as well as recordings of key historical figures, ranging from Marilyn Monroe and Fidel Castro to Mother Theresa and Salvador Dali. While British Pathé's decision to enter YouTube did not imply that its archive became totally open and free to use (using the videos elsewhere still requires licensing), it did imply that over 3,500 of hours of historic footage was suddenly made available on a platform that attracts over two billion unique **monthly visitors** from around the world. In a short period of time, a unique video collection was exposed and made visible on one of the world's most frequently visited and cross-referenced websites.

As my colleague Eggo Müller—also a member of the CADEAH research team—has previously noted, British Pathé's decision to enter YouTube was not necessarily unique because it made a vast and culturally significant archive easily accessible online (similar digitized and open archives had existed on the internet long before), but because it involved entering into a tight relationship with YouTube[15]. Openly accessible digitized archives are commonly hosted on stand-alone websites, yet British Pathé's YouTube launch meant that it opened its full collection to Google—and thereby also entered deep into the commercial logic of online platforms. This was a conscious move and part of a new business strategy that focused on tapping into YouTube's lucrative streams for advertisement revenues[16]. By placing its archive on YouTube, British Pathé sought to widen its potential audience and attract new licensing customers. In addition, it ensured that it could quickly and efficiently claim new advisement royalties—not just for its personally uploaded YouTube videos, but for every already existing or future video that would enter YouTube and contained copies of their copyright protected collection.

This is where Content ID—a cornerstone in YouTube's financial business model—enters the picture. By scanning original content, extracting its key visual and/or sonic features, and saving such features as compressed "fingerprints" of the original files, Content ID creates a reference database of copyright protected works, against which newly uploaded YouTube videos can be matched and compared. If Content ID identifies a match between a newly uploaded video and one of its reference fingerprints, it notifies the designated rights holder and gives them the choice to either 1) block the video from entering YouTube, 2) claim all future advertisement revenues that the video generates, or 3) allow the video to remain on the website, while tracking its viewing statistics[17]. Importantly, the technology behind Content ID—much like Audible Magic's system and other content identification tools—is capable of identifying not just hard-copied and identical instances of content re-use, but also manipulated and distorted content, such as video remixes or mashups. Already in 2007, this earned the content identification technique a reputation as the new “weapon in the web war over piracy” that would allegedly even out the balance between the creative industries and online copyright infringers[18].

Today, British Pathé is far from the only private archive that has decided to make its video collections available on YouTube. For instance, Cinecittá Luce—the largest film archive in Italy—entered a partnership with Google in 2012, with the goal of digitizing and making its entire collection of more than 100 000 films and 3 million photographs available on YouTube. Originally founded by Benito Mussolini in 1927, the Cinecittá Luce archives contain a rich and unique historic record of fascist propaganda and educational content, but also scenes from historic Olympic Games and everyday events such as Christmas celebrations in Holland in 1929. As of 2015, The Associated Press and British Movietone—two other private and culturally/historically significant newsreel archives—have also made more than 1 million minutes of footage available on YouTube[19], thus adding to the wide range of audiovisual archives whose content can be found on the platform.

There is much to be said about YouTube’s Content ID system but at this point, it will suffice to note that it has been a success story for YouTube/Google, as well as actors like Cinecittá Luce and British Pathé. At its core, Content ID automates the task of policing copyright abuse and provides an opportunity to monitor and control vast amounts of online content. For archives, this means a relief from the time-consuming (and nearly impossible) task of manually searching for copyright violations online, and an opportunity to monetize content in new ways. For YouTube, it provides a way to meet the demands of the creative industries, who have long called for tightened online copyright control. Importantly, Content ID’s ways of encouraging rightsowners like British Pathé to monetize user-generated videos (as opposed to blocking or removing them from the website) is also fundamental to YouTube’s business model, which would collapse if it could not maintain a steady supply of user-uploaded content. In 2018, Google also claimed that rightsholders chose to monetize copyright infringing videos (as opposed to blocking them or doing nothing) in 90 percent of the cases when Content ID had identified an instance of suspected copyright abuse[20]. In many ways, then, British Pathé’s very existence on YouTube could likely be attributed to Content ID, which (if all goes well) ensures that the archive is financially compensated whenever someone re-uses their content.

British Pathé’s presence on YouTube highlights several of the economic benefits that arise when archival institutions enter into partnerships with—and start making use of—the technical infrastructure of commercial platforms. In a time of austerity and shrunken financial support within the archival and cultural heritage sector, it is easy to see the lure and attraction of making digitized collections available and ad-funded on platforms such as YouTube. In 2018, for instance, YouTube claimed to have facilitated the payment of over $3 billion in advertisement revenues to rightsholders who chose to monetize the re-use of their content with the help of Content ID[21]. Aside from capturing a portion of such financial gains there is, perhaps, also a case to be made for the upshots of giving archival collections maximum public exposure on platforms like YouTube, and thereby lowering the barriers for public access to historical content and cultural heritage. In fact, it might be hard to imagine a more appropriate online repository for digitized audiovisual collections, if monthly visitor ratings and global outreach is mainly considered.

As previously mentioned, British Pathé’s presence on YouTube—including its access to Content ID—could also open up new and intriguing possibilities for research. What if archives such as British Pathé would not just use Content ID to safeguard copyrights, but also to map and study the presumably rich and diverse ways in which its collections are re-used online? Content ID’s ways of identifying instances of copyright abuse could just as easily be repurposed to explore how video content is remixed and re-appropriated online. When and where does British Pathé’s footage of Marilyn Monroe, Fidel Castro, Mother Theresa, and Salvador Dali resurface on YouTube? How do contemporary YouTubers make sense of, re-contextualize, and ascribe new meaning to this historical footage? How does Cinecittá Luce’s digitized collections resonate with the present and become re-inserted into new cultural memories? Content ID does not just have to be a one-purpose tool for policing copyrights—it could easily be used to trace and track cultural re-use more broadly. What if other archives and cultural heritage institutions would also make use of Content ID or the content identification technologies provided by companies like Audible Magic—either to actually publish their content on platforms like YouTube, or to simply have their archives fingerprinted and made searchable and identifiable online?

4 Will You Let Us In?

These were the questions that we asked ourselves within the CADEAH project in early 2020, as we made an attempt to contact YouTube to explore the possibility of using Content ID for research purposes. Since YouTube does not provide any contact details on its website (either in the form of email addresses, phone numbers, or chat functions), the process began with filling out a **standardized form** that potential users of Content ID are referred to. The form, which constitutes the entry point to YouTube's so-called "Content Verification Program," asks for a wide range of information, such as the name and contact details of potential clients, along with information about their relationship to the content they are trying to "protect" (are you a copyright owner? A licensed user/distributor? An agent acting on behalf of a copyright owner?). It also requests information about whether or not potential users are part of YouTube's "Partner Program" (a club that gives frequent YouTubers access to creative assistance and ad-revenue support), and if potential users have submitted any copyright takedown requests to YouTube before (with the option of specifying precisely how often this has occurred on a scale of 1-1000+ times in the last year). YouTube also asks where potential users normally host their content (such as on YouTube, or private websites), and what type of content the user is trying to protect (including categories such as advertisements, music, podcasts, audiobooks, education, entertainment, gaming, government, music, news, software, sports etc.). Supposedly, all of this information then feeds into YouTubes decision to either grant or deny someone access to Content ID.

At the moment, however, it is unclear on precisely what grounds YouTube accepts such requests. In April 2023, Content ID was accessible for rights owners who meet four somewhat loose **criteria**. First, potential users must have exclusive rights to the original material that is to be fingerprinted and evaluated. Second, the original content must be of a particular type (and for example cannot consist of mashups, "best of"s, compilations, remixes, recordings of video gameplay, software visuals or trailers, unlicensed music or video, music or video that has been licensed but without exclusivity, and recordings of performances including concerts, events, speeches, and shows). Finally, potential users of Content ID must prove that they have previously submitted "many valid takedown requests" and that they have "the resources" to manage Content ID (exactly what is meant by "many" takedown requests and "having the resources" in this particular context is unclear, however).

Several of these rules present obvious obstacles for scholars and publicly owned archives. For instance, an archive that deals with very old historical sources and/or content belonging to the public domain may not be able to claim exclusive copyrights. Furthermore, cultural heritage institutions may not be in the habit of issuing "many" takedown requests—either because they simply do not want to, or because they do not have the resources to manually look for copyright abuse online. Neither may archival institutions know how to acquire the appropriate "resources" for managing Content ID, which would make it difficult to prove they are qualified for the task.

Unsurprisingly, the fuzziness of YouTube's ways of granting access to Content ID also means that comparatively few organizations and rights holders currently use the tool. In 2018, YouTube reported that roughly 9000 actors (including movie studios, music publishers, record labels, and major network broadcasters) were using Content ID to manage and monetize their works[22]. This number is surprisingly low, given the wide range of rights owners that exist around the globe. One reason why this is the case, is likely that many copyright holders access Content ID with the help of so-called multi-channel networks—that is, third-party businesses that specialize in building commercially successful YouTube channels[23]. This, for example, was the strategy that British Pathé adopted when it released and began to monetize its collection on YouTube[24]. Multi-channel networks are commonly given a privileged and direct access to Content ID and help actors such as British Pathé to issue takedown requests and/or claim advertisement revenues. By "pooling together" multiple rights owners, they likely also shrink the total number of actors who use the tool.

Another important reason why Content ID's user base remains low is likely that YouTube is careful with granting access because of the extensive trust and authority that its users are given. For instance, those with access to Content ID can choose to issue monetization or takedown requests *by default* whenever a suspect case of copyright infringement has been detected. This means that copyright owners can issue hundreds (if not thousands) of monetization or takedown requests automatically and within very short periods of time. As a result of this fully automated way of handling copyright disputes, Content ID is also notoriously known for making fraudulent and obscure content evaluations, which has for example resulted in videos of forest sounds and purring cats being flagged as instances of copyright abuse[25]. In combination with YouTube's policy of acting on monetization and takedown requests *first* (monetizing or blocking content as soon as an issue complaint is registered), and dealing with possible counter-claims or disputes *later* (according to a better-safe-than-sorry logic for the benefit of rights owners and YouTube itself, since the company is working hard to *not* be perceived as a piracy platform), Content ID transforms into a powerful tool that could seriously stifle the freedom of speech if it is abused or misappropriated[26]. Against this background, YouTube's restrictive ways of granting access to Content ID makes sense, although it is notable that it also greatly privileges major corporate actors—or what Dustin Edwards calls "corporate authors"—as opposed to small-scale creative producers[27].

In 2018, YouTube released a lite-version of Content ID called **Copyright Match**, which utilizes the same techniques for identifying cultural content, but is stripped of the opportunity to automate takedown and monetization requests[28]. Unlike Content ID which is geared towards copyright owners in general, Copyright Match especially caters to YouTube's own so-called "content producers" (i.e., users that regularly upload videos on the platform). More specifically, Copyright Match is available for members of YouTube's **Partner Program**, which for example requires having a YouTube channel with more than 1000 subscribers and more than 4000 valid public watch hours in the last 12 months. While YouTube recently claimed that more than 1,5 million of its content creators had access to Copyright Match[29], the service is unavailable to those who are not prepared to make original content available on the platform and maintain well-frequented YouTube channels.

On February 10, 2020 I kept all of this information in mind as I carefully filled out YouTube's form for entering its "Content Verification Program," stating (to the extent that it was possible due to the form's standardized layout) that I was a scholar wanting to study video re-use in collaboration with archival partners. Eight days after my application was submitted, I received what appeared to be an automatic reply from YouTube which declined my request and instead recommended me to use the platform's online webform for manually reporting copyright infringement. That my application explicitly stated that my intention was *not* to issue takedown requests or monetize YouTube videos (but simply study video reuse), was not addressed in YouTube's reply. Neither did I, or anyone else in our research team, succeed in getting a hold of a YouTube representative through other means.

5 Recap

To summarize, we thus find ourselves in the following situation: over the past decades, millions of taxpayer money have been poured into vast digitization projects, the results of which are (partially) openly available on platforms such as Europeana.eu and EUscreen.eu. For instance, EUscreen currently hosts a collection of more than 60,000 audiovisual media items, with plans to add an additional one million videos from content partners over the next few years. It is likely to assume that some of this content is currently re-used and uploaded on major online platforms such as YouTube—either by private video creators, or by commercially driven actors, looking to collect advertisement revenues. Actors such as YouTube, in turn, are likely making profits out of remixed and re-used digitized cultural heritage collections, since they help drive online traffic and feed platforms with their most valuable asset: user-uploaded content.

Meanwhile, publicly funded archives and cultural heritage institutions are unable to get a large-scale overview of how their collections are re-used online. Manually monitoring the vast influx of new content on platforms such as YouTube would be practically unfeasible. Moreover, public archives have difficulties qualifying for access to automatic and commercial content identification services. This may, for example, be the case since they are not in the habit of issuing takedown requests, or because they cannot claim exclusive copyrights if digitized content belongs to the public domain. Furthermore, the possibility for public archives to study video re-use on a large scale from the outside (***without*** accessing services such as Content ID) is hindered by the fact that YouTube and most other commercial platforms forbid scraping, downloading, and saving content from platforms—which would be a prerequisite for conducting academic research.

At the same time, the technical systems that allow for exploring how our shared cultural heritage is broadly re-used are very much there, embodied in technical systems such as Content ID or Facebook's Rights Manager.

Just to be clear, I am not suggesting that cultural heritage institutions should use content identification tools to block or limit the re-use of digitized archive collections. What they could be used for, however, is to help us understand how alternative historical narratives and new ways dealing with the past are developing online. Ultimately, I believe that remix practices and online forms of cultural re-use should be encouraged and viewed as positive expressions of how digitization broadens who has the means to *do history*. This is a good thing that shows the fundamental strength of digitization projects and new ways of sharing historic archives openly online. But the question is if digitizing and uploading cultural heritage collections on the web should be seen as the endpoint of archival efforts, or if scholars and cultural heritage institutions also have a responsibility to explore how those collections are put to new forms of use. In many ways, digitizing cultural heritage collections is the "easy" part, with the more difficult step being to figure out good ways of curating, maintaining, and—possibly—also keeping track of the afterlife of digitized collections.

It is, of course, entirely expected that commercial businesses like Google are careful with who they invite into their technical systems and do not offer their services for free to anyone. Moreover, it makes perfect sense that the ecosystem surrounding services like Content ID is not designed to cater to humanistic researchers, but copyright owners looking to safeguard their cultural assets. To be fair, this is the fundamental reason why Content ID exists in the first place. Yet the current situation also raises several questions: to which extent do platforms like YouTube—which likely profit financially from digitized cultural heritage collections—have a responsibility to give back to the research/archival community and help facilitate cultural and historical research? How should the commercial interests of platforms such as YouTube be weighed against the public interest in learning more about how our shared historical records are put to new forms of use? What does it mean to care for digitized archival collections that are openly available online, and could monitoring possible instances of reuse qualify as a way of maintaining, protecting, and enriching knowledge about our collective cultural heritage?

To publish a privately owned archive on a platform such as YouTube—in similar ways as British Pathé did in 2014—is *one* thing, and there are certainly reasons to question if this is a good way to deal with archival content belonging to the public domain. After all, British Pathé is a private archive and has the right to use its historical footage in whatever way it likes. Just because a public cultural heritage institution may strive for openness and public exposure, however, that does not necessarily mean that it should also upload its content on commercial platforms and thereby commodify our shared cultural heritage in similar ways as British Pathé. But the key thing here is that the use of commercial content identification tools does not have to require that any content is made public on commercial platforms. In theory, it would be perfectly possible for an archive to host its collection on a non-commercial and stand-alone website *and* fingerprint its collection with the help of a commercial content identification tool, without also making the original content public on the commercial website. For instance, major Hollywood studios continuously fingerprint their archives with the help of tools like Content ID, without being forced to also make their content openly available on YouTube.

This illustrates how the use of commercial content identification techniques (such as Content ID) and the publication of content on commercial websites (such as YouTube) are two different things, and that one does not necessitate the other. Moreover, an archive would only have to expose its collection to YouTube (or whatever platform or service it chooses) on one single occasion to have its inventory fingerprinted. From then on, the only information that need to remain with the commercial platform (YouTube or other) are the content fingerprints themselves—fingerprints that are abstracted, compressed, and distorted enough to make it impossible to reproduce the original file. In other words, content fingerprinting is a non-invasive identification technique that does not require archives to share their full and original collections with commercial actors in the long term.

Here, however, we enter another thorny issue: what is the value of a content fingerprint? As previously mentioned, a video fingerprint (such as those provided by Content ID or Audible Magic) cannot be used to re-construct an original file and is more or less useless in itself—aside from the fact that it can help identify content online. So, what is the value of YouTube (or any other content identification provider) being in possession of an object (content fingerprint) and technique (content identification system) that can quickly scan and identify the whereabouts of cultural heritage online?

Let us play with the idea that YouTube would invite a public archive to use Content ID for research purposes and allow it to fingerprint a portion of its archive. In an instant, the archive would have access to a tool that could guide it towards hundreds or possibly thousands of examples of video reuse. Alternatively, the archive may discover that it could not identify one instance of video reuse at all, which in itself would be a fascinating insight that raises questions about the legitimacy of large-scale digitization projects and the means with which cultural heritage institutions facilitate access, curate digital collections, and succeed in making themselves known and relevant online.

In the example described above, however, the archive would not just have gained access to a new tool for studying video re-use. It also would have given YouTube (and in extension, Google), the tools to quickly identify archival content belonging to the commons online. Today, there are—as far as I know—no reasons to believe that YouTube is misusing this power and ability (for example to skew visibility online), but it is easy to imagine that content identification tools could be used for outright repressive purposes, such as censorship and efforts to hinder the freedom of speech. How should this potential risk of misuse be balanced against the possibility of exploring how our cultural heritage is re-used? Does it matter who has the capacity to quickly locate the whereabouts of cultural heritage online? What would an archive be giving away if it allowed an actor like YouTube to fingerprint its collections?

While it might seem like Content ID is a service that simply exists for the benefit of rightsholders and caters to their needs, it is important to remember that being in possession of vast amounts of content fingerprints—like YouTube, Audible Magic, and similar services are—carries a financial value and power in itself. As Guillaume Heuguet suggests, YouTube's use of Content ID can be described as a gradual conquest of the online sound space[30]— and, one might add, the online *audiovisual* space as well. By fingerprinting and indexing significant portions of the world's cultural productions, developers of content identification tools and services are building a capacity to scan, recognize, and regulate how immense amounts of information moves and is displayed online. In 2018, for example, YouTube alone claimed to have fingerprinted over 80 million files, which are currently stored in Content ID's reference database[31]. We should not underestimate the power that comes with having access to these fingerprints, alongside the wider ability to quickly search for the presence of original content online.

6 Final Remarks

A common catchphrase within the digital humanities is to repurpose the "methods of the medium" and make use of the multitude of techniques that are embedded in online devices[32]. There are also numerous examples of when commercial digital tools are repurposed for academic research, including the use of mundane tools like Google Ngram to notice semantic changes in texts over time, or the use of Application Programming Interfaces (or API's) to gather data from social media platforms. If an archive or team of researchers would use a tool such as Content ID, it would therefore certainly not be the first time that a commercially driven technology is re-purposed for academic research. Just because this is a common practice, however, we should automatically assume that it is unproblematic from an ethical standpoint—or refrain from engaging in "tool criticism"[33] and thinking through the outcomes that the use of a particular software solution could have.

In this paper, I have tried to highlight some of the potential advantages—and dangers—with adopting content identification tools to study video reuse. On the one hand, I have discussed how content identification techniques can offer fascinating new opportunities for research. By making use of tools such as Content ID, scholars and archives could map and monitor what happens to our collective cultural heritage online at a fundamentally new scale. On the other hand, I have discussed the difficulties of accessing such tools from a scholarly perspective and shown how existing platforms are largely built and designed for commercial partners. Furthermore, I have shown how the use of content identification tools for academic purposes would raise a series of ethical questions. Here I am, once again, primarily thinking about the long-term effects of allowing actors like Google/YouTube to fingerprint, index, and make cultural heritage identifiable and trackable at scale.

Ultimately, these issues boil down to how scholars and public archives should approach cutting-edge technologies for machine vision and listening that are provided by commercial platforms. Do the possible research outputs that could result from the use of content identification techniques (and other advanced and commercially developed digital tools) outweigh the downsides of entering into partnerships with commercial actors? What would be the best way of ensuring that we have a clear idea of how digitized archival collections circulate and are re-purposed online? For our research team, the way forward became to develop an open source toolkit for studying video reuse, which can be openly accessed on **Github**.[34]

Still, however, the lure of studying cultural re-use with the help of commercial tools such as Content ID remains, as it could deepen insights about audiovisual reuse in ways that our toolkit cannot—simply because the study of YouTube videos at scale is off limits for researchers. For instance, we may find that digitized cultural heritage footage is widely re-used in music videos, embedded into children's shows, or find a place in DIY cooking tutorials on YouTube. But it may equally be the case that digitized cultural heritage collections are heavily re-purposed for discriminatory or anti-democratic purposes (such as propaganda, history revisions, and conspiracy thinking), or that they are heavily commercialized and re-used for profit. Either way, the possibility of using technologies such as Content ID for research purposes would open intriguing new ways of tracing and mapping cultural re-appropriations of heritage archives. Importantly, the use of content identification tools could also help us ask critical questions about the role of platforms such as YouTube itself. To which extent does digitized cultural heritage actually feed into commercial video platforms? How common is it for digitized heritage belonging to the public domain to appear on websites like YouTube? Should YouTube's potential ways of re-publishing and profiting from digitized cultural heritage—without offering any financial compensation to public archival institutions—be seen as a source of concern?

At the moment, the truth of the matter is that we simply do not have a clear overview of how the digitization of European history fosters participatory engagements among the public—or how digitized audiovisual heritage is adding to the financial gains of commercial platforms.

Notes

1. Abby S. Waysdorf, "Vaporwar and Military Contents Fandom." *The Journal of Fandom Studies* 10, no 1 (2022): 19–37
2. Igor Kopytoff, "The Social Life of Things: Commodities in Cultural Perspective." In *The Social Life of Things: Commodities in Cultural Perspective*, ed. Arjun Appadurai (Cambridge: Cambridge University Press, 1986), 64–94
3. Arjun Appadurai, Arjun, ed., *The Social Life of Things: Commodities in Cultural Perspective*. (Cambridge: Cambridge University Press, 2013)
4. To address this problem, another aim of our project has involved developing an open-source toolkit for identifying visual similarities in audiovisual archives, called the Video Reuse Detector. The toolkit is open source and free to use and can be accessed on Github.
5. Google, "How Google Fights Piracy." Report. (2018) **https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Piracy_2018.pdf**
6. danah boyd and Kate Crawford. 2011. "Six Provocations for Big Data." *Paper Presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21* (2011)
7. Tarleton Gillespie, *Wired Shut: Copyright and the Shape of Digital Culture*. (Cambridge & London: MIT Press, 2007)
8. Google, "How Google Fights," 27
9. ibid, 23
10. Audible Magic, "Audible Magic's Content Identification." Core technology data sheet. (2020) **https://www.audiblemagic.com/wp-content/uploads/2019/08/B_F_Core-Technology_012720.pdf**
11. Taylor Arnold and Lauren Tilton, "Distant Viewing Toolkit: A Python Package for the Analysis of Visual Culture." *Journal of Open Source Software* 5, no 45 (2020); Kevin Kee and Timothy Compeau. *Seeing the Past with Computers: Experiments with Augmented Reality and Computer Vision for History*. (Ann Arbor, MI: digitalculturebooks: 2019); Lev Manovich, *Cultural Analytics*. (Cambridge & London: MIT Press, 2020); Wevers, Melvin Wevers and Thomas Smits, "The Visual Digital Turn: Using Neural Networks to Study Historical Images." *Digital Scholarship in the Humanities* 35, no 1 (2019): 194–207
12. YouTube for Press, "YouTube by the numbers." **https://Blog.Youtube/Press/** December 1, 2021.
13. See for example Eriksson et.al. Spotify Teardown. (Cambridge & London: MIT Press, 2019).
14. British Pathé. "British Pathé Releases 85,000 Films on YouTube." Press release (2014).

15. Eggo Müller, "'Great Stuff!': British Pathé's YouTube Channel and Curatorial Strategies for Audiovisual Heritage in a Commercial Ecosystem." *VIEW Journal of European Television History and Culture* 7, no 13 (2018): 19–30.
16. ibid
17. Google, "How Google Fights"
18. Brad Stone and Miguel Helft, "New Weapon in Web War over Piracy." *The New York Times*, February 19, 2007. **http://www.nytimes.com/2007/02/19/technology/19video.html**
19. Associated Press, "AP Makes One Million Minutes of Historical Footage Available on YouTube." Press release. (2015) **https://www.ap.org/press-releases/2015/ap-makes-one-million-minutes-of-historical-footage-available-on-youtube**
20. Google, "How Google Fights," 25
21. ibid, 13
22. Google, "How Google Fights"
23. Ramon Lobato, "The Cultural Logic of Digital Intermediaries: YouTube Multichannel Networks," *Convergence: The International Journal of Research into New Media Technologies* 22, no 4 (2016): 348-360; Patrick Vonderau, "The Video Bubble: Multichannel Networks and the Transformation of YouTube," *Convergence: The International Journal of Research into New Media Technologies* 22, no 4 (2016): 361–75.
24. British Pathé, "British Pathé Releases,"; Müller, "Great Stuff!"
25. EFF, "Takedown Hall of Shame." Electronic Frontier Foundation. Blog post. August 28, 2019. **https://www.eff.org/takedowns**; Felix Reda, "When Filters Fail- These Cases Show We Can't Trust Algorithms to Clean up the Internet." Blogpost. (2017) **https://juliareda.eu/2017/09/when-filters-fail/**
26. Evan Engstrom and Feamster,"The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools." Engine.(2017) **https://www.engine.is/the-limits-of-filtering**
27. Dustin W. Edwards, "Circulation Gatekeepers: Unbundling the Platform Politics of YouTube's Content ID." *Computers and Composition* 47, March (2018): 61–74.
28. Fabio Magagna, "Helping Creators Protect Their Content." *YouTube Official Blog* (blog). July 11, 2018. **https://blog.youtube/news-and-events/helping-creators-protect-their-content/**
29. YouTube, "How Does YouTube Protect Copyrighted Content?" October 21, 2021. **https://www.youtube.com/howyoutubeworks/our-commitments/safeguarding-copyright/**
30. Guillaume Heuguet, "Vers Une Micropolitique Des Formats: Content ID et l'administration Du Sonore." *Revue d'anthropologie Des Connaissances* 13, no 3 (2019): 817–848.
31. Google, "How We Fight," 13
32. Richard Rogers, *Digital Methods*. (Cambridge & London: MIT Press, 2013), 1
33. Karin van Es, Maranke Wieringa, and Mirko Tobias Schäfer. "Tool Criticism: From Digital Methods to Digital Methodology." In *Proceedings of the 2nd International Conference on Web Studies - WS.2 2018* (Paris, France: ACM Press, 2018): 24–27
34. See also Tomas Skotare, Pelle Snickars, and Maria Eriksson, "Tracking and Tracing Audiovisual Reuse: Introducing the Video Reuse Detector." *Journal of Digital History*, forthcoming.

Biography

Maria Eriksson is a visiting scholar in media studies at the department of art, media and philosophy at Basel University. Her research focuses on the history and politics of everyday digital technologies and she is currently involved in the research project **Modern Times 1936**, which explores and critiques how artificial intelligence interprets historic sources.

URL: **https://doi.org/10.18146/view.391**
Publisher: Netherlands Institute for Sound and Vision in collaboration with Utrecht University, University of Luxembourg and Royal Holloway University of London.