Platform Governance and Automated Enforcement: Evidence from YouTube Content ID*

Sverrir Arnórsson[1] Stefan Bechtold[2] Christian Peukert[3] Catherine E. Tucker[4]

[1]University of Zurich, Department of Business Administration
[2]ETH Zurich, Center for Law & Economics
[3]University of Lausanne, Faculty of Business and Economics (HEC)
[4]MIT Sloan School of Management

July 7, 2026

Preliminary draft. Please do not cite or circulate.

Abstract

The abundance of digital content poses challenges for copyright enforcement. Since platforms rely on scale, they also rely on automated copyright enforcement. We study the consequences of the emergence of automated enforcement tools in the context of YouTube, where the Content ID system automatically identifies copyrighted music in user-generated videos and enables rights holders to block or monetize such content. We upload a variety of sound recordings and compositions to YouTube and measure ContentID's response. Our large-scale algorithmic audit suggests that while Content ID's matching technology is generally accurate, observed enforcement systematically diverges from underlying legal entitlements. Some lawful uses are misclassified, while some infringements escape detection. These errors are asymmetric: works represented by major labels are flagged in about 90 percent of cases, compared to roughly 50 percent for non-major labels. Nearly half of uploads that should not receive Content ID flags are nonetheless flagged, often as a result of incomplete or unverified rights metadata. Moreover, the automated enforcement system fails to implement jurisdiction-specific enforcement strategies for protected works in some countries but not in others. Our paper provides the first large-scale empirical evidence on automated copyright enforcement, showing that such systems do not reliably safeguard the interests of smaller or independent creators and are subject to abuse.

*Arnórsson: sverrir.arnorsson@business.uzh.ch, Bechtold: sbechtold@ethz.ch, Peukert: christian.peukert@unil.ch, Tucker: cetucker@mit.edu. We thank Lisa George, Mark Lemley for helpful feedback. The paper has also benefited from comments by seminar and conference audiences at Bocconi University, RWTH Aachen, American Law and Economics Association Meeting, Marketing Science Conference, European Digital Platform Research Network Conference, and Digital Platform Ecosystems Forum. Bechtold and Peukert acknowledge support from the Swiss National Science Foundation for the project 10002634.

1 Introduction

The emergence of digital content venues where users upload content, means that there are many ways copyrighted material can be reused making it challenging for any one individual rights holder to enforce their copyright. For example, every minute, more than 500 hours of video are uploaded to YouTube,[1]. Digital platforms cannot individually police each piece of uploaded content, and they need the scale of this content to succeed, so therefore such platforms have developed automated copyright moderation systems to protect platforms from liability for their users' violations.

This paper presents descriptive evidence of how automated enforcement of copyright on digital platforms affects individual creators. This study focuses on YouTube Content ID, Google's automated copyright enforcement system. Content ID processes more than 2.5 billion copyright claims per year,[2] distributing over $12 billion to rights holders as of 2024.[3] This means our study focuses on one of the largest automated enforcement systems launched by a digital platform. Since Content ID was launched in 2007, similar systems for other platforms such as Meta (Facebook, Instagram) and TikTok have echoed its design.

We evaluate YouTube's automated copyright enforcement system by uploading both sound recordings and musical compositions and measuring Content ID's response. Comparing Content ID decisions with the copyright status of musical works allows us to assess the performance of this automated enforcement system at scale. Content ID issues a *flag* when it determines that an uploaded video contains music *claimed* by a rights holder. We study the divergence between this flag, which is the output of YouTube's fingerprinting technology, and the claim of legal entitlement by the rights holder in a work. Our empirical analysis distinguishes between when Content ID correctly identifies compositions and sound recordings in uploaded videos (*matching accuracy*), whether it correctly identifies whether works are protected by copyright (*enforcement accuracy*), and whether there is potential over- or under-enforcement of copyright for certain types of works or rights holders. Our study thereby analyzes whether deviations between

[1] See YouTube's 2022 Transparency Report: https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-22_2022-1-1_2022-6-30_en_v1.pdf

[2] See YouTube's Transparency Report for 2025: https://transparencyreport.google.com/youtube-copyright/intro.

[3] See YouTube's Transparency Report for 2025: https://transparencyreport.google.com/youtube-copyright/intro.

Content ID's automated enforcement decisions and the copyright status of works originate from the *algorithm*, *data*, or the *institutional design* of the automated enforcement system.

Although we find that the matching technology underlying the enforcement system is largely accurate, the system makes enforcement decisions that diverge from copyright entitlements in various ways. Instances of both over-enforcement and under-enforcement of copyright abound, and these divergences seem to result from particular design choices of the platform's copyright enforcement system. We find that larger rights holders benefit more from YouTube Content ID than smaller rights holders.

Our paper makes three main contributions. First, despite the importance of systems such as Content ID, our understanding of their practical operation remains limited. Previous research has primarily approached automated copyright enforcement from legal and theoretical perspectives, many of them focusing on fair use (Even, 2023; Soha and McDowell, 2016; Bartholomew, 2014; Lester and Pachamanova, 2017; Solomon, 2015), while others have raised awareness about the possible negative outcomes of private copyright enforcement on online platforms (Tang, 2023; Quintais et al., 2022; Grosse Ruse-Khan, 2020; Husovec and Quintais, 2021). The limited observational work has investigated copyright *takedowns* (Erickson and Kretschmer, 2018; Gray and Suzor, 2020), while more than 90% of YouTube's copyright enforcement results in *monetization* rather than removal. Our paper provides some of the first comprehensive empirical evidence on how these systems perform in practice, both in terms of precision and whether their institutional design systematically benefits certain rights holders over others. In particular, by documenting systematic patterns of over- and under-enforcement in one of the largest and oldest content moderation systems on digital platforms, we highlight the limitations of scaling copyright enforcement.

Second, our findings have implications for platform design and public policy. We identify specific mechanisms that drive these enforcement disparities, particularly with regard to access to enforcement tools and economic incentives to generate false claims. These findings echo broader literature on platform governance and digital market competitive structure by documenting how institutional arrangements can embed and amplify existing disparities.

Third, our results shed light on the allocation of revenues for creative works in the ad-supported economy. At least 50% of the approximately 15 billion[4] videos uploaded to YouTube contain music, meaning

[4]See https://tubestats.org.

that billions of videos are subject to Content ID's copyright enforcement.[5] As such, understanding how this revenue is distributed has implications for digital innovation, creative industries, and cultural access.

2 Background and Literature

2.1 Legal Background

When a user uploads a video containing music to YouTube, copyright law distinguishes between at least two works embodied in the soundtrack: the musical composition and its recorded performance.[6] Ownership in these two works often differs: The copyright in the *musical composition* is typically owned either by the original composer, by large music publishers such as Sony Music Publishing, Universal Music Publishing Group, Warner Chappell, or by smaller independent publishers. Music publishers own these rights, as composers frequently assign or exclusively license their copyrights to publishers, which administer licensing for specific rights, such as mechanical, performance, or synchronization rights. In many countries, collecting societies administer certain rights related to compositions, such as public performance rights, on behalf of composers, maintaining licensing arrangements with YouTube. In the United States, collecting societies such as ASCAP administer only public performance rights for compositions. Although the Music Modernization Act of 2018 created a blanket mandatory mechanical license for qualified digital audio services, such as Spotify or Apple Music, the Mechanical Licensing Collective does not administer rights for YouTube. In Europe, collecting societies such as SACEM in France or GEMA in Germany administer public performance and mechanical reproduction rights for compositions. They typically have licensing arrangements with YouTube and participate in Content ID, although their revenue-sharing model may differ from the standard YouTube Content ID arrangement. In Europe, licensing arrangements between collecting societies and YouTube therefore play a more important role than in the US. Therefore, regarding compositions, YouTube may require licenses from composers, music publishers, and/or collecting societies, and the entitled rights holders may differ across jurisdictions.

The rights in *recorded performances* of a musical composition are also dispersed among rights holders and across jurisdictions. In the United States, a sound recording is protected by copyright law, and artists

[5] We estimate this by sampling 1,000 random YouTube videos using the same method that TubeStats uses to estimate the number of videos on YouTube Zhou et al. (2011); McGrady et al. (2023). Then, we use an audio classification model by Gong et al. (2021) to determine whether the video contains music.

[6] Song lyrics can constitute a separate copyrighted work, but we do not analyze them in this paper;

either own the copyright in their recordings in cases of self-release or assign their copyright in recordings to record labels, such as Universal Music Group, Sony Music Entertainment, and Warner Music Group, in exchange for financing, promotion, and distribution. In Europe, recorded music is not directly protected as a copyrighted work, but protected by two "neighboring rights:" the phonogram producer's right and the performer's right. YouTube acquires the necessary licenses from record labels, which own the phonogram producer's neighboring rights and have typically acquired the performers' exclusive neighboring rights via license or assignment.

Copyright protection in *musical compositions* expires 70 years after the death of the composer. Although this copyright term has been harmonized in the US and Europe, different protection terms apply to older works, which we explore in Section 3.3.3. In the US, sound recordings typically enjoy protection for 95 years from the year of their first publication, or for 120 years from the year of their creation, whichever expires first (17 U.S.C. §§101, 114, 302). In the European Union, phonogram producers' and performers' neighboring rights generally run for 70 years from the first lawful publication of the recording (Article 3(1) of the Copyright Term Directive, European Union 2006).[7]

As a digital platform, YouTube needs to respect these rights in compositions and sound recordings. Legislators on both sides of the Atlantic have provided safe harbors that shield online service providers from copyright liability when their users upload content that infringes on third parties' copyrights. However, these safe harbors have limits. Section 512(c) of the U.S. Digital Millennium Copyright Act shields a provider such as YouTube from monetary liability only if the provider has an effective notice-and-takedown system and a repeat-infringer policy (see also Viacom Int'l, Inc. v. YouTube, Inc., 676 F.3d 19 (2d Cir. 2012)). In the European Union, YouTube can escape copyright liability if it makes best efforts to obtain licenses, uses state-of-the-art filtering technologies, responds quickly to copyright notices, and prevents re-uploads (see Article 17 of the Digital Single Market Directive and Article 6 of the Digital Services Act).

In general, determining who owns the rights to music uploaded as part of a video on YouTube is challenging. Because YouTube hosts user-generated content on a massive scale, it must determine the copyright status of music in billions of videos, distinguish between rights to compositions and sound

[7]While we explore differences in terms of protection regarding musical compositions in this study, we do not explore such differences regarding sound recordings, and treat all sound recordings as protected.

recordings, identify various rights holders, and account for regional variations in copyright law and the structure of voluntary licensing arrangements. This reflects the challenge of automating a copyright system that developed in the 19th and early 20th centuries, before the forces of digital technology and the need to scale copyright enforcement digitally were anticipated.

2.2 YouTube Content ID

YouTube developed Content ID in response to mounting legal pressure from major rights holders. In 2007, Viacom sued YouTube for $1 billion, alleging massive copyright infringement, and music labels threatened similar action (Simon, 2014; Helft and Richtel, 2006). To address these challenges while maintaining its user-generated content model, YouTube launched Content ID, an automated copyright enforcement system that aimed at satisfying rights holders' demands for protection while allowing the platform to continue operating at scale.

Content ID operates in two parts: a database of reference files provided by rights holders, and a fingerprinting technology that scans all user uploads for matches against this database. When a match is detected, the system automatically applies a predefined enforcement action set by the rights holder, such as monetizing, blocking, or tracking the video. The system can process audio and video matches. In this paper, we focus only on music within the audio track of uploaded videos. ContentID can distinguish between recording and composition rights and allows for enforcement to vary across different countries. A patent by Google employees (King et al., 2014) outlines a sequential method to distinguish between violations of rights regarding compositions and sound recordings, as illustrated in Figure A.1.

As a first step, Content ID attempts to identify a recording match (Audio ID) using signal processing methods to compare acoustic fingerprints (Weinstein and Moreno, 2007). If no recording match is found, the system then evaluates the upload for compositional similarities (Melody ID) using fingerprints calculated from the melody (Walters et al., 2012). This hierarchy does not imply that an upload can trigger only one claim. A single upload can generate multiple simultaneous claims, including both recording and composition claims, when those claims are attached to different reference entries, which YouTube calls assets. Because the database contains distinct entries from various rights holders, a single upload can interact with multiple rights simultaneously. For example, the same upload might receive one claim tied to a sound-recording asset administered by Sony Music Entertainment and another tied to a composition

asset administered by Warner Chappell. Therefore, while the detection method for a specific asset is hierarchical, observed claims on an upload need not be mutually exclusive. This claim-level distinction is separate from royalty allocation. YouTube's asset model also allows publishers to link composition assets to sound recordings, so compositional rights holders may receive royalties from a sound-recording match even when the observed Content ID claim is an Audio ID claim.[8]

The system depends fundamentally on information provided by rights holders themselves. A rights holder uploads a reference file, an audio or video recording, that YouTube scans against user uploads. All other rights information is then layered on top of this reference file for rights-administration purposes. A rights holder claiming compositional rights, for example, identifies a recording embodying the work and declares ownership of the underlying composition, enabling YouTube to allocate composition royalties when matching uploads, including covers or other recordings of the same work. Rights holders further specify in which territories they hold each type of right and what enforcement action YouTube should take on matching content (monetizing, blocking, or tracking).[9] Crucially, the platform does not require copyright expiration dates, leaving rights holders responsible for updating the system when works enter the public domain. In practice, the primary mechanism for correcting inaccuracies is the dispute process, in which individual users challenge erroneous claims, including those based on expired copyrights.

In 2025, only 0.51% of the roughly 2.5 billion copyright claims issued on YouTube were disputed by users. Of those, 67% were resolved in favor of the uploader, typically because the rights holder either released the claim or failed to respond within the 30-day review window.[10] In a small-scale study of 29 specific copyright claims, Berkowitz (2023) observed that 8 disputes were explicitly approved by rights holders, while the remaining 21 expired after the 30-day deadline. If a rights holder reinstates their claim following an initial dispute, the user may escalate the case by filing an appeal. Once an appeal is filed, the rights holder has 7 days to respond. If the rights holder rejects the appeal, they must file a formal copyright removal request (DMCA takedown) to maintain their claim, which, if successful, will result in the video's removal and a copyright strike against the channel. Throughout the dispute and appeal

[8] See YouTube's API documentation for composition assets: https://developers.google.com/youtube/partner/guides/managing_composition_assets.

[9] Further information can be found on YouTube's overview of rights management: https://developers.google.com/youtube/partner/rights_management

[10] YouTube's Transparency Report for 2025: https://transparencyreport.google.com/youtube-copyright/intro.

process, advertising revenue generated by the video is held in escrow by YouTube and subsequently paid to the prevailing party.[11]

YouTube justifies restricting access to Content ID by citing the potential for abuse of an automated enforcement system that can block, monetize, or track user uploads at scale. The platform describes Content ID as designed for copyright owners with "the most complex copyright management needs," such as record labels and movie studios, and requires applicants to own exclusive rights to a substantial body of original material that is frequently uploaded to YouTube.[12] Access to other copyright management tools is broader but less powerful (see Table 1 for an overview). YouTube's copyright removal webform is available to any copyright owner, while the Copyright Match Tool and Enterprise Copyright Match Tool are designed primarily to identify copies or potential copies and to support takedown workflows. By contrast, Content ID is the tool that allows rights holders to supply reference files, automatically scan uploads, and choose among monetization, blocking, and tracking policies.

This creates a tiered enforcement architecture rather than a system of universal access. Large rights holders, including the three major record labels, are known to participate directly in Content ID. Smaller labels and independent artists may still obtain effective access, but often only indirectly through music distributors, multi-channel networks, publishing administrators, or specialized rights-management firms that administer Content ID claims on their behalf, typically in exchange for a fee or revenue share.[13] The U.S. Copyright Office reports that stakeholders have complained that Content ID's access policy excludes smaller copyright owners and that comparable enforcement technologies may be inaccessible or unaffordable for some small creators and independent labels (U.S. Copyright Office, 2020, 43–44). At the same time, the existence of intermediary access means that non-participation should not be interpreted mechanically as exclusion. A smaller rights holder may lack direct access, may participate through an intermediary, may have only part of its catalog administered, or may rationally decide not to participate because expected revenues are below the fixed costs, administrative burden, or revenue share associated with enforcement.

[11] Help Center article covering this topic: https://support.google.com/youtube/answer/7000961.

[12] See YouTube Help, "How Content ID works," https://support.google.com/youtube/answer/2797370; YouTube Help, "About YouTube's copyright management tools," https://support.google.com/youtube/answer/9245819.

[13] YouTube itself points rights holders to third-party Content ID service providers that "act on behalf of copyright owners for a fee." See YouTube Help, "About YouTube's copyright management tools," https://support.google.com/youtube/answer/9245819, and YouTube Services Directory, https://servicesdirectory.withyoutube.com/.

Table 1: Access channels to YouTube copyright enforcement tools

Access channel	Typical users	Main functionality	Implication for our setting
Copyright removal webform	Any copyright owner	Legal takedown request for specific videos; may help prevent copies of removed videos from being reuploaded.	Broad formal access, but reactive and video-specific; does not provide the same automated monetization infrastructure as Content ID.
Copyright Match Tool and Enterprise Copyright Match Tool	Creators and rights holders with demonstrated need for recurring enforcement; enterprise access is tied to frequent valid removal requests.	Identification of copies or potential copies; review and removal workflows, including bulk removal in the enterprise version.	Provides more scalable takedown support than the webform, but remains distinct from full Content ID access and is less directly suited to monetizing user-generated uses of music.
Direct Content ID access	Large rights holders with complex copyright management needs, such as record labels, movie studios, collecting societies, and large service providers.	Reference-file matching against YouTube uploads; automated claims; right-holder policies to monetize, block, or track matched videos, potentially by territory.	Explains why works represented by major labels may be comprehensively covered by automated enforcement.
Intermediated Content ID access	Independent artists, small labels, publishers, and other rights holders represented by distributors, multi-channel networks, publishing administrators, or rights-management firms.	Third party administers Content ID references, claims, monetization, whitelisting, reporting, and disputes on behalf of rights holders.	Allows some smaller rights holders to participate, but introduces fixed costs, revenue shares, eligibility rules, and possible incomplete catalog coverage.
No effective automated access	Rights holders outside direct or intermediated Content ID arrangements, or those for whom expected benefits do not justify participation costs.	Reliance on manual search, takedown notices, or no enforcement.	Generates potential false negatives in our audit: copyrighted works may be used without a Content ID claim even though the underlying rights remain valid.

Notes: The table distinguishes formal legal access to takedown tools from effective access to automated matching and monetization. YouTube describes Content ID as reserved for rights holders with complex copyright management needs and notes that many copyright owners use third-party Content ID service providers. The categories are therefore not mutually exclusive at the rights-holder level: a small label may use the webform for some works, intermediated Content ID for others, and no automated enforcement for the remainder of its catalog.

Because YouTube does not publicly disclose which entities have been granted Content ID access, participation among smaller rights holders cannot be observed directly, which complicates efforts to assess enforcement coverage and to attribute claims to specific rights holders. What can be observed is that the monetization-capable tier is narrow. YouTube's own transparency report shows that, against billions of potential rights holders, only 7626 entities have Content ID access, and just 4454 of them actively used it in the reporting period.[14] We detail these access routes and the specific schemes and fees that distributors charge in Appendix E.

For those granted access, Content ID is financially attractive. In over 90% of cases, rights holders allow videos to remain online while redirecting advertising revenue to themselves, effectively creating an automated licensing system. As of December 2024, YouTube has distributed more than $12 billion to rights holders through the system.[15]

Despite the significance of these systems, our understanding of their practical operation remains limited. Previous research has primarily approached automated copyright enforcement from legal and theoretical perspectives (Bartholomew, 2014; Tehranian, 2011; Tang, 2023) with two observational studies that only shed light on copyright takedowns, which constitute a minority of copyright action on YouTube (Erickson and Kretschmer, 2018; Gray and Suzor, 2020). As a result, we still lack comprehensive empirical evidence on the accuracy of these systems at scale.

3 Empirical Strategy

This study aims to assess the extent to which Content ID's automated enforcement of claimed rights to compositions and sound recordings embedded in uploaded YouTube videos aligns with the rights that copyright law in the US and Europe actually grants. First, we explore the *matching accuracy* of Content ID: when Content ID flags content as infringing rights covering a third party's work, and ask whether the system correctly identified that work in Section 3.1. Second, we explore *enforcement accuracy* , that is, whether, if Content ID has correctly identified the third party's work, has the system correctly identified whether the work is protected by copyright in Section 3.2. Third, in Section 3.3. we ask whether limited enforcement accuracy leads to systematic over- or under-enforcement of copyright for certain types of

[14] See YouTube's Transparency Report for 2025, "Everyone has access" page: https://transparencyreport.google.com/youtube-copyright/everyone-has-access.

[15] YouTube, 2025 Transparency Report.

work or rights holders, and whether these systematic errors can be explained by access rules and data quality of the automated enforcement system?

3.1 Matching Accuracy

Our algorithmic audit of YouTube Content ID begins by assessing the enforcement system's matching accuracy. We are interested in whether Content ID correctly identifies a musical work when it flags an uploaded video as infringing that work's copyright. By estimating $P(\text{correct ID} \mid \text{flag})$, we assess whether the system correctly identifies the underlying musical work (both sound recording and composition) when it flags an uploaded video. Details on how we quantified this can be found in Appendix C.

Low matching accuracy means that, conditional on a Content ID flag, the system has linked the uploaded audio to the wrong reference work. Such wrong-work matches can contaminate enforcement analysis because a claim may be attached to a work different from the one actually included in the upload. This measure is distinct from enforcement accuracy. It does not capture missed flags, and a claim on non-claimable or public-domain material may reflect correct matching combined with invalid or stale rights metadata rather than a technical matching error.

We estimate Content ID's matching accuracy by uploading a video whose sound includes a musical work that we have accurately identified before. If Content ID flags this video as infringing a particular musical work, we check whether that is the work we have identified before. If the Content ID flag refers to the exact work we included in the video, we label it an accurate match.

3.2 Enforcement Accuracy

We then assess whether Content ID correctly identifies the copyright status of a musical work included in an uploaded video. Our algorithmic audit of enforcement accuracy focuses not on issues of fair use, but on the question of whether a work included in an uploaded video is protected by copyright law or not.

We assess Content ID's enforcement accuracy in three ways. First, we upload works where both recording and composition are under copyright. That is, we upload videos with audio tracks that include music where $V_R = 1$ and $V_C = 1$ and observe whether Content ID flags them. Let $F \in \{0, 1\}$ denote a Content ID flag of the relevant claim type: Audio ID for recording-rights claims and Melody ID for composition-rights claims. Second, we upload works in which neither the recording nor the composition is copyrighted. The audio tracks of our video uploads include music where $V_R = 0$ and $V_C = 0$. Third, we

upload works where the recording copyright status is fixed but the composition copyright status varies. We upload videos where the audio track includes music with public-domain (not copyright-protected) recordings (V_R=0) of copyrighted compositions ($V_C = 1$), as well as non-copyrighted compositions ($V_C = 0$).

With this, we can estimate the probability that a video receives the relevant type of Content ID flag conditional on whether the recording and composition included in the video are protected by copyright ($P(\text{flag} \mid V_R = 1, V_C = 1)$), conditional on whether neither is protected by copyright ($P(F = 1 \mid V_R = 0, V_C = 0)$), and conditional on whether the recording is not copyright protected but the composition is ($P(F = 1 \mid V_R = 0, V_C = 1)$) or is not ($P(F = 1 \mid V_R = 0, V_C = 0)$). Because the suppressed type index differs across rights, a single upload in the $V_R = 0, V_C = 1$ cell can simultaneously contribute to the composition true-positive rate if it receives a Melody ID flag and to the recording false-positive rate if it receives an Audio ID flag.

Consequently, we can assess Content ID's enforcement accuracy by estimating the true positive/negative and false positive/negative rates. For rights-specific rates, we condition on the type of claim asserted: in the $V_R = 0, V_C = 1$ cell, a Melody ID flag contributes to the composition true-positive rate, while an Audio ID flag contributes to the recording false-positive rate. We can estimate the true positive rate for recording rights using $P(F = 1 \mid V_R = 1, V_C = 1)$ and for composition rights using $P(F = 1 \mid V_R = 0, V_C = 1)$. We can estimate the false negative rate for recording rights as $P(F = 0 \mid V_R = 1, V_C = 1)$ and for composition rights using $P(F = 0 \mid V_R = 0, V_C = 1)$. Finally, we can estimate the false positive rate for recording rights as $P(F = 1 \mid V_R = 0, V_C = 1)$ and for composition rights as $P(F = 1 \mid V_R = 0, V_C = 0)$.

However, raw flag rates mix several conceptually distinct margins. A video may fail to receive a Content ID flag because the matching technology cannot detect the audio signal, because the relevant work is not represented by usable reference data in the Content ID database, because the relevant rights holder does not have effective access to Content ID, or because the rights holder chooses not to enforce. These margins matter for interpretation because the database-coverage and access channel is one of the institutional mechanisms that we study below.

To make this distinction explicit, let $R = 1$ denote that the relevant work is represented by usable Content ID reference data, and let $T = 1$ denote that the audio signal is technically detectable by the fingerprinting technology, conditional on such reference data being available. We use the parameter

$$\alpha \equiv P(T = 1 \mid R = 1, V)$$

to denote this technical-detectability component. This parameter is not estimated from our audit, because for unflagged uploads we cannot observe whether the absence of a flag reflects technical nondetection or absence of usable reference data. We therefore use evidence from Weinstein and Moreno (2007), a study by Google engineers of a music identification algorithm similar to the technology underlying Content ID, to define a plausible range for α. Their lower-bound experiments, based on degraded audio and 10-second snippets, imply $\alpha_{\min} \approx 0.90$, while their full-track experiments imply $\alpha_{\max} = 0.997$.

This definition is important because α conditions on usable reference-file availability. If $C = 1$ denotes actual correct identification by Content ID, then $P(C = 1 \mid V)$ would include both technical detectability and database coverage:

$$P(C = 1 \mid V) = P(T = 1 \mid R = 1, V) \cdot P(R = 1 \mid V) = \alpha \cdot P(R = 1 \mid V),$$

abstracting from residual metadata-labeling errors. The coverage term $P(R = 1 \mid V)$ is unobserved and is precisely part of the institutional access and database-quality channel that our paper analyzes. We therefore do not use the adjustment below to point-identify enforcement conditional on full Content ID coverage.

Instead, we use α for a narrower sensitivity adjustment: we ask how large the correct-flag rate would be after netting out purely technical nondetectability, while leaving database coverage, access frictions, and rights-holder participation choices in the measured enforcement gap. Let $F \in \{0, 1\}$ denote whether the upload receives a Content ID flag, and let $C = 1$ denote that an observed flag correctly identifies the work. We observe matching accuracy among flagged videos, $P(C = 1 \mid F = 1, V)$, and the raw flag rate,

$P(F = 1 \mid V)$. The technical-detectability-adjusted effective enforcement rate is then

$$\lambda(V;\alpha) = \frac{\overbrace{P(C = 1 \mid F = 1, V)}^{\text{matching accuracy among observed flags}} \cdot \overbrace{P(F = 1 \mid V)}^{\text{observed flag rate}}}{\underbrace{\alpha}_{\substack{\text{technical detectability} \\ \text{conditional on usable reference data}}}}. \tag{1}$$

Equation 1 should therefore be interpreted as a technical-detectability adjustment, not as an adjustment for database coverage or access. For protected works, $1 - \lambda(V = 1; \alpha)$ is the share of technically detectable uses that do not receive a correct flag, including failures due to missing or unusable reference files, incomplete Content ID participation, intermediary frictions, and rights-holder choices not to enforce. For public-domain works, $\lambda(V = 0; \alpha)$ is the corresponding adjusted over-enforcement rate, reflecting cases where technically detectable material receives a correct identification but the associated rights data or territorial metadata generate a flag that should not exist.

3.3 Structural Sources of Error

Suppose that our measurement of enforcement accuracy reveals mismatches between the claimed and the actual copyright status of a work used in an uploaded video. In that case, we are interested in whether these mismatches are systematic and can be explained by access rules (which rights holders are allowed to use Content ID to enforce their rights?) and/or data quality (on which data does Content ID base its flagging decisions?). Exploring such structural sources of error may help us to understand value distribution in digital markets.

In our empirical analysis, we focus on three distinct structural sources of error: differential enforcement by type of rights holder, fraudulent claims, and territorial variation in copyright protection. We turn to each of these sources in the following.

3.3.1 Different Types of Rights Holders

We are interested in whether Content ID flags content owned by major rights holders differently from content owned by smaller rights holders. Specifically, we estimate $P(\text{flag} \mid V = 1,\ \text{type})$ between different types of rights holders: major labels, which are known to participate in Content ID ($P(inDB) \approx 1$),[16]

[16] We classify labels as major or non-major by evaluating whether a label matched either Universal Music Group, Sony Music Entertainment, or Warner Music Group. Further details can be found in Section B.2.

versus other rights holders, where participation in Content ID is uncertain ($P(inDB) < 1$). If we upload videos that include a copyrighted work published by major record labels known to participate in Content ID ($P(inDB) \approx 1$), $P(\text{flag}V = 1,\ \text{major})$ directly measures copyright enforcement conditional on full coverage. If Content ID flagging differs systematically between major and smaller rights holders $P(\text{flag} \mid V = 1, \text{small}) < P(\text{flag} \mid V = 1,\ \text{major})$, we can derive whether the flagging errors are systematic, originating from restricted participation in Content ID or incomplete reference data in Content ID.

A lower participation rate among smaller rights holders can itself arise from several distinct mechanisms—exclusion from Content ID despite a wish to participate, access costs imposed by intermediaries, lack of awareness of the system, or a deliberate choice not to enforce—which carry different welfare and policy implications. We cannot observe an individual rights holder's intent and therefore do not attribute the entire gap to involuntary exclusion. We discuss this interpretive ambiguity, together with evidence that some rights holders who wish to participate are nonetheless kept out, in Appendix E.

To further assess heterogeneity in flagging, we focus on differences between major and smaller labels with the following specification:

$$\text{Receives a flag}_r = \beta_0 + \beta_1 \cdot \text{Is published by a major label}_r + \epsilon_r \tag{2}$$

where r indexes individual recordings. In an extended model, we incorporate publication-country fixed effects. We construct publication country from the first two characters of the recording's International Standard Recording Code (ISRC), which identify the ISRC registration authority or territory (Int, 2019). These fixed effects should therefore be interpreted as ISRC-registration-territory fixed effects—a proxy for the market in which the recording is registered—rather than Spotify availability-market, Content ID claim-territory, or artist-location fixed effects. We include them to account for systematic differences in enforcement rates across markets, as rights holders may have varying incentives to participate in Content ID based on YouTube's market share in different countries. In addition, these fixed effects absorb registration-territory-specific legal and institutional differences, which are not conceptually relevant to our paper. We also examine enforcement heterogeneity by label size more broadly, measuring size by the number of tracks per label and comparing labels above and below the 99th percentile.

3.3.2 Fraudulent Claims

In addition to affording preferential treatment to larger rights holders over smaller ones, automated copyright enforcement systems may also be vulnerable to fraudulent claims. In particular, we examine whether compositions that are no longer protected by copyright are nonetheless claimed by rights holders in Content ID, and whether Content ID enforces such claims.

To test for this, we rely on the Classical Old sample, which we treat as non-claimable for Content ID purposes: the underlying compositions are in the public domain, and the recordings are released under non-exclusive Creative Commons licenses that YouTube's policies exclude from Content ID claiming. Thus, while the recordings may remain copyrighted in a formal legal sense, any Content ID flag against this sample constitutes over-enforcement under the platform's own rules. By examining the distribution of claimants behind these false positives, we can assess whether erroneous claims are dispersed across many rights holders — suggesting unsystematic metadata errors — or concentrated among a small number of entities, which would be more consistent with deliberate or negligent misrepresentation of rights. This distinction matters for policy: dispersed errors call for better data hygiene, while concentrated patterns suggest that the platform's verification mechanisms are insufficient to deter abuse by specific actors.

3.3.3 Territorial Variation in Copyright Protection

As mentioned in Section 2.1, copyright protection in musical compositions nowadays expires 70 years after the death of the composer. However, this copyright term has expanded significantly at various (and different) points in time in the US and Europe, which significantly complicates an automated determination of a composition's term of protection. Most relevant for our study, the US copyright term for compositions evolved from a system of 28 years of protection after publication (with a 28-year renewal option) under the 1909 Copyright Act, to a system of 50 years after the composer's death under the 1976 Copyright Act, and finally to a system of 70 years after death under the 1998 Sonny Bono Copyright Term Extension Act. In particular, for compositions published between 1923 and 1963, accurately calculating the copyright term requires determining the first publication date (in the United States and abroad), verifying the presence of a proper copyright notice and timely copyright renewals, identifying the composer's nationality, and accounting for copyright restorations for foreign works. Many of these steps can be challenging. Determining the correct first publication date, for example, can be difficult

because archival records may reflect different editions or revisions rather than the true first publication. We provide more information on how we address this challenge in Appendix D.

For our study, these intricacies of US copyright law mean that a composition may be in copyright in one jurisdiction (such as the European Union) while being in the public domain in another (such as the United States): a composition published in the early 20th century by a European composer might have entered the US public domain because of its publication history, and yet have remained under copyright in the EU based on the composer's date of death.

We exploit territorial variations in copyright terms by uploading recordings of compositions that are still protected in the EU but vary in their US copyright status: some remain protected in the US, while others have entered the US public domain. We then examine whether the resulting Melody ID claims respect these territorial boundaries or apply indiscriminately. We estimate:

$$\text{US flag}_{ca} = \beta_0 + \beta_1 \cdot \text{In copyright (US)}_c + \epsilon_{ca}, \tag{3}$$

where c indexes compositions and a indexes assets (distinct entries in the Content ID database, each representing a different rights holder's claim on composition c). The dependent variable equals 1 if the flag applies to the United States. The coefficient β_1 captures the difference in US claiming rates between compositions that are still in copyright in the US and those that are not; in the pooled specification, β_0 gives the baseline rate at which public domain compositions receive US flags.

Because we observe multiple compositions by the same composer with varying US copyright status, we can add composer fixed effects to compare claiming rates within composers. This controls for time-invariant composer characteristics, such as fame, catalog size, or the attentiveness of particular rights holders, that might jointly influence both which compositions are recorded and how Content ID handles them. The identifying assumption is that the selection of which compositions to record is not systematically related to their US copyright status in ways that independently affect claiming behavior. This is plausible because all compositions in our sample remain protected in the EU, so record labels must obtain licenses regardless of US status. The decision to record a particular composition is therefore driven by musical and commercial considerations, not US copyright boundaries.

4 Data

To assess the power and limits of automated copyright enforcement, we designed a large-scale algorithmic audit of YouTube Content ID. As described in Section 3, we isolated deviations between automated enforcement and ground truth in three steps: matching accuracy, enforcement accuracy, and structural sources of error. In Section 4.1, we provide an overview of the setup of our algorithmic audit. In Section 4.2, we describe the samples of works that we used in the audit.

4.1 Setup

To assess the accuracy of YouTube Content ID, we designed a large-scale audit study that systematically compares observed enforcement patterns with known copyright status. Our approach involves uploading videos that include music works with known legal status to YouTube and tracking which uploads receive copyright flags from Content ID. By comparing flags received against the copyright status of the included composition and sound recording, we can identify both over-enforcement (false positives) and under-enforcement (false negatives, see Table **??**).

For our audit, we prepared 15,740 blank-screen videos with audio recordings of various compositions. We uploaded all videos to YouTube as private videos using the platform's official API, for which we received an extended quota through the YouTube Researcher Program. We distributed uploads over several weeks to minimize server load.[17] Since we always kept our uploads private, no third party could watch them. However, we could still verify that Content ID had scanned the uploads for violations and monitor whether it issued any flags on our uploads. We then collected detailed information about each flag, including the claimant's identity, the specific rights asserted (composition vs. sound recording), the geographical scope of enforcement, and whether the rights holder wanted to monetize, block, or track the upload. Appendix B provides more information on our uploading and monitoring infrastructure.

4.2 Samples

To assess matching accuracy, enforcement accuracy, and potential structural sources of errors in Content ID (see Section 3), we used three different samples of works in our uploaded videos. We carefully designed

[17] When analyzing music by copyright status, we use each work's status at the time the corresponding video was uploaded.

these samples to benefit from information about rights holders of works, about whether works were still protected by copyright, and to exploit regional variation in copyright terms.

We uploaded 4,000 videos with sound recordings from all genres that are copyright-protected ("Contemporary" sample). As we knew the ground truth for these works, that is who their rights holders are, we wanted to see whether Content ID's fingerprinting technology correctly identifies the rights holders. We complemented this sample with 790 uploaded videos containing sound recordings of public-domain classical compositions in Creative Commons-licensed recordings that are non-claimable for Content ID purposes ("Classical Old" sample). If Content ID flags these videos, which it should not, we can investigate whether Content ID correctly identifies the composition and its composer. Regarding enforcement accuracy, we use these two samples (Contemporary and Classical Old) to assess whether Content ID can correctly identify whether a work is protected or otherwise claimable through Content ID. Regarding structural sources of error, we use the *Contemporary* sample to analyze whether Content ID enforces works owned by different types of rights holders differently, the *Classical Old* sample to examine whether fraudulent claims occurred on Content ID, and a sample of 630 early-20th-century classical compositions that are still copyright-protected in the EU and vary in US copyright status ("Classical US-EU" sample) to explore territorial variation in copyright protection using composition × asset observations (on these three structural sources of error, see Section 3.3).

Table 2 provides an overview of these three samples.

Contemporary Sample. This sample focuses on works where we have reliable information on ground truth. We need works which are still protected by copyright and for which we know the rights holders. In the *Contemporary* sample, we used "Every Noise at Once" (ENAO),[18] a website that provides a broad genre-based overview of tracks on Spotify, organized in 6,201 genre playlists. We obtained 30-second audio snippets of these tracks from Spotify. To ensure coverage across the popularity distribution, we stratified our sample using Spotify's popularity metric (see Figure B.1) and preserved the share of major labels observed in the ENAO frame. When a sampled track did not have a usable preview, we replaced it with a track from the same popularity bucket and label type. To determine whether the sound recordings

[18] https://everynoise.com.

Table 2: Summary of Musical Samples Used in the Study

Name	N	Period	Genre	Focus	Status	Audit Role	Sources
Contemporary	4,000	1938–2023	All	Sound recording	Protected	Matching Accuracy; Enforcement Accuracy; Structural Sources of Error	Spotify; ENAO
Classical Old	790	1521–1921	Classical	Composition & sound recording	Non-claimable	Matching Accuracy; Enforcement Accuracy; Structural Sources of Error	Gardner; Krueger; Ishizaka
Classical US–EU	630	1895–1967	Classical	Composition	EU: protected US: varying	Structural Sources of Error	YouTube uploads

Note: Column "N" shows the number of observations per sample. The unit of observation is the individual video for the *Contemporary* and *Classical Old* samples, and the composition for the *Classical US–EU* sample. Column "Period" refers to the release date for sound recordings, and year of creation for compositions. Column "Focus" describes whether our study analyzed Content ID flags on compositions and/or sound recordings in the respective samples. Column "Status" shows whether the works in focus are copyright-protected or claimable for Content ID purposes. Column "Audit Role" shows the steps for which we use the three samples. Column "Sources" describes the sources of works used in the three samples, which are described in more detail in the following text and in Appendix B.

contained in the snippets are still copyright-protected, we relied on the metadata provided by Spotify, which lists the rights holder for each sound recording. We then uploaded blank videos with these Spotify snippets to YouTube. Among those videos that Content ID flagged, we measured the percentage in which Content ID correctly identified the underlying work, using the Spotify track title as our ground truth (see Appendix C for the matching procedure). Because Content ID claims are often administered by distributors or intermediaries, the named claimant need not coincide with the Spotify rights holder even when the correct recording is identified; we therefore assess matching accuracy at the level of the work rather than the claimant. We created 4,000 videos in this sample, which were uploaded to YouTube in December 2023. Further details on this sample can be found in Appendix B.

We use this sample to evaluate Content ID along several dimensions. First, we assess enforcement accuracy by observing whether the system detects the recordings at all, since every recording in this sample are copyright-protected, and any unflagged upload represents a false negative (see Table **??**). Conditional on detection, we examine matching accuracy by analyzing whether Content ID correctly identifies the sound recording, using the Spotify track title as ground truth for the underlying work.

Finally, we investigate structural sources of error by examining how enforcement and matching accuracy vary with recording label characteristics.

Classical Old Sample. This sample examines whether Content ID's automated enforcement leads to over-enforcement, including false positives as depicted in Table **??**. This requires works that should not receive Content ID flags. The *Classical Old* sample is non-claimable for Content ID purposes because it combines public-domain compositions with Creative Commons-licensed recordings that YouTube's policies exclude from Content ID claiming.[19] Any flags on videos that include recordings of such compositions therefore represent false positives. The rate at which these works receive Content ID flags allows us to quantify how often Content ID restricts content that should remain freely available. Many of the works in the sample stem from the Isabelle Stewart Gardner Museum, which has released recordings of classical music compositions that are in the public domain and whose sound recordings are released under a Creative Commons license. In total, we created 790 videos with such sound recordings for this sample, which were uploaded between October and November 2023. Section B.6 in the Appendix provides further information on this sample.

Using this sample, we primarily evaluate the enforcement accuracy by observing whether these videos receive flags at all, since none of them should. For the flags we do receive, we also assess matching accuracy by checking whether the flag at least identifies the correct underlying composition and composer.

Classical US-EU Sample. As described in Section 3.3.3, there are cases where early 20th-century compositions are still protected by copyright law in the EU, while they ran out of copyright in the US. This provides an interesting test for an algorithmic audit of structural sources of errors, as Content ID issues region-specific flags. If a video with such a composition is uploaded to YouTube, Content ID should flag it for the EU but not for the US. We exploit this fact by collecting a sample of recordings of early-20th-century compositions that feature this regional variation in copyright protection.

Collecting such a sample faces a legal challenge. Determining the copyright status of early-20th-century compositions is far from trivial. In the EU, copyright lasts 70 years after the composer's death, regardless of publication history. We therefore focus on recordings of classical compositions by early-

[19] See YouTube's support page outlining what qualifies for Content ID: https://support.google.com/youtube/answer/1311402.

20th-century composers who died less than 70 years ago. This ensures that these compositions are still copyright-protected in the EU. Among these compositions, we need variation in US copyright status, with some works still protected in the US and others in the US public domain. Under US copyright law, the exact term of protection depends on factors such as publication dates, renewal filings, and foreign publication history, meaning that works by the same composer can have different copyright statuses (see also Section 3.3.3). This within-composer variation allows us to test whether Content ID distinguishes between protected and unprotected compositions, holding the composer constant.

Determining the true copyright status of an early-20th-century composition under US copyright law is typically a task reserved for highly paid copyright lawyers. Determining the true copyright status of hundreds of early-20th-century compositions is, however, infeasible for our study, as it would likewise be infeasible for YouTube to determine the true copyright status of all early-20th-century compositions contained in its videos. To approximate the true copyright status of these compositions and to adopt an approach that YouTube could implement at scale, we use three methods to determine the copyright status of early-20th-century compositions under U.S. law: (i) publication dates from the International Music Score Library Project (IMSLP),[20] a large public archive of scores with metadata on publication history; (ii) IMSLP's jurisdiction-specific copyright notes, which incorporate factors such as renewal filings and foreign publications; and (iii) GPT-5 with web search enabled, which we used to retrieve factual information and apply the rules governing US copyright term. Classification results were broadly consistent across methods. More details on the copyright term verification process can be found in Appendix D.

Following this methodology, we identified early-20th-century compositions that are still under copyright in the EU, some of which are also under copyright in the US, while others have entered the US public domain. For instance, all of Shostakovich's works remain under EU copyright until 2045, yet his early compositions – such as his Symphony No. 1 (published 1927) – have entered the US public domain, while later works – such as his Symphony No. 5 (published 1939) – remain protected in both jurisdictions. We then identified YouTube videos that included recordings of such compositions. Overall, we successfully identified recordings of 630 compositions written by 22 composers, yielding 1,696 composition × asset observations for the territorial analysis.As explained in Section B.7.6, a composition can be claimed by

[20] https://imslp.org.

the same rights holder several times for different jurisdictions, and a composition can also contain several movements, leading to several composition × asset observations. All videos containing such recordings were uploaded between February and April 2025. Further details on the selection of recordings and the upload process are provided in Appendix B.7.

We use this sample to examine the third of our structural sources of error: territorial variation in copyright protection (see Section 3.3.3). We evaluate whether Content ID enforcement respects territorial copyright boundaries by flagging videos in the *Classical US-EU* sample in the EU, but not in the US, holding the composer constant.

5 Results from the Algorithmic Audit

In reporting our results, we begin by assessing Content ID's matching and enforcement accuracy (see Section 3): when Content ID flags content as infringing rights covering a third party's work, has the system correctly identified that work? If yes, has the system correctly identified whether that work is protected by copyright? We explore both questions in Section 5.1, and turn to structural sources of error in Section 5.2.

5.1 Matching and Enforcement Accuracy

To assess Content ID's matching and enforcement accuracy performance, we rely on the *Contemporary* and *Classical Old* samples, as described in Section 4.2. Table 3 reports the share of uploaded videos that were flagged, by copyright or Content ID claimability status, and Table 4 reports matching accuracy, that is, whether Content ID identified the correct work.

Matching Accuracy. Among the 4,000 videos of our *Contemporary* sample, 2,186 were flagged by Content ID, receiving an Audio ID flag, ideally identifying the sound recording and the composition of the video's soundtrack. Because the *Contemporary* sample consists of copyrighted commercial recordings, a successful match will always occur at the Audio ID stage, and Content ID only proceeds to Melody ID when no recording match is found (see Section 2.2). All flags in this sample are therefore Audio ID flags. Among the flagged videos, Content ID correctly identified the sound recording —– and thereby the underlying composition — in 97.9% of the cases. Direct verification of the exact reference recording was not possible for any flag because YouTube reports claim metadata but does not provide the reference

Table 3: Content ID Enforcement Accuracy

Status	Claimed	Not claimed	Total
Protected *(Contemporary)*	54.65% (2,186)	45.35% (1,814)	4,000
Non-claimable *(Classical Old)*	45.44% (359)	54.56% (431)	790
Total videos	2,545	2,245	4,790

Note: A video is counted as claimed if it received at least one Content ID claim. Columns report estimated enforcement accuracy without controlling for observable differences across works and rights holders. The *Contemporary* sample consists of copyright-protected recordings, which a perfectly accurate system would always claim, whereas the *Classical Old* sample is non-claimable for Content ID purposes, as it combines public-domain compositions with non-exclusively licensed recordings. Whether the claimed work was correctly identified (matching accuracy) is reported separately in Table 4.

Table 4: Content ID Matching Accuracy

Sample	Audio ID claims	Melody ID claims	Correct work	Incorrect work
Contemporary (protected)	2,186	0	2,141 (97.9%)	45 (2.1%)
Classical Old (non-claimable)	295	122	386 (92.6%)	31 (7.4%)

Note: This table reports how accurately Content ID identified the claimed work, conditional on a claim being issued; the unit of observation is a claim. In the *Contemporary* sample, we received only Audio ID claims (recording rights), as Content ID processes an Audio ID claim before a Melody ID; a correct Audio ID claim on a sound recording also refers to the correct underlying composition. For the *Classical Old* sample, we differentiate between flags for sound recordings (Audio ID) and compositions (Melody ID), as Content ID might issue an Audio ID flag for a video and refer to a different recording of the same composition. We sometimes received multiple Audio ID or Melody ID claims per video, because recordings of classical works are sometimes split into several tracks (e.g., movements of a symphony), and because the same composition or recording can be claimed by more than one rights holder, each asserting distinct rights over the work. While 93% of the claims in the *Classical Old* sample referred to the correct underlying composition, all of them matched an incorrect recording.

audio file on which the claim is based. The first-stage title-and-artist matches therefore provide metadata-based evidence of recording-level correspondence, while residual cases may include both true recording matches with inconsistent metadata and different recordings of the same song. Because the GPT-assisted residual validation used a work-level prompt, it may count the latter as matches. [21] This indicates that the matching algorithm itself performs well: when Content ID has a reference file in its database and flags the soundtrack in an uploaded video, it nearly always correctly identifies the works. The observed high technical accuracy is consistent with Google's own research (Weinstein and Moreno, 2007). A general limitation is that we observe only the metadata labels attached to Content ID claims, not the underlying reference files in the Content ID database. A claim may correctly identify the soundtrack in a video, yet attribute it to the wrong work if the reference file itself is mislabeled. We therefore cannot always distinguish a matching error from a labeling error. Therefore, the actual matching accuracy of Content ID's fingerprinting technology might be higher than what we can observe.

We repeated this claim-level analysis for our *Classical Old* sample, which consists of Creative Commons-licensed recordings of public-domain compositions. Among the 417 flags issued against videos in this sample, the system correctly identified the underlying composition in 386 flags (92.6%). However, it consistently attributed the soundtrack to the wrong sound recording. For example, a 1960s performance of a Johann Sebastian Bach suite might be identified as a 2010 recording by a different artist. The system, therefore, appears to match musical content reliably but struggles to correctly attribute new recordings of classical compositions when a corresponding reference file already exists in its database, tending instead to classify the new recording as matching an existing recording.

Enforcement Accuracy. While our audit with the *Contemporary* sample reveals a very high matching accuracy, Content ID's enforcement accuracy in this sample is much lower. As Table 3 shows, among the 4,000 videos we uploaded in this sample, 1,814 (45.4%) were not flagged by Content ID, even though they contained copyright-protected content. The adjustment in Equation 1 should be read as a technical-detectability adjustment rather than a full correction for database coverage. Combining the observed flag rate with matching accuracy among flagged videos, and then applying the technical-detectability range from Weinstein and Moreno (2007), yields an adjusted effective enforcement rate between 53.7%

[21] Appendix C discusses this limitation.

and 59.5%. The corresponding adjusted under-enforcement rate is therefore between 40.5% and 46.3%. This rate should not be interpreted as a purely technical false-negative rate. It includes the institutional sources of under-enforcement that are central to our analysis, such as missing or unusable reference files, incomplete participation in Content ID, intermediary frictions, and rights-holder choices not to enforce.

In contrast, among the 790 videos uploaded as part of the *Classical Old* sample, 359 videos (45.4%) were flagged by Content ID, even though the sample is non-claimable for Content ID purposes. Here the flag rate is itself the error. Applying the same technical-detectability adjustment yields an adjusted over-enforcement rate between 42.2% and 46.8%. This rate captures over-enforcement net of purely technical nondetectability. It is therefore best interpreted as evidence that, when the system is technically capable of matching the musical content, the reference database or associated rights metadata can still generate claims on material that should remain freely usable.

Overall, given the high matching accuracy documented above, the lower enforcement accuracy may not stem from any technical inability to identify content, but from gaps in Content ID's database or imprecise metadata, leading to over- and under-enforcement. We explore these factors in the following section.[22]

5.2 Structural Sources of Errors

As described in Section 3.3, we focus on three distinct structural sources of error: differential enforcement by type of rights holder, fraudulent claims, and territorial variation in copyright protection. We turn to each of these sources below.

5.2.1 Different Types of Rights Holders

We are interested in whether Content ID flags content owned by major rights holders differently from content owned by smaller rights holders. To examine whether under-enforcement affects all rights holders equally, we focus on rights in sound recordings and distinguish between the major record labels – Universal Music Group, Sony Music Entertainment, Warner Music Group, and their subsidiaries – and independent labels. Major labels are known to participate directly in Content ID, while independent labels' participation is less common and often unclear. We regress an indicator of whether a video con-

[22] Our data also enables us to observe how rights holders chose to enforce their claims once Content ID has determined that an uploaded video infringes their rights. Almost all rights holders choose to monetize their claim rather than block the video or merely track usage. For more information, see Appendix A.

Table 5: Enforcement by Type of Rights Holder and Popularity

	(1)	(2)	(3)
Published by a major label	0.371***		0.323***
	(0.027)		(0.033)
Spotify popularity		0.745***	0.673***
		(0.046)	(0.050)
Major × Popularity			−0.418***
			(0.141)
(Constant)	0.573***	0.604***	0.576***
	(0.025)	(0.024)	(0.023)
Num.Obs.	3986	3986	3986
R2 Adj.	0.076	0.094	0.117
R2 Within	0.045	0.063	0.088
R2 Within Adj.	0.044	0.063	0.087
FE: Publication country	X	X	X

Note: Linear probability model with publication-country fixed effects in every column; the constant shown is the mean of the estimated fixed effects. Dependent variable: 1 if a recording from the Contemporary sample was flagged by Content ID. Independent labels are the omitted base category. Spotify popularity is mean-centered and expressed as a 0–1 fraction of its native 0–100 scale, so the popularity coefficient is the change in flag probability across the full popularity range, the major-label coefficient is the enforcement gap at average popularity, and the constant is the independent enforcement rate for a track of average popularity. Standard errors clustered by country.

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

taining a sound recording from the *Contemporary* sample receives a claim on rights holder characteristics in a linear probability model, and examine how this gap varies with the recording's popularity.

Column (1) of Table 5 reports the baseline rights-holder gap. Independent-label recordings receive a Content ID flag in 57.3 percent of cases on average (the constant, computed as the mean of the publication-country fixed effects), whereas major-label recordings are 37.2 percentage points more likely to be flagged—an enforcement rate of roughly 94 percent. Column (2) shows that enforcement also rises strongly with a recording's popularity: moving from the least to the most popular recording, across the full 0–100 Spotify popularity scale, raises the probability of a flag by 74 percentage points.

Column (3) interacts the two. Because Spotify popularity enters as a mean-centered fraction of the full scale, the major-label coefficient is the enforcement gap at average popularity: 32.3 percentage points, implying that major-label recordings of typical popularity are flagged about 90 percent of the time (57.6% + 32.3%). The popularity slope now describes independent labels: moving across the full popularity range raises their probability of a flag by 67 percentage points, while the negative interaction (−42 percentage points) shows that the slope for majors is far flatter—a rise of only about 25 percentage points across the same range, roughly a third as steep. In other words, major-label content is enforced at a consistently high rate regardless of its popularity, whereas independent-label content is protected primarily when it is popular enough to be detected and flagged. The rights-holder gap is therefore largest for obscure recordings and narrows as popularity rises, consistent with major labels operating comprehensive Content ID coverage while independents' protection depends on a recording attracting sufficient attention. These patterns are robust to replacing the major-label indicator with a broader measure of label size—whether a label ranks in the top 1% by number of tracks published: large labels are likewise substantially more likely to receive a claim, though the gap is smaller in magnitude (Appendix A, Table A.2).

Another way to study differential enforcement by type of rights holder is to examine the distribution of rights holders associated with Content ID claims. Table 6 shows that the major record labels are the rights holders that are behind most claims against works in the *Contemporary* sample. Tracks that we identified as published by the three major labels – Universal Music Group, Sony Music Entertainment, and Warner Music Group – together account for 14.6% of claims, although they only represent around 8% of the tracks in the sample. This overrepresentation is consistent with the demand-side structure of the recorded music market. Streaming consumption is highly concentrated in a small head of popular tracks: in the UK streaming market, the top 1% of tracks account for between 75% and 80% of streams, while the top 10% account for 95% to 97% (Hesmondhalgh et al., 2021). Major-label recordings are also disproportionately represented in this popular head. The top 0.1% of tracks contains around nine times as many major-owned as non-major-owned tracks, and even among the top 10% of tracks, major-owned tracks outnumber non-major-owned tracks by roughly three to one (Hesmondhalgh et al., 2021). More recent industry evidence points in the same direction: the vast majority of newly delivered tracks come from independent and DIY distribution, while a very small fraction of available tracks accounts for a large

Table 6: Distribution of Claims by Rights Holder in the Contemporary Sample

No.	Content owner name	Number of claims	Percentage
1	**Sony Music Entertainment**	**480**	**22.0%**
2	Believe Music	225	10.3%
3	**Universal Music Group**	**196**	**9.0%**
4	**Warner Music Group**	**123**	**5.6%**
5	CD Baby CO	118	5.4%
6	TuneCore	87	4.0%
7	ONErpm	68	3.1%
	...		
297	wataryproduction	1	0.0%

Note: This table lists the rights holders behind claims against recordings of the *Contemporary* sample when uploaded to YouTube. The unit is a Content ID claim: "Number of claims" is the number of the 2186 distinct claims in the sample that name the given rights holder, and "Percentage" expresses this as a share of those 2186 claims. Because a single claim can name more than one rights holder, the percentages do not sum to 100%. The Orchard Music, a subsidiary of Sony Music Entertainment, is collapsed into Sony Music Entertainment. Rights holders are grouped by ownership at the time of data collection; CD Baby has since been acquired by Universal Music Group (via the Downtown Music Holdings acquisition completed February 2026).

share of actual streaming consumption (Luminate, 2026). Thus, the fact that major-label tracks generate a higher share of claims than their share of tracks suggests that claims are partly demand-weighted: major-label repertoires occupy a relatively small part of the supply of available music, but a much larger part of the music that is actually consumed and reused. At the same time, this interpretation should not be read as implying that demand alone explains the overrepresentation of major-label tracks among claims. In the Content ID setting, claims also depend on rightsholders' access to the system, whether their reference files are included in the matching database, the quality and coverage of fingerprints, enforcement choices by rightsholders, and the propensity of uploaders to reuse particular repertoires. Major-label tracks may therefore be overrepresented among claims both because they are more likely to be consumed and reused, and because the institutional infrastructure for detecting and claiming these uses is more complete for major-label repertoires. However, among the rights holders in Table 6 they are named in 36.0% of claims. As discussed in Section 2.1, this may be caused by smaller labels entering into distribution and licensing agreements with the major labels, likely driven by the latter's sophisticated infrastructure, which allows them to easily enforce music rights via Content ID.

These results suggest that Content ID's under-enforcement stems primarily from incomplete effective coverage rather than technical detection failures. That coverage gap may arise because independent labels lack direct access to Content ID, because they access the system only through intermediaries that charge fees or revenue shares, because intermediaries do not administer all works equally, or because some rightsholders rationally choose not to participate when expected revenues are low. The key point is that automated enforcement is not applied symmetrically across the catalog: major-label recordings appear to be comprehensively covered, while independent-label recordings are covered only selectively.

5.2.2 Fraudulent Claims

We now examine whether we can find fraudulent claims in our samples, which could contribute to over-enforcement. In particular, we are interested in whether compositions that are no longer protected by copyright are nonetheless claimed by rights holders and, consequently, flagged by Content ID. We focus on the *Classical Old* sample, which includes recordings of compositions that are in the public domain.

Among the claims raised in this sample, Table 7 shows that a single entity accounts for a disproportionate share of claims: "*LatinAutorPerf*" is named in 81 of the 417 claims (19.4%), more than twice

Table 7: Distribution of Claims by Rights Holder in the Classical Old Sample

No.	Content owner name	Type	Number of claims	Percentage
1	LatinAutorPerf	Publisher	81	19.4%
2	UMPG Publishing	Publisher	40	9.6%
3	Kontor New Media Music	Label	39	9.4%
4	LatinAutor	Publisher	30	7.2%
5	Wise Music Group	Publisher	30	7.2%
6	Hexacorp (music publishing)	Publisher	28	6.7%
7	The state51 Conspiracy	Label	27	6.5%
	...			
9	**Universal Music Group**	**Label**	**21**	**5.0%**
	...			
27	**Sony Music Entertainment**	**Label**	**6**	**1.4%**
	...			
59	**Warner Music Group**	**Label**	**2**	**0.5%**
	...			
77	[Merlin] IDLA Distribution	Label	1	0.2%

Note: This table lists the rights holders that were behind claims against recordings of the *Classical Old* sample when uploaded to YouTube. The unit is a Content ID claim: "Number of claims" is the number of the 417 distinct claims in the sample that name the given rights holder, and "Percentage" expresses this as a share of those 417 claims. Because a single claim can name more than one rights holder, a claim can be counted toward several rows, so the percentages do not sum to 100%. The most frequent claimants are predominantly music publishers asserting composition rights, while major record labels (Universal Music Group, Sony Music Entertainment, Warner Music Group), asserting recording rights, account for a comparatively marginal share of claims.

as many as the next most frequent claimant. This entity systematically asserts rights over compositions it does not own or represent. It also appears to be affiliated with "*LatinAutor*" and "*UBEM (Uniao Brasileira de Editoras de Musica)*", which do not extend its reach but claim alongside it: every claim naming LatinAutor or UBEM is one that already names LatinAutorPerf. Multiple online complaints from creators document LatinAutorPerf's baseless claims across diverse content, including public domain classical music, original compositions, and Creative Commons-licensed works.[23] In our sample, the claims made by LatinAutorPerf and affiliated claimants are based on incorrect information about composers. For example, a Joseph Haydn composition – composed over 200 years ago – is labeled as composed by Bela Bartók, who died in 1945. The entities behind LatinAutorPerf and related accounts appear to have access to Content ID and can enter what YouTube calls "composition shares," pretending to represent the rights of (typically deceased) composers. This pattern is consistent with copyfraud: the false assertion of copyright over public domain or otherwise uncontrolled works (Mazziotti, 2010). Copyfraud occurs at a meaningful scale despite YouTube's restricted access policy suggesting that verification mechanisms are insufficient to deter abuse, and that simply broadening access to address under-enforcement could exacerbate fraudulent claims.[24]

5.2.3 Territorial Variation in Copyright Protection

As explained in Section 3.3.3, copyright terms for individual works may differ between countries, as the details of copyright terms have not been harmonized worldwide. In particular, there are cases where an early-20th-century composition is out of copyright in the United States, but still protected in the European Union. A perfectly functioning automated enforcement system should recognize these differences and, if a video containing such a composition is uploaded, issue a flag controlling its distribution in the European Union while allowing its distribution in the United States.

[23] See Reddit discussions at https://www.reddit.com/r/audioengineering/comments/14sije4/false_youtube_content_id_claims_by_latinautorperf/, https://www.reddit.com/r/COPYRIGHT/comments/1b5p3kt/bogus_copyright_claims_by_latinautorperf_uniao/, https://www.reddit.com/r/youtube/comments/bjfv3z/someone_or_a_company_named_latinautor_copyrighted/, https://www.reddit.com/r/PartneredYoutube/comments/1bidysa/false_copyright_claims/.

[24] It is noteworthy that legitimate rights holders, such as the three major record labels Universal Music Group, Sony Music Entertainment, and Warner Music Group, also appear in the list of claimants in Table 7, although the uploaded videos in our *Classical Old* sample should not receive Content ID claims. It appears that this is mainly due to mismatched recordings, that is, Content ID recognizing the correct composition, but associating the recording with another recording that is represented by a record label (on this problem, see Section 5.1).

Because YouTube reports copyright flags on a per-country basis, we can test whether Content ID's enforcement patterns accurately respect US copyright boundaries. Using the *Classical US-EU* sample, we examine whether Content ID distinguishes between copyrighted and public domain compositions within the same composer's catalog. The sample contains compositions with varying US copyright statuses while remaining protected under EU law. As outlined in Section 4.2 and described in more detail in Appendix D, we rely on three methods (publication dates, jurisdiction-specific copyright notes, and GPT-5) to automatically determine the term of protection for compositions under US law – an approach that closely mirrors how YouTube itself could assess copyright terms at scale.

Table 8: Claim Results: US–EU Sample

Copyright status	Claimed in US	Not claimed in US	Total
Protected in US *(Classical US-EU sample)*	654 (62.0%)	401 (38.0%)	1,055
Not protected in US *(Classical US-EU sample)*	315 (49.1%)	326 (50.9%)	641
Total claims	969	727	1,696

Note: Each observation is a composition × asset pair. The sample includes 630 compositions and 1,696 composition × asset observations. A single composition may be matched by multiple assets (e.g., when a recording is split into movements), each of which can differ in territorial scope. Copyright status is determined using GPT-5 with web search.

Table 8 summarizes the flags that we received on the uploads from this sample. All the compositions in the sample could legitimately be claimed in the EU, but only some in the US. We analyze whether the territorial scope of observed flags tracks where the composition is still protected under copyright. The clearest pattern is over-enforcement in the United States: 49 percent of flags involving compositions that have entered the public domain in the US nonetheless apply there. Conversely, among flags involving compositions that are still protected in the US, 62 percent cover the United States, while 38 percent are limited to other territories.[25] Taken together, these results show that territorial claim metadata only

[25]These narrower flags are not necessarily legally wrong, since the same composition may be represented by different publishers in different countries. We therefore interpret this pattern as incomplete US coverage among observed Content ID assets rather than direct evidence that each non-US flag has an erroneous territorial scope.

Table 9: Enforcement Accuracy Within Composers

	Year		IMSLP		GPT-5	
	(1)	(2)	(3)	(4)	(5)	(6)
Copyrighted in US	0.169***	0.089**	0.135**	0.062*	0.128**	0.063*
	(0.055)	(0.039)	(0.057)	(0.038)	(0.052)	(0.036)
(Constant)	0.468***	0.505***	0.491***	0.531***	0.491***	0.521***
	(0.053)	(0.039)	(0.053)	(0.041)	(0.053)	(0.039)
Num.Obs.	1488	1488	1620	1620	1696	1696
R2 Adj.	0.027	0.145	0.017	0.160	0.015	0.156
R2 Within Adj.		0.007		0.003		0.003
FE: Composer		X		X		X

Note: Linear probability model. Dependent variable: 1 if a claim covers the US. Unit of observation is composition × asset. Standard errors clustered by composer in columns (1), (3), and (5), and by composition in columns (2), (4), and (6). Columns (2), (4), and (6) include composer fixed effects; the constant shown is the mean of the estimated fixed effects. The coefficient "Copyrighted in the US" covers works that are actually still copyright-protected in the US. The copyright status is determined using publication year in columns 1-2, IMSLP copyright notices in columns 3-4, and GPT-5 with web search in columns 5-6. For more information, see Section 4.2 and Appendix D.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

weakly tracks composition-level US copyright status, with especially strong evidence that Content ID frequently issues US flags to works that should not be flagged there.

We next examine these patterns more systematically, using the specification in Equation 3. Table 9 presents results at the composition × asset level, where the dependent variable equals one if the observed flag covers the United States.

The pooled specifications (Columns 1, 3, 5) restate the descriptive pattern in Table 8, with composition × asset observations involving still-protected compositions 13 to 17 percentage points more likely to carry a US flag, while roughly half of observations involving public-domain compositions are flagged in the United States. Much of this gap reflects composer-level confounders. Once these are absorbed in the within-composer specifications (Columns 2, 4, 6), which compare works within the same composer's catalog, the difference narrows to 6 to 10 percentage points, so that conditional on the composer, Content ID applies US flags at nearly the same rate whether or not the underlying composition remains under US

copyright. Even after conditioning on copyright status, roughly half of all public-domain composition × asset observations continue to carry US flags.

Importantly, this over-enforcement does not appear to reflect a failure of Content ID's detection capabilities. Our filtering procedure retains claims whose asset titles match the uploaded compositions under the claim-to-composition verification procedure described in Appendix B.7.6; human checks were used to assess this procedure's accuracy and to resolve non-standard model outputs, not to comprehensively classify the sample by hand. Conditional on this filtering, the observed claims identify the relevant composition. However, Content ID applies territorial enforcement based on metadata provided by rights holders, who appear to define the scope of their claims at the composer- or catalog-level rather than tracking the copyright status of individual compositions across jurisdictions. The over-enforcement problem is therefore not one of detection but of metadata granularity: the institutional process by which rights holders specify territorial scope does not reflect the underlying legal boundaries prescribed by copyright laws, which vary across jurisdictions.

6 Discussion and Conclusions

Our empirical analysis documents a contrast between the technical capabilities of YouTube's fingerprinting technology and the actual enforcement outcomes that Content ID produces. When the system has reference files in its database, it identifies the correct work with high precision. At the same time, about half of infringements affecting smaller rights holders remain unaddressed, about half of the public domain works in our classical sample are nonetheless flagged, and Content ID is unable to consistently apply country-specific copyright rules. Enforcement is strongest for major rights holders with direct access to Content ID, while independent labels and individual creators experience both under-enforcement of their own rights and over-enforcement against their works.

These patterns suggest that the main constraints are not technological. YouTube decides who can access Content ID, which works can be registered, and how much effort it invests in verifying that the information rights holders put into the reference database is accurate and up to date. In an appendix, we present a simple model that formalizes these choices. This model clarifies how a platform that maximizes its own payoff will typically restrict access to automated enforcement and underinvest in data quality, even if this leads to both over- and under-enforcement from a legal perspective.

6.1 Policy Implications

Our study reveals a trade-off that policymakers face when entrusting digital platforms with enforcing legal rules. On the one hand, platforms can provide a degree of automation and scalability that traditional state enforcement mechanisms cannot match. On the other hand, the incentives of digital platforms may prevent them from implementing an enforcement system that fully aligns with policymakers' objectives. While legislators on both sides of the Atlantic have devoted considerable effort to establishing a liability framework for digital platforms that includes safe harbors and procedures for processing claims, the regulatory framework has placed far less emphasis on questions of access.

Policy intervention need not require unrestricted access to full Content ID functionality. One approach would be to lower the cost of intermediated access, for example, through collective access mechanisms. A second approach would be to require transparent, objective, and appealable access criteria for direct Content ID participation. A third approach would be to separate detection from remedies: smaller rightsholders could receive broader access to matching and reporting tools, while monetization, blocking, and takedown authority remain subject to stronger verification requirements. A fourth approach would regulate the intermediary layer itself, through fee transparency, portability of rights data, audit obligations, and dispute-resolution standards.

However, such a regulatory regime does not come without costs. Broader access and stronger verification increase operational costs, which would have to be borne by the platform operator, the rights holders, and/or the users. Providing a definitive answer as to how this trade-off between the costs and benefits of alternative platform regimes should be resolved is beyond the scope of this paper. Our aim is more modest: through an algorithmic audit, we demonstrate that such a trade-off exists. We also show that the assessment of automated enforcement systems requires system-wide algorithmic audits, which may be integrated into risk assessment and mitigation procedures that regulators are increasingly introducing for digital platforms (see, e.g., Articles 34 and 35 of the European Union's Digital Services Act).

6.2 Conclusion

In this paper, we present the first large-scale algorithmic audit of the world's largest automated copyright enforcement system: YouTube Content ID. We find that while Content ID's matching technology is generally accurate, its enforcement accuracy is lower: works represented by major labels are claimed in

about 90 percent of cases, compared to roughly 50 percent for non-major labels. Content ID is often unable to consistently enforce copyrights on works that are protected in some jurisdictions but not in others. We relate these findings to particular design choices of YouTube Content ID: access to the system is reserved for large rights holders, and automated enforcement relies on data provided by rights holders, which does not always reflect actual legal entitlements. We further explain how platform incentives can lead to such access and data quality policies.

These findings have important implications for platform governance more broadly. As platforms increasingly have to act as regulators implementing legally mandated enforcement systems, the institutional design of these systems shapes how laws are applied in practice. When access to enforcement tools is restricted based on organizational size or resources, automated systems can inadvertently favor incumbents and larger participants, even when the underlying technology performs accurately. Moreover, when automated enforcement systems rely on metadata provided by rights holders, the systems' enforcement accuracy will deteriorate if the metadata is incorrect.

These factors may explain why automated enforcement systems have not, so far, enabled a truly disintermediated marketplace for content creators. Rather than eliminating intermediaries, Content ID appears to reorganize intermediation around access to platform enforcement infrastructure: large labels and collecting societies participate directly, specialized distributors and rights-management firms provide indirect access, and smaller rightsholders outside these channels experience weaker enforcement.

References

(2019). "International Standard Recording Code."

Aguiar, L., Waldfogel, J., and Zeijen, A. (2024). "Platform power struggle: Spotify and the major record labels."

Bartholomew, T. B. (2014). "The death of fair use in cyberspace: YouTube and the problem with Content ID." *Duke L. & Tech. Rev.*, *13*, 66.

Berkowitz, A. E. (2023). "Algorithmic (in) tolerance: Experimenting with beethoven's music on social media platforms." *Transactions of the International Society for Music Information Retrieval*, *6*(1).

Erickson, K., and Kretschmer, M. (2018). "This video is unavailable." *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, *9*, 75.

European Union (2006). "Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the Term of Protection of Copyright and Certain Related Rights, as amended by Directive 2011/77/EU." Official Journal of the European Union, L 372, pp. 12–18.

Even, A. M. (2023). "Keeping the good faith: YouTube, fair use, and the DMCA." In *Boston College Intellectual Property and Technology Forum*, vol. 2023, 1–25.

Gong, Y., Chung, Y.-A., and Glass, J. (2021). "Ast: Audio spectrogram transformer." *arXiv preprint arXiv:2104.01778*.

Gray, J. E., and Suzor, N. P. (2020). "Playing with machines: Using machine learning to understand automated copyright enforcement at scale." *Big Data & Society*, *7*(1), 205395172091996.

Grosse Ruse-Khan, H. (2020). "Automated copyright enforcement online: From blocking to monetization of user-generated content."

Helft, M., and Richtel, M. (2006). "Venture Firm Shares a YouTube Jackpot." *The New York Times*.

Hesmondhalgh, D., Osborne, R., Sun, H., and Barr, K. (2021). "Music creators' earnings in the digital era." Tech. rep., UK Intellectual Property Office.

Husovec, M., and Quintais, J. P. (2021). "Too small to matter? On the copyright directive's bias in favour of big right-holders." *Global Intellectual Property Protection and New Constitutionalism. Hedging Exclusive Rights, Tuomas Mylly and Jonathan Griffiths (Eds.), Oxford University Press (2021).*

King, D., Salem, G., Wang, Y., and Wiseman, M. (2014). "Media Rights Management Using Melody Identification."

Lester, T., and Pachamanova, D. (2017). "The dilemma of false positives: Making content id algorithms more conducive to fostering innovative fair use in media creation." *24*.

Luminate (2026). "2025 year-end music report." Tech. rep., Luminate.

Mannapperuma, M., Schofield, B., Yankovsky, A., Bailey, L., and Urban, J. M. (2014). "Is it in the Public Domain? A Handbook for Evaluating the Copyright Status of a Work Created in the United States Between January 1, 1923 and December 31, 1977." https://www.law.berkeley.edu/archive/files/FINAL_PublicDomain_Handbook_FINAL(1).pdf.

Mazziotti, G. (2010). "New licensing models for online music services in the European Union: From collective to customized management." *Colum. JL & Arts*, *34*, 757.

McGrady, R., Zheng, K., Curran, R., Baumgartner, J., and Zuckerman, E. (2023). "Dialing for Videos: A Random Sample of YouTube." *Journal of Quantitative Description: Digital Media*, *3*.

Nimmer, M. B., and Nimmer, D. (2025). *Nimmer on Copyright*. New York: LexisNexis, loose-leaf service, updated regularly.

Quintais, J. P., Mezei, P., Harkai, I., C Magalhães, J., Katzenbach, C., Schwemer, S. F., and Riis, T. (2022). "Copyright content moderation in the EU: An interdisciplinary mapping analysis." *Available at SSRN 4210278*.

Simon, A. R. (2014). "Contracting in the dark: Casting light on the shadows of second level agreements." *Wm. & Mary Bus. L. Rev.*, *5*, 305.

Soha, M., and McDowell, Z. J. (2016). "Monetizing a meme: YouTube, content ID, and the harlem shake." *Social Media+ Society*, *2*(1), 2056305115623801.

Solomon, L. (2015). "Fair users or content abusers: The automatic flagging of non-infringing videos by content id on youtube." *Hofstra L. Rev.*, *44*, 237.

Tang, X. (2023). "Privatizing Copyright." *Michigan Law Review*, (121.5), 753.

Tehranian, J. (2011). *Infringement Nation: Copyright 2.0 and You*. OUP USA.

U.S. Copyright Office (2020). "Section 512 of Title 17: A Report of the Register of Copyrights." Tech. rep., U.S. Copyright Office, Washington, D.C.

Walters, T. C., Ross, D. A., and Lyon, R. F. (2012). "The intervalgram: An audio feature for large-scale cover-song recognition." In *International Symposium on Computer Music Modeling and Retrieval*, 197–213, Springer.

Weinstein, E., and Moreno, P. (2007). "Music Identification with Weighted Finite-State Transducers." In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, II–689–II–692, Honolulu, HI: IEEE.

Zhou, J., Li, Y., Adhikari, V. K., and Zhang, Z.-L. (2011). "Counting YouTube videos via random prefix sampling." In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, 371–380, New York, NY, USA: Association for Computing Machinery.

A Additional Results and Figures

As mentioned in Section 5.1, our data allows us to observe how rights holders chose to enforce their claims once Content ID has determined that an uploaded video infringes their rights. Across the three samples of our study, Table A.1 reports how rights holders chose to enforce their claims. Across all samples, monetization is the dominant strategy. 94.9% of claims in the *Contemporary* sample, 96.4% in the *Classical Old* sample, and 100% in the *Classical US-EU* resulted in the rights holder running ads on the video rather than blocking it. Blocking was rare, occurring in only 5.0% of *Contemporary* claims and 2.6% of *Classical Old* claims. This pattern suggests that Content ID functions primarily as a revenue-extraction mechanism rather than a takedown tool.

Table A.1: Rights Holder Enforcement Policy by Sample

Policy	Contemporary	Classical Old	Classical US-EU
Monetize	2,074 (94.9%)	402 (96.4%)	1,696 (100.0%)
Block	109 (5.0%)	11 (2.6%)	0 (0.0%)
Track only	3 (0.1%)	4 (1.0%)	0 (0.0%)
Total	2,186	417	1,696

Note: This table reports the enforcement policy chosen by rights holders for claims in each sample. "Monetize" means the rights holder chose to run ads on the video and collect revenue. "Block" means the video was blocked in some or all territories. "Track only" means the rights holder monitors views without monetizing or blocking.

The differential-enforcement result in Section 5.2.1 is not specific to the three major labels. Table A.2 re-estimates the rights-holder gap using a broader measure of label size, classifying a label as *large* if it ranks in the top 1% of labels in the ENAO data by number of tracks published. Columns (1) and (2) reproduce the major-label specification (without and with publication-country fixed effects), while columns (3) and (4) replace the major indicator with the large-label indicator. Large labels are 20.9 percentage points more likely to receive a flag than other labels (column 3), compared with 38.4 percentage points for majors (column 1)—the same direction but smaller in magnitude, consistent with major labels operating more comprehensive Content ID programs than other high-volume publishers. The estimates are robust to

Table A.2: Enforcement by Type of Rights Holder

	(1)	(2)	(3)	(4)
Published by a major label	0.384***	0.371***		
	(0.033)	(0.027)		
Published by a large label			0.209***	0.183***
			(0.034)	(0.017)
(Constant)	0.512***	0.573***	0.477***	0.530***
	(0.035)	(0.025)	(0.037)	(0.025)
Num.Obs.	4000	3986	4000	3986
R2 Adj.	0.048	0.076	0.039	0.060
R2 Within		0.045		0.028
R2 Within Adj.		0.044		0.028
FE: Publication country		X		X

Note: Linear probability model. Dependent variable: 1 if a recording from the *Contemporary* sample was flagged by Content ID. Standard errors clustered by country. Publication-country fixed effects are included in Columns (2) and (4); the constant shown is the mean of the estimated fixed effects. A label is classified as large if it is among the top 1% of labels in the ENAO dataset in terms of number of tracks published.

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

the inclusion of publication-country fixed effects (columns 2 and 4), with a modest attenuation suggesting that cross-country variation in Content ID adoption accounts for part of the association.

Figure A.1 provides an overview of Content ID's enforcement system.

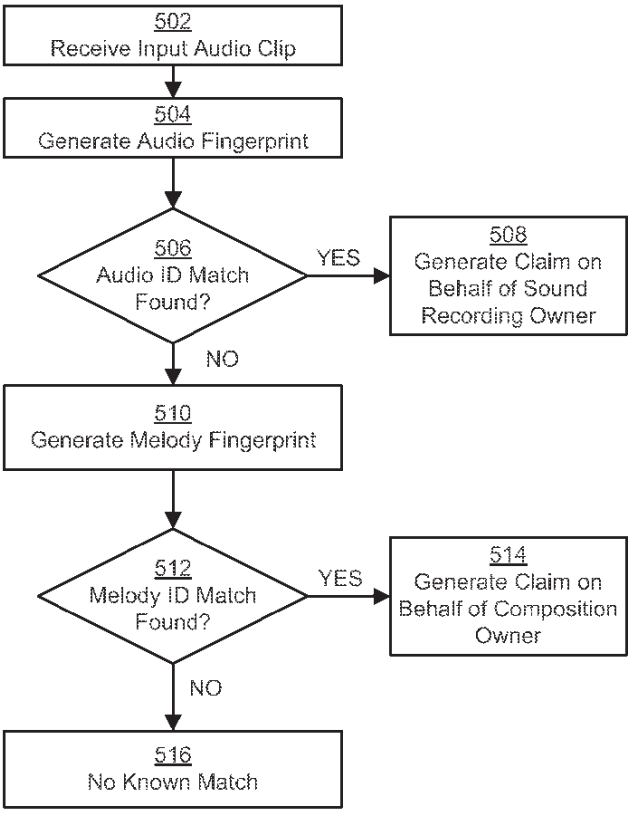

Figure A.1: Process description of ContentID system, from King et al. (2014), Figure 5.

B Data Sources

In this Appendix, we further describe the data collection process and the sources that we used.

B.1 Enforcement Data

To capture comprehensive enforcement data, we developed a custom monitoring system using a local proxy to intercept structured JSON data transmitted from YouTube's backend to the user interface. This approach provided two key advantages. First, it allowed us to collect detailed, machine-readable information about each claim, including claimant identity, specific rights asserted (composition vs. recording), and geographical scope of enforcement. Second, it enabled us to confirm that Content ID had processed every uploaded video, as the system reports processing status even when no claims are made. This verification was essential to distinguish between videos that received no claims because they were not detected and videos that were analyzed but deemed non-infringing.

We used Spotify as a source of copyrighted music. Aside from being the most popular streaming service on the global market, it also offered an API which offers detailed information about tracks. This included popularity metrics, publication year and location, along with copyright information, which indicated which label published the track, or distributed it.

B.2 Major Label Classification

Based on the information provided from Spotify, we classified each track as either being published or distributed by a major label. In line with Aguiar et al. (2024), we estimated this by evaluating whether the copyright information contained strings identifying one of the three major record labels (Warner, Sony, Universal) as well as their subsidiaries identified using Wikipedia. We further validated this using Hannes Datta's `musicMetadata` package[26] and by manually validating all positive matches, ensuring a minimal rate of false-positives.

B.3 Audio Snippets

Audio-snippet coverage was high across the final sample and did not differ materially by label type, so selection on snippet availability is unlikely to drive our results.

[26] https://github.com/hannesdatta/musicMetadata.

B.4 Selecting Samples from Every Noise at Once

To obtain a broad sampling frame for music on Spotify, we utilized the playlists generated by Every Noise at Once (ENAO).[27] Attempting to create a sample through other means would have been infeasible, primarily because Spotify does not offer an index of the songs available on the platform. Furthermore, the official playlists available through Spotify are often curated for commercial purposes or personalized for individual users, meaning they do not necessarily reflect the full diversity of music available on the platform. ENAO addresses this problem by offering genre playlists across Spotify's genre taxonomy. When the data was collected, every artist on Spotify was assigned to at least one of 6,201 genres based on a combination of listener data and other properties. For each genre, ENAO offered a playlist of tracks commonly listened to by typical listeners of that genre.

B.5 Contemporary Sample

By combining all of these playlists, we could therefore get an ENAO-based sampling frame of tracks available on Spotify, consisting of 907,473 tracks. To reduce the sample to a size that could be easily uploaded to YouTube, we created a smaller sample of 4,000 tracks. We stratified across quartiles of Spotify's popularity metric (see Figure B.1) and sampled so as to preserve the proportion of major-label tracks in the ENAO frame. Because not every Spotify track had a usable 30-second preview, we replaced tracks without usable previews with tracks from the same popularity bucket and label type (major vs independent). The resulting Contemporary sample is therefore designed to be representative of the ENAO-based Spotify track frame, subject to the limitations of ENAO coverage and Spotify preview availability.

B.6 Classical Old Sample

Sources included the Isabella Stewart Gardner Museum's concert series and collections by pianists Bernd Krueger and Kimiko Ishizaka.

B.6.1 The Isabella Gardner Museum music collection

Our primary source of classical music, where the recording is licensed under a Creative Commons license and the recorded composition has entered the public domain, is the recordings released by the Isabella

[27] https://everynoise.com.

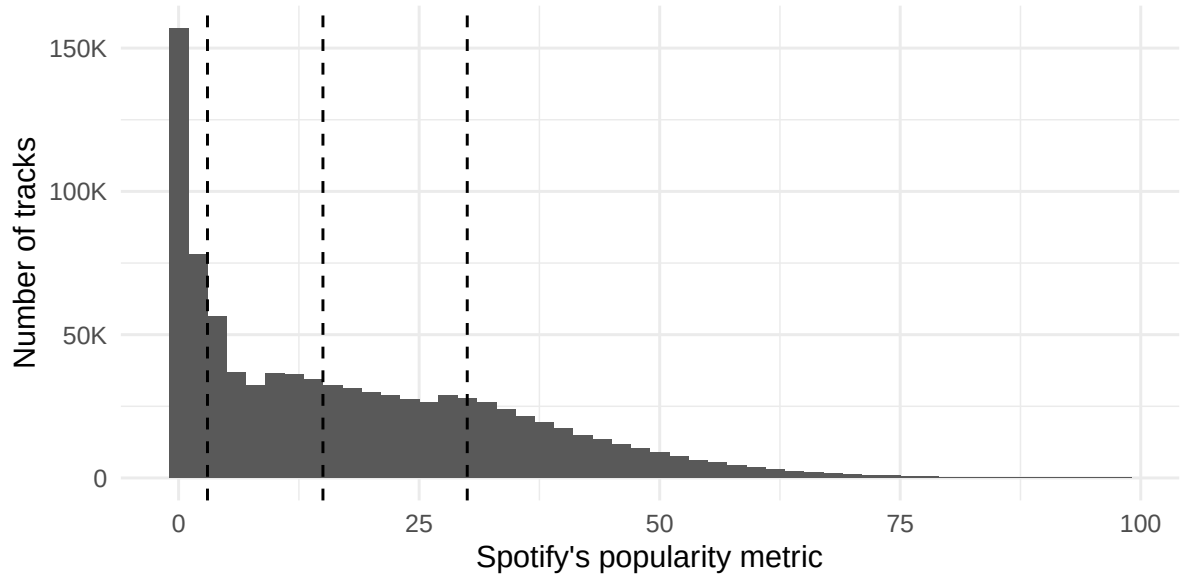

Figure B.1: A histogram showing the distribution of Spotify's popularity metric for all tracks listed on Every Noise at Once. Vertical lines are quartile marks.

Gardner Museum.[28] In total, we used 378 recordings, mostly chamber music (trios, quartets, sonatas) and solo piano pieces. These are pieces regularly performed in chamber concerts around the world. These recordings often feature performances by world-renowned musicians such as Colin Carr, Ray Chen, or Charlie Albright. As an illustration, we list some representative recordings below:

- *Johannes Brahms:* Sonata for Clarinet and Piano in F Minor Op. 120 No. 1, performed by David Deveau and Richard Stoltzman
- *Ludwig van Beethoven:* Piano Sonata No. 14 in C-Sharp Minor, Op. 27, No. 2 ("Moonlight"), performed by Gleb Ivanov.
- *Felix Mendelssohn:* Song Without Words, Op. 109, performed by Colin Carr and Thomas Sauer.
- *Franz Joseph Haydn:* String Quartet in B-Flat Major, Op. 50, No. 1, performed by Musicians from Marlboro.
- *Maurice Ravel:* La Valse, performed by Jean-Frédéric Neuburger.

[28] https://www.gardnermuseum.org/experience/music

B.6.2 Bernd Krueger

Since 1996, Bernd Krueger has been notating public domain piano compositions in notation software and making them available on his website.[29] He publishes these works in various formats, ranging from raw Musical Instrument Digital Interface (MIDI) files, along with synthesized versions in an MP3 format. Most importantly, he releases all his files under a Creative Commons license (CC-BY-SA), meaning that any reproductions cannot be claimed by rights holders on YouTube (see Section 2.2). As the interpretation is produced using music notation software, it is mechanical in nature, but remains enjoyable because Bernd Krueger has adjusted the playback to account for musical elements such as ritardando and rubato. Examples of these recordings include Mozart sonatas, piano works by Debussy, Liszt études, and Mussorgsky's Pictures at an Exhibition.

B.6.3 Kimiko Ishizaka

Japanese pianist Kimiko Ishizaka has released professionally recorded interpretations of three Bach piano works on her Bandcamp page:[30] *Das Wohltemperierte Klavier*, *Die Kunst der Fuge*, and the *Goldberg Variations*. These recordings are released under the most permissive Creative Commons license (CC0), and the compositions are in the public domain (Bach died in 1750).

B.6.4 Note on Differing Recording Formats

The recordings from the Gardner Museum are often of entire compositions, whereas recordings from other sources are always a single section, movement, or track. As this could affect the probability of a track being claimed (if it has more sections, there is a bigger pool of tracks that Content ID could match), we counted the number of different sections in each work from the Gardner Museum. We use this variable only as an unreported nuisance control in auxiliary specifications involving the *Classical Old* sample; including it does not affect the substantive results, so we do not report its coefficient in the main tables. Section counts were first manually annotated by cross-referencing each recording with its IMSLP entry, and subsequently verified using GPT-5.4 with web search. The two methods agreed on 275 out of 378 recordings. In 24 cases where they disagreed, we reviewed the IMSLP entry again and resolved the discrepancy. The remaining 79 recordings – typically excerpts, arrangements, or collections – required

[29] http://piano-midi.de.
[30] https://kimikoishizaka.bandcamp.com.

manual verification against IMSLP or comparable recordings on Spotify; of these, 4 required listening to the recording directly.

The prompt in Listing B.1 was used for the LLM verification. Web search capabilities were activated.

Listing B.1: Prompt used to determine movement count for the Gardner Recordings

```
You are a musicology research assistant.
I will give you the title and composer of a classical music recording. Your task:

1. Search IMSLP (imslp.org) for this composition.
2. Determine how many separate tracks this composition would be split into on a
    streaming \
platform like Spotify or Apple Music.

Rules:
- Count MOVEMENTS, not internal sections. A single-movement piece that has
    internal sections \
(e.g. an aria with a recitative introduction, or a cantata performed as one
    continuous piece) \
counts as 1. A sonata with 3 movements counts as 3.
- The key question is: how many separate tracks would this appear as on a
    streaming platform?
- Do NOT count lost, incomplete, or dubious movements that would not appear in a
    standard \
modern performance or recording.
- Variations: each variation is typically a separate track on streaming platforms
    . Count the \
introduction, each individual variation, and the coda/finale as separate tracks.
- If the recording title indicates it is an EXCERPT, a single movement from a
    larger work, \
an arrangement, or a collection/selection of pieces, report no_sections as null
    and set \
is_excerpt to true. These need manual verification.
- If you are not confident enough in the number of sections, report no_sections
    as null. \
It is better to report null than to guess.
- Use IMSLP as your primary source. If IMSLP does not have the composition, say
    so.\
```

B.7 Classical US-EU Sample

This section describes the data collection pipeline used to construct the *Classical US-EU* sample. The goal was to collect YouTube recordings of classical compositions by early-20th-century composers whose works have mixed copyright status: still protected in the EU (where the copyright term is life plus 70 years), but potentially in the public domain in the US (where copyright depends on publication dates, renewal filings, and other factors). The pipeline involved five main stages: composer selection,

composition extraction, YouTube video search and verification, video upload and claim collection, and claim-to-composition matching.

B.7.1 Composer Selection

We queried Wikidata for composers meeting the following criteria: (i) their occupation was listed as composer or a subclass thereof, (ii) they were born before 1910 and died after 1954, (iii) they have an English Wikipedia article, and (iv) they have at least five notable works listed on Wikidata. This yielded 337 candidate composers.

For each composer, we searched for their page on the International Music Score Library Project (IMSLP) using the Google Custom Search API. We verified that each search result pointed to the correct composer using GPT-4o-mini, which flagged 99 mismatches (e.g., searches returning category pages for genres or instrumentation rather than the composer). After this filtering, 238 composers had valid IMSLP pages. From these, we selected 25 composers for the final sample based on the availability of compositions with varying copyright status.

B.7.2 Composition Extraction

For each selected composer, we scraped their IMSLP category page to extract a list of compositions, yielding 2,121 compositions across 25 composers. For each composition, we collected the title, URL, copyright notice, and composition year. We also parsed the structured metadata table on each composition's IMSLP page into a JSON object containing fields such as work title, alternative titles, opus and catalog numbers, movements, instrumentation, year of composition, and dedication.

B.7.3 YouTube Video Search and Verification

We searched the YouTube Data API for recordings of each composition, using the composition title as a query. Each search was restricted to the US region and returned up to five results, excluding YouTube's auto-generated "Topic" channels. This produced 7,013 candidate video-composition pairs (6,267 unique source videos). Some videos appeared in search results for multiple compositions; for instance, a recital video containing several works by the same composer. We deduplicated at the video-composition level before sending pairs for verification, but did not exclude videos matched to multiple compositions, as the

claim-to-composition matching stage (described below) later classified which specific composition each claim referred to.

To verify that each video actually contained a recording of the target composition, we used GPT-4o via the OpenAI API. The model was given the composition's title, composer, and structured metadata, along with the video's title and description, and asked to return `true` or `false`. The full system prompt can be found in Listing B.2.

Listing B.2: Prompt used to verify whether a videos contained our recording of interest

```
You are an assistant that determines whether a video contains a recording
of a specific musical composition. You will be given:

1. The title and composer of the musical composition, as well as general
   information about the composition in JSON format, including the title
   and composer
2. Information about a YouTube video, the title and description

Return one of these values:
- 'true' - The video contains a recording of the composition (partial
  or full)
- 'false' - The video does not contain a recording of this specific
  composition or if there is not enough information to determine

Notes for matching:
- Alternative naming conventions (opus numbers, catalog numbers,
  translations into other languages) should be considered matches
  (return true)
- Consider that the video title may omit some special characters
  (non-ASCII characters, diacritics, etc.), but that is okay if it
  matches closely enough
- Arrangements and transcriptions should not be considered matches
  (return false)
- Different performers or instrumentation of the same work should be
  considered matches (return true)
- Only return the boolean value true or false, without any quotes or
  additional content
```

The prompt also included six few-shot examples: three matches and three non-matches. The match examples demonstrated that the model should accept titles in different languages (e.g., Czech and English for a Hába string quartet), partial title matches (e.g., "Kyrie te Deum Laudamus" for a full mass setting by Perosi), and French-language matches (Tailleferre's *Jeux de plein air*). The non-match examples showed cases where the video featured a different work by the same composer (Goossens's Violin Concerto instead of Rhythmic Dance), where a composition was mentioned in a video description but not performed

(Aubert's Pie Jesu mentioned biographically in a video of his piano Esquisses), and where two works by the same composer had superficially similar names (Delage's *Schumann...* vs. *Tout allégresse*). Here, "different instrumentation" refers to different recordings or performer/ensemble descriptions of the same composition, not derivative arrangements or transcriptions, which were excluded.

We manually verified a random sample of 153 video-composition pairs. The model achieved 93% accuracy overall, with 100% precision and 91% recall (F1 = 0.95). All 11 observed errors were false negatives (cases in which the model rejected a valid match), meaning that we observed no incorrectly matched videos in the validation sample. The verification model accepted 4,748 pairs (4,714 unique source videos) covering 1,519 compositions as likely containing the target work.

Because classical works are often performed and recorded by different artists, we uploaded up to five source videos per composition that the verification model classified as likely containing the target work. The upload pipeline processed 4,707 of these source videos.

B.7.4 Isolating Composition Claims

Our goal was to determine whether compositions themselves, independent of their sound recordings, are subject to copyright claims on Content ID in the United States. However, Content ID presents a technical challenge: the system first attempts an Audio ID match and only proceeds to Melody ID if no recording match is found (Figure A.1). As a result, an unmodified recording generally does not reveal whether the underlying composition would independently trigger a Melody ID claim. To work around this limitation, we employed an iterative approach to isolate claims on the composition. After uploading each recording, we checked for claims. When a video received a recording claim, we applied audio transformations (low-pass filtering to remove high frequencies and pitch modulation to alter the spectral signature) and re-uploaded the modified version. These transformations degrade the recording's acoustic fingerprint while preserving the underlying melodic and harmonic structure, which Content ID uses to identify the composition (via Melody ID). We repeated the transformation process up to six times per recording until we either (a) received a composition claim, (b) received no claim at all, or (c) exhausted the iteration limit. Of the 4,707 recordings processed through this pipeline, 1,820 (38.7%) resulted in a Melody ID (composition) claim, 1,519 (32.3%) received no claim at all, and 1,368 (29.1%) exhausted the

iteration limit. These last recordings continued to receive only AudioID (recording) claims despite the audio transformations, indicating that the recording fingerprint could not be fully removed.

B.7.5 Claim-to-Composition Matching

After collecting Content ID claims, we needed to determine whether each claim referred to the composition we had uploaded or to an unrelated work. This is necessary because Content ID claims identify an "asset" by title, which may use different naming conventions, languages, or refer to individual movements rather than the full work.

We used GPT-4o via the OpenAI API to match each claim's asset title to the IMSLP composition. The system prompt can be seen in Listing B.3.

Listing B.3: Prompt used to verify whether a copyright claim referenced our composition of interest

```
You are an assistant that determines whether a copyright claim is
referring to a certain composition. You will be given the following:

1. The title and composer of the musical composition, as well as general
   information about the composition in JSON format, including the title,
   composer, and movements.
2. Information about the copyright claim. This consists of the
   'asset_title', which is the title of the asset, which could refer to
   the name of the composition, the name of the specific movement that
   is being claimed, or both.

Return one of these values:
- 'true' - if the copyright claim is referring to the specific
  composition in question
- 'false' - if the copyright claim is not referring to that composition
- 'null' - if there is not enough information available

Notes for matching:
- You will find information on the movements in the composition in the
  JSON formatted data, which could be useful if the copyright claim does
  not name the work explicitly.
- It could be that the composition is listed in different languages
  between the claim and the baseline information. If it's a match after
  a reasonable translation, consider it to be a match.
```

The prompt included two few-shot examples. The match example showed a claim with asset title "1. Prologue" correctly matched to Martinů's *La revue de cuisine* by cross-referencing the movement list in the composition metadata. The non-match example showed a claim for "Choros No. 5 'Alma brasileira' " correctly rejected as a match for Villa-Lobos's *A lenda do caboclo*, despite both being works by the

same composer. The model output determined the claim-to-composition classification in almost all cases; human checks were used to assess the procedure's accuracy and to resolve non-standard model outputs. Of the 5,507 asset–composition records evaluated by the model, 4,523 (82.1%) were classified as referring to the target composition, involving 1,049 compositions and 4,065 distinct assets. The remaining 984 records were rejected as mismatches.

B.7.6 Final Sample

We found 7,013 candidate video-composition pairs (6,267 unique source videos) for 1,879 candidate compositions by 25 composers. Video-composition verification accepted 4,748 pairs (4,714 unique source videos) covering 1,519 compositions, and 4,707 verified source videos entered the upload and filtering pipeline. The upload pipeline produced claims for 3,188 source videos: 1,820 eventually yielded a Melody ID (composition) claim and 1,368 continued to yield only AudioID (recording) claims after the iteration limit. We used GPT-4o to match each claim's asset title to the target composition (as described above). After automated claim-to-composition verification, 1,049 compositions had at least one claim classified as referring to the uploaded composition.

The final sample was restricted to compositions where (i) the video-composition match was classified as correct, (ii) the claim-to-composition match was classified as correct, (iii) at least one matched claim was a Melody ID (composition) claim rather than only an AudioID (recording) claim, and (iv) copyright status could be determined using GPT-5 with web search (see Appendix D). Among the 1,049 compositions with at least one matched claim, GPT-5 could determine US copyright status for 1,000 compositions. Of those, 631 had at least one matched Melody ID claim; the remaining 369 were matched to the correct composition asset, but only through AudioID (recording fingerprint), not through Melody ID (composition fingerprint). We dropped one singleton composer, yielding a final sample of 630 compositions by 22 composers, with 1,696 composition $\times$ asset observations. A single composition can be claimed by multiple assets from different rights holders, or by the same rights holder multiple times when the composition contains several movements. These assets may differ in territorial scope. We therefore use the composition $\times$ asset pair as the unit of observation.

C Claim Verification

When Content ID generates a flag against an uploaded video, the flag identifies a specific reference asset in the system's database. However, a flag against the correct video does not necessarily mean that the system identified the correct underlying work (sound recording or composition). To assess matching accuracy, we verified whether each claim's reference asset corresponded to the composition or recording that we had actually uploaded. We applied different verification strategies to the *Contemporary* and *Classical Old* samples, reflecting the different metadata available in each case.

C.1 Contemporary Sample

For the *Contemporary* sample, we compared the Spotify track title of the uploaded recording with the asset title reported in the Content ID claim. We applied a two-stage verification process.

In the first stage, we performed automated string and artist matching. We normalized both the track title and the asset title by converting to lowercase, decomposing Unicode characters and stripping diacritical marks, removing parenthetical expressions (such as "feat." tags or remix labels), removing punctuation while preserving characters from all scripts, and collapsing whitespace. We then checked whether either normalized title string was contained within the other. For claims that passed the title check, we additionally verified that at least one artist from the Spotify track metadata overlapped with the Content ID asset artists, using the same normalization and allowing for name reordering (e.g., "Claudio Fasoli" matching "Fasoli Claudio") and ignoring uninformative labels such as "Various Artists." This procedure resolved 1,989 of 2,186 claims (91%), all as matches. The normalization handled differences in capitalization, diacritics (e.g., "Dobrý Pocit" vs. "Dobry pocit"), parenthetical formatting (e.g., "Girl On Top - Raving George Remix" vs. "Girl On Top (Raving George Remix)"), and non-Latin scripts including Cyrillic, Chinese, Japanese, Korean, Thai, Hebrew, and Arabic.

In the second stage, we submitted the remaining 197 claims to OpenAI's GPT-5.4 model with web search enabled. For each claim, we provided the Spotify track title and record label, along with the Content ID asset title and asset artists, and asked the model to determine whether the two referred to the same underlying song. The model returned a structured response with a binary match result, a confidence level, and a brief reasoning. These residual cases included cross-language title differences (e.g., "LOVE LETTER RETURNS" vs. the Japanese title), abbreviated or reformatted titles (e.g., "3 Japanese Dances:

3. Dance with Swords" vs. "Three Japanese Dances: No. 3, Dance with Swords"), and entirely unrelated assets matched in error. The system prompt used for this stage is shown in Listing C.1.

Listing C.1: Prompt used to evaluate accuracy of claims against the Contemporary sample

```
You are an expert music librarian. A researcher uploaded a song to YouTube
and received a Content ID copyright claim. Your task is to determine whether
the claim refers to the **same underlying song/work** that was uploaded.

You will receive:
1. The Spotify track title and record label of the uploaded song.
2. The Content ID claim's asset title and asset artists.

Important considerations:
- The "asset artists" field lists the performer or rights holder who
  registered the reference in Content ID. This may differ from the
  original artist.
- Titles may differ slightly due to formatting, language, featured-artist
  tags, remix/version suffixes, or transliteration.
- A claim is a match even if the performers differ, as long as it is the
  same underlying song (e.g. a cover version claiming the original is still
  the same work).
- A claim is NOT a match if it is for a completely different song that
  happens to have a similar or short title.

Use web search if you need to confirm whether two titles refer to the same
song.

Always commit to "yes" or "no". Use the confidence field to express
uncertainty.
```

Of the 197 claims evaluated by the model, 152 were judged to refer to the same underlying song as the uploaded track, and 45 were judged to be incorrect matches. Combined with the 1,989 string-matched claims, this yields 2,141 of 2,186 Contemporary claims (97.9%) where Content ID identified the correct work. Direct verification of the exact reference recording was not possible for any claim because YouTube reports claim metadata but does not provide the reference audio file on which the claim is based. The first-stage procedure therefore provides metadata-based evidence of recording-level correspondence through title and artist overlap, but it cannot independently verify the reference audio. Conversely, residual claims without title-and-artist overlap may still involve the same recording if the Content ID asset metadata is incomplete or inconsistent. The additional limitation of the second-stage GPT step is that its prompt asked whether the claim referred to the same underlying song or work, rather than

explicitly requiring the same sound recording. If any of the 152 GPT-accepted residual claims referred to a different recording of the same composition, the recording-level matching accuracy would be lower.

C.2 Classical Old Sample

For the *Classical Old* sample, automated string matching was not feasible because classical music metadata is substantially more heterogeneous. The same composition may appear under different names, translations, opus numbers, or movement labels, and the Content ID asset title typically lists a performer rather than the composer. We therefore submitted all claims to GPT-5.4 with web search enabled.

For each claim, we provided the model with the uploaded composition's title and composer name, along with structured work and movement information where available. On the claim side, we provided the asset title and asset artists. Importantly, we excluded the claim's own composition metadata fields (composition title and composition writers) from this verification step, as these fields come from a separate metadata layer that is sometimes inconsistent with the asset itself. By evaluating against the asset title alone, we ensured that the verification reflected Content ID's actual matching behavior rather than errors in metadata entries.

The model was instructed to use web search to resolve ambiguous identifiers such as opus numbers, catalog numbers (BWV, K., WoO, D.), and movement names. As with the Contemporary sample, it returned a binary match result, confidence level, and reasoning for each claim. The system prompt can be found in Listing C.2.

Listing C.2: Prompt used to evaluate accuracy of claims against the Classical Old sample

```
You are an expert in classical music. A researcher uploaded a recording of a
classical composition to YouTube and received a Content ID copyright claim.
Your task is to determine whether the claim refers to the **same underlying
musical composition** that was uploaded.

You will receive:
1. Information about the uploaded composition: composer, title, and
   optionally the parent work name, movement/part name, and number of
   sections.
2. Information from the Content ID claim: asset title, asset artists, and
   optionally the claim's own composition title and composition writers
   fields.

Important considerations:
- Classical compositions have many names, translations, and catalog numbers
  (e.g. "Moonlight Sonata" vs "Piano Sonata No. 14" vs "Sonata quasi una
```

```
  fantasia").
- Opus numbers and catalog numbers (BWV, K., Op., WoO, D., etc.) are
  strong identifiers.
- The claim's asset_artists field lists the performer, not the composer.
  The composer may appear in the composition_writers field instead.
- Movements of a larger work may be claimed separately. A claim for a
  specific movement matches if it is a movement of the uploaded
  composition.
- A claim for a DIFFERENT movement, opus number, or work by the same
  composer is NOT a match.
- Use web search to look up opus numbers, catalog numbers, or movement
  names when needed to resolve ambiguity.

Always commit to "yes" or "no". Use the confidence field to express
uncertainty.
```

Although this system prompt lists the claim's own composition title and composition writers as *optional* inputs, these fields were not supplied in the Classical Old accuracy pass. The information provided on the claim side was limited to the asset title and asset artists, consistent with the description above. The claim's composition metadata is instead evaluated in a separate pass, reported below (see Listing C.3 and Table C.1). The reported matching accuracy is therefore based on the asset title and artists alone and does not depend on the composition metadata.

Of the 417 Classical Old claims, 386 (92.6%) were judged to match the correct composition, and 31 (7.4%) were judged incorrect matches. Breaking these down by claim type: of the 295 Audio ID claims (matched by audio fingerprint), 277 (93.9%) were correct, and 18 (6.1%) were incorrect; of the 122 Melody ID claims (matched by melodic similarity), 109 (89.3%) were correct, and 13 (10.7%) were incorrect.

C.3 Composition Metadata Consistency

Content ID claims on composition rights (Melody ID claims) carry an additional metadata layer: a composition title and composition writers field. Unlike the asset title, which is paired with the reference audio recording, these composition metadata fields are entered separately and can be incorrect even when the underlying match is correct.

To quantify the prevalence of metadata errors, we conducted a separate verification pass on the 122 Melody ID claims that carried composition metadata. In this pass, we asked GPT-5.4 to evaluate whether the composition title and writers were consistent with the asset title, without reference to the uploaded work. This allowed us to distinguish between cases where Content ID's matching technology identified the

wrong work (a genuine technical error) and cases where it identified the correct work but the accompanying composition metadata was incorrect (a data quality issue). The system prompt is shown in Listing C.3.

Listing C.3: Prompt to evaluate composition metadata information for the Classical Old sample

```
You are an expert in classical music metadata. A Content ID claim on
YouTube has two layers of metadata:
1. Asset-level: the asset title and asset artists (paired with the actual
   audio recording).
2. Composition-level: a separate composition title and composition writers
   field (entered separately, sometimes incorrectly).

Your task is to determine whether these two layers are **consistent** -
i.e., do the composition title and composition writers describe the same
work as the asset title?

Use web search to look up works, opus numbers, or catalog numbers when
needed.

Answer "yes" if consistent, "no" if they describe different works.
```

Of the 122 Melody ID claims, the model judged 44 (36.1%) to have consistent composition metadata and 78 (63.9%) to have inconsistent metadata — that is, the composition title or writers fields described a different work than the asset title.

Table C.1 crosses the asset-level verification with the metadata consistency check. Of the 109 claims where Content ID matched the correct composition, only 38 (34.9%) had consistent composition metadata; the remaining 71 (65.1%) had metadata describing a different work despite the correct audio match. Among the 13 incorrect asset matches, 6 had consistent metadata, and 7 had inconsistent metadata. The high rate of metadata inconsistency among correctly matched claims indicates that composition metadata errors are primarily a data quality issue in rights registration rather than a failure of the matching technology itself.

Table C.1: Cross-tabulation of asset-level match correctness and composition metadata consistency for the 122 Melody ID claims in the Classical Old sample.

	Consistent metadata	Inconsistent metadata	Total
Correct asset match	38	71	109
Incorrect asset match	6	7	13
Total	44	78	122

D Copyright Verification

Determining the precise copyright status of musical compositions is a complex task, particularly for the United States. In the European Union, copyright in compositions lasts for 70 years after the author's death. This copyright term typically also applies retroactively, i.e., to works that were composed before the European copyright term reached its current length. In the United States, the copyright term for compositions varies. For compositions created on or after Jan. 1, 1978, a copyright term of 70 years after the author's death applies, leading to the same copyright term as in the EU. Older compositions are, in general, protected for 95 years after their publication. But the details of the rules are complex. As a result, for any composition published between 1931 and 1977 (calculated at the time of this writing, i.e. in 2026), determining whether the composition is still in copyright in the US requires the following information (some of which may not be relevant in individual cases):

- When the first publication took place;
- Whether or not the first publication took place in the United States;
- When and where publications in a foreign country took place;
- Whether or not a copyright notice in the proper form was included on the work;
- Whether or not a copyright renewal notice was filed in the US;
- Nationality of the composer;
- Year of death of the composer (important for posthumous works);
- Whether or when the work lapsed in the public domain in his/her home country.

For an extensive treatment of determining copyright status, see (Nimmer and Nimmer, 2025, chapter 9). A practical handbook to determine copyright status is provided by Mannapperuma et al. (2014).

As mentioned in Section 4.1, determining the true copyright status of an early-20th-century composition under US copyright law is typically a task reserved for highly paid copyright lawyers. Providing a legally fully accurate determination at scale is effectively infeasible for a large-scale quantitative study,

and would likely be equally impracticable for a platform such as YouTube. To approximate the true copyright status of these compositions and to adopt an approach that YouTube could implement at scale, we use three methods to determine the copyright status of compositions. This approach allows us to assess whether our findings are sensitive to the underlying classification approaches.

D.1 Publication Year

The simplest way to determine when a composition was published, as the general rule for the compositions that are relevant for our study, is that they are in copyright for 95 years after the first publication (17 U.S.C. §304). Our source for this date is the International Music Score Library Project (IMSLP), more specifically the field "First Publication" which is available for 89% of the compositions in our sample. However, this date is not always correct, as it is based not on the first published manuscript, but on the oldest one IMSLP hosts.

D.2 IMSLP Copyright Information

AS a next step, we turn to more detailed information provided by IMSLP. For compositions with a complicated copyright term across countries, IMSLP includes tailored information about where the copyright term has likely expired, and where it has not. This information is provided by IMSLP contributors but takes more factors into consideration than just publication year.[31] This information was available for 95% of the compositions in our sample. However, this might not generalize to a broader set of compositions, as these copyright summaries are most often only included when there is ambiguity around the copyright term.

D.3 GPT-5 and Web Search

As a more general alternative to relying on composition-specific information from IMSLP, we turned to OpenAI's GPT-5 reasoning model and paired it with web search capabilities. With the search option, the model could both retrieve relevant information about the composition and reason about the copyright status. The cost of verifying the copyright status came to less than $0.50 per composition using OpenAI's API. We show the prompt used in listing D.1

Listing D.1: Prompt used to verify copyright status for the Classical US-EU sample

[31] https://imslp.org/wiki/Public_domain.

```
You are an expert legal assistant who is supposed to determine the U.S. copyright
    status
of a provided musical composition as of February 2025.
Your analysis must be thorough and follow a two-step process:

Step 1: Factual Research
First, find and clearly state the following essential information. This data is
   critical
for the legal analysis.
* Composer's Full Name, Life Dates and Nationality
* Date of Creation
* Date and Place of First Publication
* Whether and when the composition was registered with the U.S. copyright office
   and
  whether and when a renewal notice was filed
* Whether a copyright notice in the proper form was included in the publication
* Whether or not the copyright has lapsed in the home country of the composers if
    they
  are not U.S citizens

Step 2: Legal Analysis and Conclusion
Based on the facts gathered in Step 1, provide a detailed explanation of the work
   's
copyright status in the United States. Your reasoning must address all relevant
   points
of U.S. copyright law, including:
* The copyright status of works first published before 1930.
* Rules for works published between 1930 and 1977, specifically addressing
   copyright
  notice and renewal requirements and place of publication.
* The copyright term for works created or published on or after January 1, 1978.
* The impact of the composer's nationality (French) and the potential for
   copyright
  restoration for foreign works under the Uruguay Round Agreements Act (URAA).
* Copyright notice
* Renewal notice
* Place of publication
* Composer's nationality
* Potential copyright restoration for foreign works under the Uruguay Round
   Agreements
  Act (URAA)
* Also consider that rules on copyright term, renewal, and restoration are
   different
  for works published before 1930, before 1978, and after

Finally, provide a clear conclusion stating whether the work is in the public
   domain
in the U.S. or is likely still protected by copyright.
Also, provide a confidence score (low, medium, high) for your prediction. If the
   work
is still protected by copyright, provide the year when this protection expires.

## Possible data sources
```

```
For information on copyright term, the following pages are good sources:
- Cornell Library guide on copyright term
  https://guides.library.cornell.edu/copyright/publicdomain
- IMSLP page on what is in the public domain
  https://imslp.org/wiki/Public_domain
```

D.4 Comparison Between Different Methods

Figure D.1 shows the classification patterns across the three methods listed above. We compute this comparison at the composition level for the 630 analyzed compositions in the *Classical US–EU* sample. GPT-5 provided copyright terms for all compositions in the sample, classifying 375 as still protected in the United States and 255 as in the public domain. The three methods agree closely: nominal Krippendorff's α is 0.943 across all three methods. Pairwise agreement is similarly high, with $\alpha = 0.969$ for publication year and IMSLP, $\alpha = 0.967$ for publication year and GPT-5, and $\alpha = 0.935$ for IMSLP and GPT-5. The two largest classification patterns are full agreement that a composition remains protected (295 compositions) and full agreement that it is in the public domain (224 compositions). Only 25 of the 626 compositions with at least two available classifications contain a substantive disagreement among the available methods.

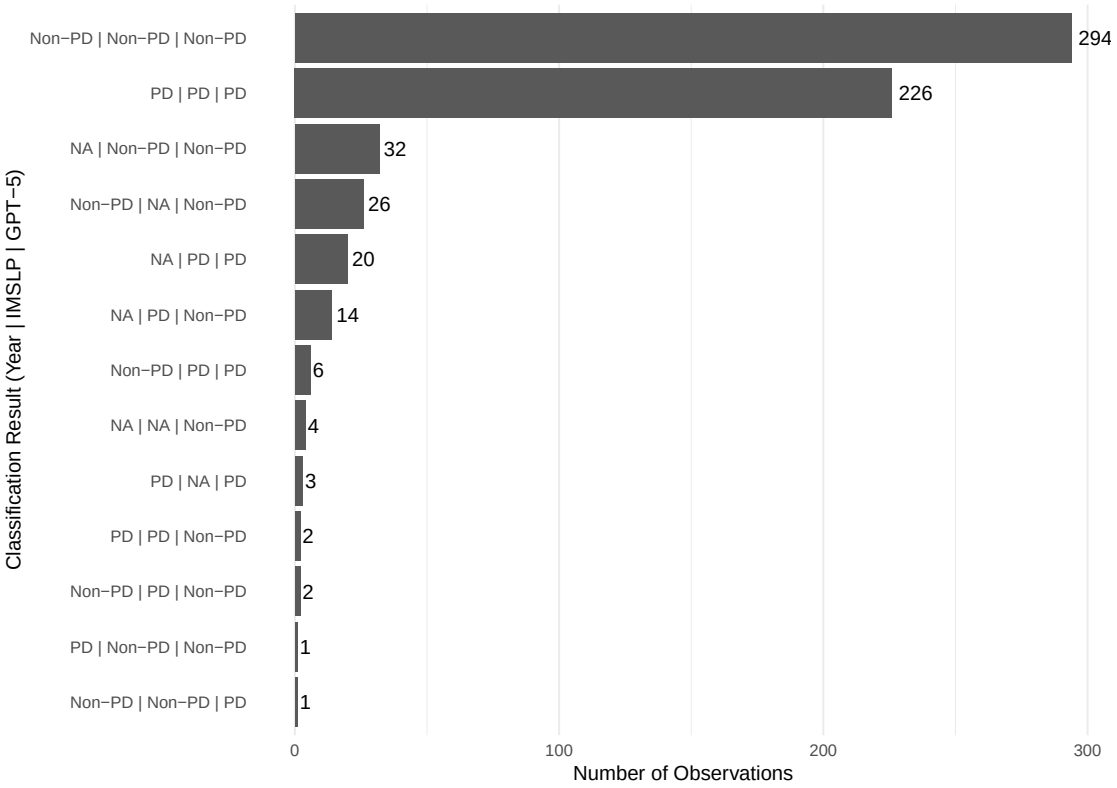

Figure D.1: Frequency of classification patterns across the three methods (publication year, IMSLP, GPT-5). *"Non-PD / Non-PD / Non-PD" denotes that all three methods classified the composition as not being in the public domain, whereas "NA / PD / Non-PD" indicates that the copyright status could not be determined using publication year, IMSLP denoted the composition as in the public domain, but GPT-5 classified it as not in the public domain.*

E Access to Content ID in Practice

As discussed in Section 2.2, Content ID is not openly available. YouTube restricts it to "copyright owners with the most complex copyright management needs" and does not publicly disclose the full roster of rights holders granted access. This appendix details the institutional arrangement through which rights holders obtain access in practice. This arrangement provides useful context for interpreting our results, since differences in access across types of rights holders are among the mechanisms that could shape the differential-enforcement patterns we examine in Section 3.3.1. We first describe the ladder of copyright-management tools that YouTube offers (Section E.1), then the three routes through which rights holders reach the monetization-capable tier (Section E.2), and finally the schemes that distributors offer and their economics (Section E.3).

E.1 The Ladder of YouTube Copyright Tools

Content ID sits at the top of a tiered set of copyright-management tools, and only this top tier permits automated monetization. In ascending order of capability, YouTube offers four tools. The public *Webform* is available to any rights holder but supports only after-the-fact DMCA copyright takedown requests. The *Copyright Match Tool* automatically detects re-uploads of a creator's own videos, but it acts only on full-video copies and offers removal rather than monetization. Its high-volume counterpart, the *Enterprise Copyright Match Tool*—which replaced the Content Verification Program in early 2025—adds a bulk interface for searching out and removing infringing uploads at scale.[32] Finally, Content ID itself is the only tier that fingerprints partial matches against a reference database and allows a rights holder to monetize, block, or track matching uploads.[33] The decisive distinction for our purposes is that the lower tiers can only remove infringing content, whereas Content ID is the only tool that turns a match into advertising revenue. A rights holder without Content ID access can therefore have an infringing video taken down, but cannot monetize it, and so is shut out of the automated licensing market that accounts for the overwhelming majority of copyright value on the platform (Section A).

The four tiers differ enormously in how many rights holders can reach them. According to YouTube's Copyright Transparency Report, while the basic Webform is available to billions of users and the Copyright

[32] YouTube's Copyright Transparency Report labels this tier the "Enterprise Webform"; the two names refer to the same tool.
[33] See YouTube's overview of its copyright-management tools at https://support.google.com/youtube/answer/9245819.

Match Tool to more than four million, only 7626 entities have access to Content ID—the sole tool that permits automated monetization—and just 4454 of them actively used it.[34] This figure is of the same order of magnitude as the "over 9,000 rightsholders" reported by the U.S. Copyright Office in 2020 (U.S. Copyright Office, 2020), indicating that access to the monetization tier has not meaningfully broadened over time even as the platform has grown. Against billions of potential rights holders worldwide, fewer than eight thousand direct participants—of which only about three in five are active—is a striking measure of how restricted the monetization-capable tier remains.

E.2 Three Routes to the Monetization Tier

In practice, a rights holder can reach the Content ID tier through one of three routes. The first is a direct partnership, under which the rights holder applies to YouTube, is vetted against the "complex needs" criteria, and administers its own reference files, claims, and disputes.[35] YouTube treats a history of valid DMCA takedown requests as the primary indicator of eligibility, which mechanically favors incumbents that already operate enforcement operations at scale. In practice, this route is open to the three major record labels and to a limited set of large publishers, aggregators, and studios.

The second route is intermediated access through a distributor or aggregator. Because the major labels and large distributors hold direct Content ID partnerships, they can enroll the catalogs of the smaller artists and labels they distribute as assets under their own Content ID accounts. An independent artist who could never qualify for Content ID directly thus obtains access by signing with a distributor, which administers the reference file and claims on the artist's behalf and remits a share of the resulting revenue.[36] This is the route available to the typical independent artist, and it is the mechanism behind our assumption that participation among smaller rights holders is partial and uncertain ($P(\text{in DB}) < 1$). While YouTube does not name the rights holders that hold Content ID access, it publishes a directory of approved service providers, the YouTube Services Directory, and the entries offering Content ID administration are

[34] See YouTube's Transparency Report for 2025, "Everyone has access" page: https://transparencyreport.google.com/youtube-copyright/everyone-has-access. Figures are as reported by Google for the most recent reporting period available at the time of writing.

[35] YouTube states that Content ID eligibility turns on holding exclusive rights to material that can be claimed and on "demonstrated need"; see "Get started with Content ID," https://support.google.com/youtube/answer/1311402.

[36] YouTube's own documentation describes how distributors ("digital aggregators") deliver music and manage Content ID rights on artists' behalf, charging either a flat fee or a percentage of royalties collected; see "Music aggregators," https://support.google.com/youtube/answer/9105565.

overwhelmingly music distributors, aggregators, and multi-channel networks rather than individual rights holders.[37] The publicly disclosed face of Content ID access is thus the layer of intermediaries, not the rights holders on whose behalf they act.

The third route is a multi-channel network (MCN), which can enroll affiliated channels' content under the network's Content ID access.[38] MCNs are most relevant for creators whose primary output is YouTube videos rather than released recordings, and we note them for completeness.

A structural feature of the second route is that these intermediaries are themselves large, consolidated firms—frequently the major labels or companies they own, and otherwise a small number of publicly listed distribution groups. Sony Music owns the distributors The Orchard and AWAL; Universal Music Group owns Ingrooves, FUGA, and PIAS, and in February 2026 completed its acquisition of Downtown Music Holdings, the parent of CD Baby and the publishing administrator Songtrust;[39] and TuneCore belongs to Believe, a large publicly listed distributor that is independent of the majors but is itself a major consolidator of independent-artist distribution. The disintermediation that automated enforcement was expected to deliver (Section 1) is therefore incomplete. Smaller rights holders reach the enforcement system largely by routing through—and paying—the very incumbents—major labels and large distribution groups alike—that the technology was supposed to make unnecessary.

This institutional structure bears directly on how we interpret a small rights holder's absence from Content ID. A recording can go unclaimed for several distinct reasons. The rights holder may have been excluded despite wishing to participate; may be able to reach the system only through a distributor whose fees or revenue share make participation uneconomic for a low-revenue catalog; may be unaware that Content ID exists; or may have made a deliberate choice not to enforce. Only the first two of these turn a missing claim into an enforcement failure in the sense relevant to our analysis. The latter two

[37] See the YouTube Services Directory (formerly the Creator Services Directory), https://servicesdirectory.withyoutube.com. Providers offering Content ID administration are classified there as multi-channel networks or "à la carte" service providers.

[38] YouTube defines MCNs as third-party service providers that affiliate with multiple channels to offer services including digital rights management; see "Multi-channel network (MCN) overview," https://support.google.com/youtube/answer/2737059.

[39] On Universal's $775 million acquisition of Downtown Music Holdings (parent of CD Baby and Songtrust), see https://www.billboard.com/pro/virgin-music-group-acquires-downtown-775-million-universal/; on Sony's acquisition of AWAL, see https://www.billboard.com/articles/business/9519146/sony-music-acquires-kobalt-awal-neighbouring-rights/; and on the broader wave of major-label acquisitions of independent distributors, see https://www.billboard.com/pro/major-record-labels-2024-acquiring-indie-music-companies/ (all Billboard).

would instead reflect the system's limited scope. Because we observe a recording's presence or absence in the database but not the rights holder's intent, we cannot cleanly separate these cases at the level of an individual recording, and we therefore do not attribute the full participation gap to involuntary exclusion.

Two considerations nonetheless indicate that involuntary exclusion is real rather than hypothetical. First, as described above, access is conditional on either qualifying for a direct partnership—which YouTube limits to holders of a "substantial body of original material that is frequently uploaded"—or paying an intermediary, so that even an informed rights holder who wishes to participate may be unable to do so on terms that make economic sense for a modest catalog. Second, the U.S. Copyright Office's 2020 review of Section 512 records that rights holders themselves contest this exclusion. Commenters complained that the eligibility policy means "the little guy need not apply," argued that "every artist should be entitled to this service, to register their music once and for all," and objected that platforms "should not be allowed to continue to offer new solutions only to large or preferred rights holders" (U.S. Copyright Office, 2020). The Office itself characterized the deployed tools as "DMCA+ systems that are primarily open to larger content owners." These accounts establish that at least some rights holders who wish to participate are kept out, even though the precise share cannot be determined. We accordingly read the differential-enforcement estimates in Section 3.3.1 as reflecting a combination of genuine exclusion, prohibitive access costs, and—to an unknown degree—non-participation by choice or unawareness, rather than exclusion alone.

E.3 Distributor Schemes and Their Economics

Distributors price Content ID access in two ways that often combine. They charge an enrollment fee for placing a recording in the reference database, and they take a commission on the advertising revenue that Content ID subsequently collects. The arrangements are publicly documented on the distributors' own websites and are summarized in Table E.1.

DistroKid treats Content ID as a paid add-on, where enrollment costs $4.95 per single and $14.95 per album per year, recurring annually for as long as the recording remains enrolled, and the service retains

20% of the Content ID advertising revenue it collects.[40] TuneCore bundles Content ID with distribution at no separate per-release charge but likewise retains 20% of the resulting advertising revenue, paying the artist 80%.[41] CD Baby includes Content ID with its distribution at no separate enrollment fee but retains a substantially larger 30% of the revenue under its "Social Video Monetization" program, paying the artist the remaining 70%.[42] LANDR offers Content ID only to paying subscribers—its distribution plans run from $23.99 to $44.99 per year—and collects a flat 20% commission on Content ID royalties regardless of the subscriber's plan.[43] Thus, while the headline pricing varies—some distributors charge a separate annual fee for Content ID, others fold it into distribution, and others gate it behind a subscription—all of them retain a percentage of the monetization revenue, in addition to any commission they take on streaming income. For an independent artist, a single distributor already retains 20–30% of the Content ID revenue it collects—consistent with the up-to-30% figure cited in Section 2.2—typically on top of the recurring per-release fees or subscription charges described above. This commission is, moreover, charged only on the revenue that reaches the distributor after YouTube has taken its own share of the gross advertising revenue, so the artist's effective take of what advertisers pay is the product of the two cuts, well below the 70–80% split that the distributors headline.

Two implications follow for the interpretation of our results. First, access to automated enforcement for smaller rights holders is conditional on a paid commercial relationship with an intermediary, which means that coverage in Content ID is correlated with an artist's willingness and ability to pay rather than with the underlying legal entitlement. Second, because enrollment is opt-in and carries a cost—whether a per-release fee, a subscription, or a revenue share—catalogs that are old, low-revenue, or held by rights holders who never engaged a participating distributor are systematically less likely to appear in

[40] The 20% commission is stated in DistroKid, "What is YouTube Content ID?", https://support.distrokid.com/hc/en-us/articles/360013535314; the per-release add-on fee is listed in DistroKid, "What Are Album Extras?", https://support.distrokid.com/hc/en-us/articles/360013534274. Because the fee recurs per release, the cost of maintaining a catalog in Content ID grows over time even if the artist stops releasing new music.

[41] TuneCore, "Getting started with YouTube Content ID," https://support.tunecore.com/hc/en-us/articles/115006507207, which states that TuneCore sends the artist 80% of the revenue it collects (keeping a 20% commission) and charges no annual subscription fee.

[42] CD Baby, "Understanding your Social Video Monetization revenue," https://support.cdbaby.com/hc/en-us/articles/211095083. The 70/30 split applies to what CD Baby receives from the platform; YouTube takes its own cut before revenue reaches CD Baby. This is distinct from CD Baby's lower commission on streaming distribution.

[43] Plan prices from LANDR, "How much does it cost to distribute music with LANDR?", https://support.landr.com/hc/en-us/articles/31618416509975; the Content ID commission is described in LANDR, "How do I get paid for my music on YouTube?", https://support.landr.com/hc/en-us/articles/4821707948951.

Table E.1: Representative Distributor Schemes for YouTube Content ID Access

Distributor	Owner	Enrollment fee	Revenue share
DistroKid	Independent	Add-on: \$4.95/single or \$14.95/album per year	20%
TuneCore	Believe	Included (no separate fee)	20%
CD Baby	Universal (Downtown)	Included (no separate fee)	30%
LANDR	Independent	Paid subscription (\$23.99–44.99/year)	20%

Note: This table summarizes how four widely used music distributors price access to YouTube Content ID, based on the distributors' own published terms (see footnotes in Section E.3). The "Enrollment fee" is the charge for placing a recording in the reference database; the "Revenue share" is the commission the distributor takes on the advertising revenue Content ID collects, separate from any commission on streaming income. Figures are indicative of terms as of mid-2026 and are subject to change. The ownership column illustrates that intermediated access frequently routes through entities owned by major rights holders.

the reference database. Both mechanisms push in the same direction as the differential-enforcement and under-enforcement patterns we document in Section 5.

F Composition-Level Copyright Classification

This appendix lists every composer and composition in the *Classical US–EU* sample together with the US copyright status assigned by each of the three classification methods described in Appendix D: the publication-year rule ("Year"), IMSLP's jurisdiction-specific copyright note ("IMSLP"), and GPT-5 with web search ("GPT-5"). "Protected" denotes that the composition is still under US copyright, "Public domain" denotes that the US term has expired, and "—" denotes that the method could not assign a status (for the year and IMSLP methods, the required information was unavailable). Compositions are grouped by composer and ordered by the composer's year of death. The sample comprises 630 compositions by 22 composers; all remain protected under EU copyright law. These 630 are the subset, of a broader pool of 1,000 compositions sampled for the regional-compliance study and verified with the same three methods, for which we located uploaded videos and observed Content ID flags (and for which the claim was verified to reference the correct composition). The inter-method agreement reported in Appendix D is computed on these 630 analyzed compositions and can be reproduced directly from the classifications listed below.

Table F.1: US copyright classification of every composition in the Classical US–EU sample, by method.

Composition	Pub. year	Year	IMSLP	GPT-5
Reinhold Glière (1875–1956), Russian				
Ballade, Op.4	1903	Public domain	Public domain	Public domain
String Octet, Op.5	1903	Public domain	Public domain	Public domain
8 Pieces, Op.39	1909	Public domain	Public domain	Public domain
12 Esquisses, Op.47	1910	Public domain	Public domain	Public domain
12 Album Leaves, Op.51	1911	Public domain	Public domain	Public domain
12 Duos for 2 Violins, Op.49	1911	Public domain	Public domain	Public domain
Symphony No.2, Op.25	1912	Public domain	Public domain	Public domain
Symphony No.3 'Ilya Murometz', Op.42	1912	Public domain	Public domain	Public domain
The Sirens, Op.33	1912	Public domain	Public domain	Public domain
The Red Poppy, Op.70	1933	Protected	Protected	Protected
Concerto for Coloratura and Orchestra, Op.82	1944	Protected	Protected	Protected
The Bronze Horseman, Op.89	1950	Protected	Protected	Protected
Horn Concerto, Op.91	1951	Protected	Protected	Protected
25 Preludes, Op.30	—	—	Public domain	Public domain

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Alexander Goedicke (1877–1957), Russian				
Prelude, Op.20	1909	Public domain	Public domain	Public domain
2 Preludes and Fugues for Organ, Op.34	1926	Public domain	Public domain	Public domain
Etude concertante, Op.49	1946	Protected	Protected	Protected
Erich Wolfgang Korngold (1897–1957), Austrian				
Piano Trio, Op.1	1910	Public domain	Public domain	Public domain
Schauspiel-Ouverture, Op.4	1912	Public domain	Public domain	Public domain
Violin Sonata, Op.6	1913	Public domain	Public domain	Public domain
Sinfonietta, Op.5	1914	Public domain	Public domain	Public domain
6 Einfache Lieder, Op.9	1916	Public domain	Public domain	Public domain
String Sextet, Op.10	1917	Public domain	Public domain	Public domain
Piano Quintet, Op.15	1924	Public domain	Public domain	Public domain
3 Gesänge, Op.18	1925	Public domain	Public domain	Public domain
3 Lieder, Op.22	1930	Protected	Protected	Protected
Suite for Left Hand Piano and Strings, Op.23	1930	Protected	Protected	Protected
String Quartet No.2, Op.26	1937	Protected	Protected	Protected
The Adventures of Robin Hood	1938	Protected	Protected	Protected
The Sea Hawk	1940	Protected	Protected	Protected
4 Lieder nach Shakespeare, Op.31	1941	Protected	Protected	Protected
Tomorrow, Op.33	1942	Protected	Protected	Protected
Songs of the Clown, Op.29	1943	Protected	Protected	Protected
Violin Concerto, Op.35	1945	Protected	Protected	Protected
Straussiana	1954	Protected	Protected	Protected
5 Lieder, Op.38	1956	Protected	Protected	Protected
Symphony in F-sharp major, Op.40	1977	Protected	—	Protected
Baby-Serenade, Op.24	—	—	Public domain	Protected
String Quartet No.1, Op.16	—	—	Public domain	Public domain
String Quartet No.3, Op.34	—	—	Protected	Protected
Florent Schmitt (1870–1958), French				
Psaume XLVII, Op.38	1909	Public domain	Public domain	Public domain
La Tragédie de Salomé, Op.50	1912	Public domain	Public domain	Public domain
Lied et Scherzo, Op.54	1912	Public domain	Public domain	Public domain
Une semaine du petit elfe Ferme-l'Oeil, Op.58	1913	Public domain	Public domain	Public domain
Dionysiaques, Op.62	1917	Public domain	Public domain	Public domain

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Antoine et Cléopâtre, Suite No.1, Op.69a	1920	Public domain	Public domain	Public domain
Sonate libre, Op.68	1920	Public domain	Public domain	Public domain
Mirages, Op.70	1925	Public domain	Public domain	Public domain
In Memoriam, Op.72	1937	Protected	Protected	Protected
Saxophone Quartet, Op.102	1948	Protected	Protected	Protected
Symphony No.2, Op.137	1959	Protected	Protected	Protected
Ralph Vaughan Williams (1872–1958), English				
The House of Life	1903	Public domain	Public domain	Public domain
Fain Would I Change That Note	1907	Public domain	Public domain	Public domain
The Wasps	1909	Public domain	Public domain	Public domain
5 Mystical Songs	1911	Public domain	—	Public domain
Let all the World in Every Corner Sing	1911	Public domain	Public domain	Public domain
Fantasia on Christmas Carols	1912	Public domain	Public domain	Public domain
5 English Folk Songs	1913	Public domain	Public domain	Public domain
The Wasps, Aristophanic suite	1914	Public domain	Public domain	Public domain
A Sea Symphony (Symphony No.1)	1918	Public domain	Public domain	Public domain
On Christmas Night	1919	Public domain	Public domain	Protected
4 Hymns	1920	Public domain	Public domain	Public domain
Loch Lomond	1921	Public domain	Public domain	Public domain
Lord, Thou hast been our refuge	1921	Public domain	Public domain	Public domain
Phantasy Quintet	1921	Public domain	Public domain	Public domain
Suite of 6 Short Pieces	1921	Public domain	Public domain	Public domain
Let Us Now Praise Famous Men	1923	Public domain	Public domain	Public domain
String Quartet No.1	1923	Public domain	Public domain	Public domain
Toccata Marziale	1924	Public domain	Public domain	Public domain
6 Studies in English Folksong	1927	Public domain	Public domain	Public domain
Concerto Accademico	1927	Public domain	Public domain	Public domain
Flos Campi	1928	Public domain	Public domain	Public domain
Prelude and Fugue in C minor	1930	Protected	Protected	Protected
Sir John in Love	1930	Protected	Protected	Protected
Symphony No.4 in F minor	1935	Protected	Protected	Protected
Suite for Viola and Small Orchestra	1936	Protected	Protected	Protected
5 Variants of ‘Dives and Lazarus’	1940	Protected	Protected	Protected
Oboe Concerto	1947	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Symphony No.6 in E minor	1948	Protected	Protected	Protected
3 Shakespeare Songs	1951	Protected	Protected	Protected
Partita	1951	Protected	Protected	Protected
Sinfonia Antartica (Symphony No.7)	1953	Protected	Protected	Protected
Hodie	1954	Protected	Protected	Protected
Symphony No.8 in D minor	1956	Protected	Protected	Protected
Variations for Brass Band	1957	Protected	Protected	Protected
Symphony No.9 in E minor	1958	Protected	Protected	Protected
Dona Nobis Pacem	1969	Protected	Protected	Protected
In The Fen Country	1969	Protected	Protected	Protected
Flourish for Wind Band	1972	Protected	—	Protected
Piano Quintet in C minor	2002	Protected	—	Protected
2 Hymn-Tune Preludes	—	—	Protected	Protected
2 Poems by Seumas O'Sullivan	—	—	Public domain	Public domain
A London Symphony (Symphony No.2)	—	—	Public domain	Public domain
Along the Field	—	—	Public domain	Protected
Benedicite	—	—	Public domain	Protected
Concerto Grosso	—	—	Protected	Protected
Festival Te Deum	—	—	Protected	Protected
Hugh the Drover	—	—	Public domain	Protected
Piano Concerto in C major	—	—	Public domain	Protected
Riders to the Sea	—	—	Protected	Protected
Romance for Harmonica	—	—	Protected	Protected
Songs of Travel	—	—	Public domain	Protected
Tuba Concerto	—	—	—	Protected
Bohuslav Martinů (1890–1959), Czech				
String Quartet No.2, H.150	1927	Public domain	Public domain	Public domain
Duo for Violin and Cello, H.157	1928	Public domain	Public domain	Public domain
Film en miniature, H.148	1929	Public domain	Public domain	Public domain
8 Preludes, H.181	1930	Protected	Protected	Protected
La revue de cuisine, H.161	1930	Protected	Protected	Protected
String Quintet, H.164	1930	Protected	Protected	Protected
Violin Sonata No.1, H.182	1930	Protected	Protected	Protected
Nocturnes, H.189	1931	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
String Quartet No.3, H.183	1931	Protected	Protected	Protected
7 Arabesques, H.201/201a	1932	Protected	Protected	Protected
7 Études rhythmiques, H.202	1932	Protected	Protected	Protected
Concerto for String Quartet and Orchestra, H.207	1932	Protected	Protected	Protected
Rhythmische Etüden, H.202	1932	Protected	Protected	Protected
Impromptu, H.166	1934	Protected	Protected	Protected
Intermezzo, H.261	1937	Protected	Protected	Protected
Double Concerto for 2 String Orchestras, Piano and Timpani, H.271	1941	Protected	Protected	Protected
Etudes and Polkas, H.308	1946	Protected	Protected	Protected
String Sextet, H.224	1948	Protected	Protected	Protected
3 Madrigals, H.313	1949	Protected	Protected	Protected
Cello Sonata No.1, H.277	1949	Protected	Protected	Protected
Concertino for Piano Trio and String Orchestra, H.232	1949	Protected	Protected	Protected
Serenade No.1, H.217	1949	Protected	Protected	Protected
Serenade No.3, H.218	1949	Protected	Protected	Protected
Variations on a Theme of Rossini, H.290	1949	Protected	Protected	Protected
Violin Concerto No.2, H.293	1949	Protected	Protected	Protected
Les rondes, H.200	1950	Protected	Public domain	Protected
Symphony No.4, H.305	1950	Protected	Protected	Protected
Trio for Flute, Cello and Piano, H.300	1950	Protected	Protected	Protected
Violin Sonata No.3, H.303	1950	Protected	Protected	Protected
Les madrigaux, H.266	1951	Protected	Protected	Protected
Piano Quartet No.1, H.287	1951	Protected	Protected	Protected
String Trio No.2, H.238	1951	Protected	Protected	Protected
3 Písně posvátné, H.339	1953	Protected	Protected	Protected
Concerto for 2 Pianos and Orchestra, H.292	1953	Protected	Protected	Protected
Sinfonietta giocosa, H.282	1953	Protected	Protected	Protected
Concerto da camera, H.285	1955	Protected	Protected	Protected
Cello Sonata No.3, H.340	1957	Protected	Public domain	Public domain
Clarinet Sonatina, H.356	1957	Protected	Protected	Protected
Piano Quintet No.2, H.298	1957	Protected	Public domain	Public domain
Symphony No.6, H.343	1957	Protected	Protected	Protected
Trumpet Sonatina, H.357	1957	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Viola Sonata, H.355	1958	Protected	Public domain	Public domain
Nonet No.2, H.374	1959	Protected	Protected	Protected
Promenades, H.274	1959	Protected	Protected	Protected
Sonata for Flute, Violin and Piano, H.254	1959	Protected	Protected	Protected
Variations sur un thème Slovaque, H.378	1960	Protected	Public domain	Public domain
Concerto for Flute, Violin, and Orchestra, H.252	1961	Protected	Protected	Protected
Piano Trio No.2, H.327	1961	Protected	Protected	Protected
2 Pieces pour clavecin, H.244	1962	Protected	Protected	Protected
Serenade, H.334	1962	Protected	Protected	Protected
Bergerettes, H.275	1963	Protected	Protected	Protected
Piano Trio No.3, H.332	1963	Protected	Protected	Protected
Duo for Violin and Cello, H.371	1964	Protected	Protected	Protected
Overture, H.345	1965	Protected	Protected	Protected
Musique de chambre No.1, H.376	1966	Protected	Protected	Protected
Sextet, H.174	1966	Protected	Protected	Protected
Fantasia for Theremin, Oboe, String Quartet and Piano, H.301	1973	Protected	—	Protected
Motýli a rajky, H.127	1973	Protected	—	Protected
String Quartet No.1, H.117	1973	Protected	—	Protected
Cello Concerto No.2, H.304	1978	Protected	—	Protected
Concerto for 2 Violins and Orchestra, H.329	1980	Protected	—	Protected
Sonata da camera, H.283	1980	Protected	—	Protected
Thunderbolt P-47, H.309	1989	Protected	—	Protected
5 Madrigal Stanzas, H.297	—	—	Protected	Protected
Concertino, H.143	—	—	Public domain	Protected
Flute Sonata, H.306	—	—	Protected	Protected
Harpsichord Concerto, H.246	—	—	Protected	Protected
Oboe Concerto, H.353	—	—	Protected	Protected
Serenade No.2, H.216	—	—	Protected	Protected
Serenade No.4, H.215	—	—	Protected	Protected
Sinfonietta La Jolla, H.328	—	—	Protected	Protected
Ernest Bloch (1880–1959), American (Swiss-born)				
Schelomo, B.39	1916	Public domain	Public domain	Public domain
3 Poèmes juifs, B.36	1918	Public domain	Public domain	Public domain

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Psalm 22, B.38	1919	Public domain	Public domain	Public domain
Violin Sonata No.1, B.42	1922	Public domain	Public domain	Public domain
Piano Quintet No.1, B.43	1924	Public domain	Public domain	Public domain
Concerto Grosso No.1, B.59	1925	Public domain	Public domain	Public domain
From Jewish Life, B.54	1925	Public domain	Public domain	Public domain
Méditation hébraïque, B.55	1925	Public domain	Public domain	Public domain
Nuit exotique, B.57	1925	Public domain	Public domain	Public domain
Paysages, B.62	1925	Public domain	Public domain	Public domain
Violin Sonata No.2, B.58	1925	Public domain	Public domain	Public domain
Prélude, B.63	1929	Public domain	Public domain	Public domain
Avodath Hakodesh, B.68	1934	Protected	Protected	Protected
Visions et prophéties, B.70a	1936	Protected	Protected	Protected
Voice in the Wilderness, B.70	1936	Protected	Protected	Protected
String Quartet No.2, B.76	1947	Protected	Protected	Protected
6 Preludes, B.79	1948	Protected	Public domain	Public domain
Concertino for Flute and Viola, B.80	1951	Protected	Protected	Protected
Meditation and Processional, B.82	1954	Protected	Protected	Protected
Suite modale, B.95	1958	Protected	Protected	Protected
Suite No.1 for Solo Violin, B.99	1959	Protected	Protected	Protected
Suite No.2 for Solo Violin, B.100	1959	Protected	Protected	Protected
Baal Shem, B.47	—	—	Public domain	Public domain
Heitor Villa-Lobos (1887–1959), Brazilian				
Pequena suíte, W064	1913	Public domain	Public domain	Public domain
Suíte infantil No.2, W067	1913	Public domain	Public domain	Public domain
O canto do cisne negro, W122, 123	1917	Public domain	Public domain	Public domain
A lenda do caboclo, W166, 188	1920	Public domain	—	Public domain
A prole do bebê No.1, W140	1920	Public domain	Public domain	Public domain
Carnaval das crianças, W157	1920	Public domain	Public domain	Public domain
Chôros No.1, W161	1920	Public domain	Public domain	Public domain
Danças características africanas, W085	1920	Public domain	Public domain	Public domain
Suíte floral, W117	1920	Public domain	Public domain	Public domain
Simples coletânea, W134	1922	Public domain	Public domain	Public domain
Histórias da carochinha, W148	1923	Public domain	Public domain	Public domain
Chôros No.5, W207	1925	Public domain	Public domain	Public domain

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Cirandinhas, W210	1925	Public domain	Public domain	Public domain
Cirandas, W220	1926	Public domain	Public domain	Public domain
Saudades das selvas brasileiras, W226	1927	Public domain	Public domain	Public domain
Chôros No.3, W206	1928	Public domain	Public domain	Public domain
Chôros No.4, W218	1928	Public domain	Public domain	Public domain
Chôros No.7, W199	1928	Public domain	Public domain	Public domain
3 Poemas indígenas, W223, 224	1929	Public domain	Public domain	Public domain
Sonata fantasia No.1, W051	1929	Public domain	Public domain	Public domain
2 Chôros bis, W227	1930	Protected	Protected	Protected
Momoprecoce, W240, 259	1934	Protected	Protected	Protected
Ciclo brasileiro, W374	1940	Protected	Protected	Protected
As três Marias, W411	1941	Protected	Protected	Protected
Poema singelo, W434	1943	Protected	Protected	Protected
Bachianas brasileiras No.6, W392	1946	Protected	Protected	Protected
Bachianas brasileiras No.5, W389-391	1947	Protected	Protected	Protected
Bachianas brasileiras No.1, W246	1948	Protected	Protected	Protected
Caixinha de música quebrada, W256	1948	Protected	Protected	Protected
String Quartet No.5, W263	1948	Protected	Protected	Protected
String Quartet No.6, W399	1948	Protected	Protected	Protected
Uirapuru, W133	1948	Protected	Protected	Protected
String Quartet No.8, W446	1949	Protected	Protected	Protected
12 Estudos, W235	1952	Protected	Protected	Protected
Assobio a Jato, W493	1953	Protected	Protected	Protected
Bachianas brasileiras No.3, W388	1953	Protected	Protected	Protected
Quinteto em forma de chôros, W231	1953	Protected	Protected	Protected
String Quartet No.1, W099	1953	Protected	Protected	Protected
Preludes, W419	1954	Protected	Protected	Protected
Chôros No.6, W219	1955	Protected	Protected	Protected
Hommage à Chopin, W474	1955	Protected	Protected	Protected
Suite populaire brésilienne, W020	1955	Protected	Protected	Protected
String Quartet No.4, W129	1956	Protected	Protected	Protected
String Quartet No.7, W435	1956	Protected	Protected	Protected
String Trio, W460	1956	Protected	Protected	Protected
New York Skyline Melody, W407, 408	1957	Protected	Protected	Protected
Sexteto místico, W131	1957	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
String Quartet No.9, W457	1957	Protected	Protected	Protected
Bendita sabedoria, W543	1958	Protected	Protected	Protected
Duo for Oboe and Bassoon, W535	1958	Protected	Protected	Protected
Ciranda das sete notas, W325	1961	Protected	Protected	Protected
Fantasia para saxophone, W490	1963	Protected	Protected	Protected
String Quartet No.11, W481	1966	Protected	Protected	Protected
Bachianas brasileiras No.9, W449	1969	Protected	Protected	Protected
Distribuição de flores, W381, 575	1970	Protected	Protected	Protected
Quinteto instrumental, W538	1970	Protected	Protected	Protected
Valsa da dor, W316	1972	Protected	—	Protected
String Quartet No.17, W537	1977	Protected	—	Protected
Bachianas brasileiras No.7, W432	1978	Protected	—	Protected
Piano Concerto No.4, W505	1979	Protected	—	Protected
String Quartet No.16, W526	1981	Protected	—	Protected
Piano Concerto No.1, W453	1984	Protected	—	Protected
Chôros No.9, W232	1987	Protected	—	Protected
Bachianas brasileiras No.4, W264, 424	—	—	Protected	Protected
Guitar Concerto, W501, 502	—	—	Protected	Protected
Improviso No.7, W096	—	—	Public domain	Public domain
Modinhas e canções, W441, 563	—	—	—	Protected
Simples, W040	—	—	Public domain	Protected
String Quartet No.15, W523	—	—	—	Protected
Ernő Dohnányi (1877–1960), Hungarian				
Piano Quintet No.1, Op.1	1902	Public domain	Public domain	Public domain
4 Rhapsodien, Op.11	1904	Public domain	Public domain	Public domain
Serenade, Op.10	1904	Public domain	Public domain	Public domain
Waltz, Op.3	1906	Public domain	Public domain	Public domain
String Quartet No.2, Op.15	1907	Public domain	Public domain	Public domain
Humoresken in Form einer Suite, Op.17	1908	Public domain	Public domain	Public domain
Violin Sonata, Op.21	1913	Public domain	Public domain	Public domain
Suite in the Olden Style, Op.24	1914	Public domain	Public domain	Public domain
Violin Concerto No.1, Op.27	1920	Public domain	Public domain	Public domain
Pastorale on a Hungarian Christmas Song	1922	Public domain	Public domain	Public domain
Variationen über ein Kinderlied, Op.25	1922	Public domain	Public domain	Public domain

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Ruralia hungarica, Op.32a	1925	Public domain	Public domain	Public domain
Ruralia hungarica, Op.32b	1925	Public domain	Public domain	Public domain
Ruralia hungarica, Op.32c	1927	Public domain	Public domain	Public domain
String Quartet No.3, Op.33	1927	Public domain	Public domain	Public domain
Symphonic Minutes, Op.36	1935	Protected	Protected	Protected
6 Pieces for Piano, Op.41	1945	Protected	Protected	Protected
Piano Concerto No.2, Op.42	1948	Protected	Protected	Protected
Sextet, Op.37	1948	Protected	Protected	Protected
American Rhapsody, Op.47	1954	Protected	Public domain	Public domain
Aria for Flute and Piano, Op.48 No.1	1962	Protected	Protected	Protected
Passacaglia for Flute Solo, Op.48 No.2	1963	Protected	Protected	Protected
String Sextet	2005	Protected	—	Protected
Jacques Ibert (1890–1962), French				
6 Pièces pour harpe à pédales	1917	Public domain	Public domain	Public domain
Ballade	1917	Public domain	Public domain	Public domain
En barque, le soir...	1917	Public domain	Public domain	Public domain
Le vent dans les ruines	1918	Public domain	Public domain	Public domain
Histoires	1922	Public domain	Public domain	Public domain
2 Mouvements	1923	Public domain	Public domain	Public domain
Escales	1924	Public domain	Public domain	Public domain
Les rencontres	1924	Public domain	Public domain	Public domain
Jeux	1925	Public domain	Public domain	Public domain
Cello Concerto	1926	Public domain	Public domain	Public domain
3 Pièces brèves	1930	Protected	Protected	Protected
Suite symphonique	1932	Protected	Protected	Protected
Flute Concerto	1934	Protected	Protected	Protected
Entr'acte	1937	Protected	Protected	Protected
Petite suite en 15 images	1944	Protected	Protected	Protected
5 Pièces en trio	1947	Protected	Protected	Protected
2 Interludes	1949	Protected	Protected	Protected
Caprilena	1951	Protected	Protected	Protected
Ghirlarzana	1951	Protected	Protected	Protected
Impromptu	1951	Protected	Protected	Protected
Louisville-concert	1954	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Hommage à Mozart	1957	Protected	Protected	Protected
Toccata sur le nom d'Albert Roussel	1961	Protected	Protected	Public domain
John Ireland (1879–1962), English				
Berceuse	1903	Public domain	Public domain	Public domain
Phantasie	1908	Public domain	Public domain	Public domain
Cavatina	1911	Public domain	Public domain	Public domain
Sursum Corda	1911	Public domain	Public domain	Public domain
Decorations	1915	Public domain	Public domain	Public domain
Sea Fever	1915	Public domain	Public domain	Public domain
Preludes	1917	Public domain	Public domain	Public domain
Rhapsody	1917	Public domain	Public domain	Public domain
Earth's Call	1918	Public domain	Public domain	Public domain
If There were Dreams to Sell	1918	Public domain	Public domain	Public domain
Spring Sorrow	1918	Public domain	Public domain	Public domain
The Forgotten Rite	1918	Public domain	Public domain	Public domain
The Towing Path	1919	Public domain	Public domain	Public domain
Piano Sonata	1920	Public domain	Public domain	Public domain
2 Piano Pieces	1921	Public domain	Public domain	Protected
Equinox	1923	Public domain	Public domain	Public domain
Cello Sonata	1924	Public domain	Public domain	Public domain
2 Pieces for Pianoforte	1925	Public domain	Public domain	Public domain
Prelude	1925	Public domain	Public domain	Public domain
5 Poems by Thomas Hardy	1927	Public domain	Public domain	Public domain
A Downland Suite	1932	Protected	Public domain	Protected
Piano Concerto	1932	Protected	Protected	Protected
Songs Sacred and Profane	1934	Protected	Protected	Protected
Greenways	1938	Protected	Protected	Protected
Legend	1938	Protected	Protected	Protected
3 Pastels	1941	Protected	Protected	Protected
Sarnia	1941	Protected	Protected	Protected
Epic March	1942	Protected	Protected	Protected
Fantasy-Sonata	1945	Protected	Protected	Protected
Columbine	1951	Protected	Protected	Protected
London Pieces	—	—	Public domain	Public domain

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Francis Poulenc (1899–1963), French				
Mouvements perpétuels, FP 14a	1919	Public domain	Public domain	Public domain
Rapsodie nègre, FP 3	1919	Public domain	Public domain	Public domain
Sonata for 2 Clarinets, FP 7	1919	Public domain	Public domain	Public domain
Sonata for Piano Four Hands, FP 8	1919	Public domain	Public domain	Public domain
Cocardes, FP 16	1920	Public domain	Public domain	Public domain
Le bestiaire, FP 15	1920	Public domain	Public domain	Public domain
Valse for l'Album des Six, FP 17	1920	Public domain	Public domain	Public domain
Promenades, FP 24	1923	Public domain	Public domain	Public domain
Sonata for Clarinet and Bassoon, FP 32a	1924	Public domain	Public domain	Public domain
Sonata for Horn, Trumpet and Trombone, FP 33a	1924	Public domain	Public domain	Public domain
5 Poèmes de Pierre Ronsard, FP 38a	1925	Public domain	Public domain	Public domain
Napoli, FP 40	1926	Public domain	Public domain	Public domain
Trio for Oboe, Bassoon and Piano, FP 43	1926	Public domain	Public domain	Public domain
Pastourelle, FP 45	1928	Public domain	Public domain	Public domain
Concert champêtre, FP 49	1929	Public domain	Public domain	Public domain
Pièce brève sur le nom d'Albert Roussel, FP 50	1929	Public domain	Public domain	Public domain
3 Pièces, FP 48	1931	Protected	Protected	Protected
3 Poèmes de Louise Lalanne, FP 57	1931	Protected	Protected	Protected
4 Poèmes de Guillaume Apollinaire, FP 58	1931	Protected	Protected	Protected
Aubade, FP 51	1931	Protected	Protected	Protected
5 Poèmes de Max Jacob, FP 59	1932	Protected	Protected	Protected
Le bal masqué, FP 60	1932	Protected	Protected	Protected
Concerto for 2 Pianos, FP 61	1933	Protected	Protected	Protected
Feuillets d'album, FP 68	1933	Protected	Protected	Protected
Valse-improvisation sur le nom de BACH, FP 62	1933	Protected	Protected	Protected
Villageoises, FP 65	1933	Protected	Protected	Protected
Presto in B-flat major, FP 70	1934	Protected	Protected	Protected
Badinage, FP 73	1935	Protected	Protected	Protected
À sa guitare, FP 79a	1935	Protected	Protected	Protected
7 Chansons, FP 81	1936	Protected	Protected	Protected
Les soirées de Nazelles, FP 84	1937	Protected	Protected	Protected
Tel jour, telle nuit, FP 86	1937	Protected	Protected	Protected
2 Marches et un intermède, FP 88	1938	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
3 Poèmes de Louise de Vilmorin, FP 91	1938	Protected	Protected	Protected
2 Poèmes de Guillaume Apollinaire, FP 94	1939	Protected	Protected	Protected
4 Motets pour un temps de pénitence, FP 97	1939	Protected	Protected	Protected
La grenouillère, FP 96	1939	Protected	Protected	Protected
Le portrait, FP 92	1939	Protected	Protected	Protected
Organ Concerto, FP 93	1939	Protected	Protected	Protected
Priez pour paix, FP 95	1939	Protected	Protected	Protected
Fiançailles pour rire, FP 101	1940	Protected	Protected	Protected
Française d'après Claude Gervaise, FP 103	1940	Protected	Protected	Protected
Banalités, FP 107	1941	Protected	Protected	Protected
Ce doux petit visage, FP 99	1941	Protected	Protected	Protected
2 Poèmes de Louis Aragon, FP 122	1944	Protected	Protected	Protected
Métamorphoses, FP 121	1944	Protected	Protected	Protected
Violin Sonata, FP 119	1944	Protected	Protected	Protected
2 Mélodies de Guillaume Apollinaire, FP 127	1945	Protected	Protected	Protected
Figure humaine, FP 120	1945	Protected	Protected	Protected
Léocadia, FP 106	1945	Protected	Protected	Protected
Mélancolie, FP 105	1945	Protected	Protected	Protected
Sextet, FP 100	1945	Protected	Protected	Protected
Un soir de neige, FP 126	1945	Protected	Protected	Protected
Le disparu, FP 134	1947	Protected	Protected	Protected
8 Chansons françaises, FP 130	1948	Protected	Protected	Protected
4 Petites prières de Saint-François d'Assise, FP 142	1949	Protected	Protected	Protected
Cello Sonata, FP 143	1949	Protected	Protected	Protected
Histoire de Babar le petit éléphant, FP 129	1949	Protected	Protected	Protected
Hymne, FP 144	1949	Protected	Protected	Protected
Mazurka, FP 145	1949	Protected	Protected	Protected
La fraîcheur et le feu, FP 147	1951	Protected	Protected	Protected
Sinfonietta, FP 141	1951	Protected	Protected	Protected
Stabat Mater, FP 148	1951	Protected	Protected	Protected
4 Motets pour le temps de Noël, FP 152	1952	Protected	Protected	Protected
L'embarquement pour Cythère, FP 150	1952	Protected	Protected	Protected
Thème varié, FP 151	1952	Protected	Protected	Protected
Sonata for 2 Pianos, FP 156	1954	Protected	Protected	Protected
Dernier poème, FP 163	1957	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Le travail du peintre, FP 161	1957	Protected	Protected	Protected
2 Improvisations, FP 170	1958	Protected	Protected	Protected
Flute Sonata, FP 164	1958	Protected	Protected	Protected
Laudes de Saint-Antoine de Padoue, FP 172	1959	Protected	Protected	Protected
Elegy, FP 175	1960	Protected	Protected	Protected
Gloria, FP 177	1960	Protected	Protected	Protected
La courte paille, FP 178	1960	Protected	Protected	Protected
La Dame de Monte-Carlo, FP 180	1961	Protected	Protected	Protected
Sarabande pour guitare, FP 179	1961	Protected	Protected	Protected
7 Répons des ténèbres, FP 181	1962	Protected	Protected	Protected
Oboe Sonata, FP 185	1963	Protected	Protected	Protected
10 Improvisations, FP 63	—	—	Protected	Protected
3 Intermezzos, FP 71/118	—	—	Protected	Protected
3 Novelettes, FP 47/173	—	—	Protected	Protected
5 Impromptus, FP 21	—	—	Public domain	Public domain
8 Chansons gaillardes, FP 42	—	—	Public domain	Public domain
Airs chantés, FP 46	—	—	Public domain	Protected
Clarinet Sonata, FP 184	—	—	Protected	Protected
Le gendarme incompris (suite), FP 20b	—	—	Public domain	Public domain
Les animaux modèles, FP 111	—	—	Protected	Protected
Les mamelles de Tirésias, FP 125	—	—	Protected	Protected
Paul Hindemith (1895–1963), German				
3 Pieces for Cello and Piano, Op.8	1917	Public domain	Public domain	Public domain
String Quartet No.2, Op.10	1919	Public domain	Public domain	Public domain
Violin Sonata, Op.11 No.2	1920	Public domain	Public domain	Public domain
Nusch-Nuschi Tänze	1921	Public domain	Public domain	Public domain
Tanzstücke, Op.19	1921	Public domain	Public domain	Public domain
Violin Sonata, Op.11 No.1	1921	Public domain	Public domain	Public domain
1922, Op.26	1922	Public domain	Public domain	Public domain
Die junge Magd, Op.23b	1922	Public domain	Public domain	Public domain
Kammermusik No.1, Op.24 No.1	1922	Public domain	Public domain	Public domain
Kleine Kammermusik, Op.24 No.2	1922	Public domain	Public domain	Public domain
Tuttifäntchen	1922	Public domain	Public domain	Public domain
Viola Sonata, Op.11 No.4	1922	Public domain	Public domain	Public domain

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Cello Sonata, Op.25 No.3	1923	Public domain	Public domain	Public domain
String Quartet No.4, Op.22	1923	Public domain	Public domain	Public domain
Der Dämon, Op.28	1924	Public domain	Public domain	Public domain
String Quartet No.5, Op.32	1924	Public domain	Public domain	Public domain
String Trio No.1, Op.34	1924	Public domain	Public domain	Public domain
Violin Sonata, Op.31 No.1	1924	Public domain	Public domain	Public domain
Violin Sonata, Op.31 No.2	1924	Public domain	Public domain	Public domain
Kammermusik No.4, Op.36 No.3	1925	Public domain	Public domain	Public domain
Konzertmusik für Blasorchester, Op.41	1927	Public domain	Public domain	Public domain
Schulwerk für Instrumental-Zusammenspiel, Op.44	1927	Public domain	Public domain	Public domain
Kammermusik No.7, Op.46 No.2	1928	Public domain	Public domain	Public domain
Kleine Sonate, Op.25 No.2	1929	Public domain	Public domain	Public domain
Trio for Viola, Heckelphone and Piano, Op.47	1929	Public domain	Public domain	Public domain
Konzertmusik für Klavier, Blechbläser und Harfen, Op.49	1930	Protected	Protected	Protected
Konzertmusik für Streichorchester und Blechbläser, Op.50	1931	Protected	Protected	Protected
Plöner Musiktag	1932	Protected	Protected	Protected
Anekdoten für Radio	1934	Protected	Protected	Protected
Mathis der Maler Symphony	1934	Protected	Protected	Protected
Violin Sonata in E major	1935	Protected	Protected	Protected
Der Schwanendreher	1936	Protected	Protected	Protected
Piano Sonata No.2	1936	Protected	Protected	Protected
Piano Sonata No.3	1936	Protected	Protected	Protected
Trauermusik	1936	Protected	Protected	Protected
Flute Sonata	1937	Protected	Protected	Protected
Organ Sonata No.1	1937	Protected	Protected	Protected
Organ Sonata No.2	1937	Protected	Protected	Protected
3 Leichte Stücke	1938	Protected	Protected	Protected
Symphonic Dances	1938	Protected	Protected	Protected
Bassoon Sonata	1939	Protected	Protected	Protected
Clarinet Quartet	1939	Protected	Protected	Protected
Oboe Sonata	1939	Protected	Protected	Protected
Sonata for Piano 4 Hands	1939	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Violin Concerto	1939	Protected	Protected	Protected
Cello Concerto	1940	Protected	Protected	Protected
Clarinet Sonata	1940	Protected	Protected	Protected
Harp Sonata	1940	Protected	Protected	Protected
Nobilissima Visione, Konzert-Suite	1940	Protected	Protected	Protected
Viola Sonata, IPH 172	1940	Protected	Protected	Protected
Violin Sonata in C major	1940	Protected	Protected	Protected
English Horn Sonata	1942	Protected	Protected	Protected
6 Chansons	1943	Protected	Protected	Protected
Ludus Tonalis	1943	Protected	Protected	Protected
Amor and Psyche	1944	Protected	Protected	Protected
Apparebit repentina dies	1947	Protected	Protected	Protected
The Four Temperaments	1947	Protected	Protected	Protected
Cello Sonata in E major	1949	Protected	Protected	Protected
String Quartet No.7	1949	Protected	Protected	Protected
Concerto for Woodwinds, Harp and Orchestra	1950	Protected	Protected	Protected
Double Bass Sonata	1950	Protected	Protected	Protected
Symphony in B-flat	1951	Protected	Protected	Protected
Cum natus esset	1952	Protected	Protected	Protected
Des Todes Tod, Op.23a	1953	Protected	Protected	Protected
Sonata for 4 Horns	1953	Protected	Protected	Protected
Alto Horn Sonata	1956	Protected	Protected	Protected
Tuba Sonata	1957	Protected	Protected	Protected
8 Pieces for Solo Flute	1958	Protected	Protected	Protected
Octet	1958	Protected	Protected	Protected
Pittsburgh Symphony	1959	Protected	Protected	Protected
Organ Concerto	1964	Protected	Protected	Protected
Plöner Musiktag Suite	1969	Protected	Protected	Protected
Viola Sonata, Op.25 No.4	1976	Protected	—	Protected
Minimax	1978	Protected	—	Protected
Melancholie, Op.13	1994	Protected	—	Protected
9 English Songs	—	—	Protected	Protected
Cardillac, Op.39	—	—	Public domain	Public domain
Cello Sonata, Op.11 No.3	—	—	Public domain	Public domain
Clarinet Quintet, Op.30	—	—	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Concerto for Trumpet, Bassoon and Strings	—	—	Protected	Protected
Das Marienleben, Op.27	—	—	Public domain	Protected
Hin und zurück, Op.45a	—	—	Public domain	Public domain
Kammermusik No.3, Op.36 No.2	—	—	Public domain	Public domain
Kammermusik No.5, Op.36 No.4	—	—	Public domain	Public domain
Kammermusik No.6, Op.46 No.1	—	—	Protected	Protected
Mathis der Maler	—	—	Protected	Protected
Piano Sonata No.1	—	—	Protected	Protected
Septet	—	—	Protected	Protected
Tuttifäntchen Suite	—	—	Public domain	Protected
When Lilacs last in the Door-yard Bloom'd	—	—	Protected	Protected
Ernst Toch (1887–1964), German				
Piano Concerto, Op.38	1926	Public domain	Public domain	Public domain
Violin Sonata, Op.44	1928	Public domain	Public domain	Public domain
Cello Sonata, Op.50	1929	Public domain	Public domain	Public domain
Pinocchio	1937	Protected	Protected	Protected
Henry Cowell (1897–1965), American				
Suite for Violin and Piano, HC 397	1926	Public domain	Public domain	Public domain
Fiddler's Jig, HC 771	1956	Protected	Protected	Protected
Persian Set, HC 838	1957	Protected	Protected	Protected
Air and Scherzo, HC 897	—	—	Protected	Protected
Zoltán Kodály (1882–1967), Hungarian				
Magyar népdalok	1906	Public domain	Public domain	Public domain
9 Pieces, Op.3	1910	Public domain	Public domain	Public domain
Adagio	1910	Public domain	Public domain	Public domain
7 Pieces, Op.11	1921	Public domain	Public domain	Public domain
Serenade, Op.12	1921	Public domain	Public domain	Public domain
Sonata for Solo Cello, Op.8	1921	Public domain	Public domain	Public domain
String Quartet No.2, Op.10	1921	Public domain	Public domain	Public domain
Cello Sonata, Op.4	1922	Public domain	Public domain	Public domain
Duo for Violin and Cello, Op.7	1922	Public domain	Public domain	Public domain
Méditation sur un motif de Claude Debussy	1925	Public domain	Public domain	Public domain
Háry János (suite)	1927	Public domain	Public domain	Public domain
2 Zoborvidéki népdal	1932	Protected	Protected	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Fölszállott a páva	1941	Protected	Protected	Protected
Kállai kettős, K.148	1952	Protected	Protected	Protected
Laudes organi	1966	Protected	Protected	Protected
Cello Sonatina	—	—	Public domain	Protected
Missa brevis	—	—	—	Protected
Healey Willan (1880–1968), English				
Magnificat and Nunc dimittis in B-flat major	1906	Public domain	Public domain	Public domain
Magnificat and Nunc dimittis in E-flat major	1912	Public domain	Public domain	Public domain
Introduction, Passacaglia and Fugue, B.149	1919	Public domain	Public domain	Public domain
Magnificat and Nunc dimittis, Tone I	1948	Protected	Protected	Protected
6 Chorale Preludes, Set I	1950	Protected	Protected	Protected
Mario Castelnuovo-Tedesco (1895–1968), Italian				
Cipressi, Op.17	1921	Public domain	Public domain	Public domain
La sirenetta e il pesce turchino, Op.18	1921	Public domain	Public domain	Public domain
Guitar Concerto No.1, Op.99	1954	Protected	Protected	Protected
Guitar Quintet, Op.143	1959	Protected	Protected	Protected
Sonatina Canonica, Op.196	1971	Protected	Protected	Protected
33 Shakespeare Songs, Op.24	—	—	Public domain	Public domain
Greeting Cards, Op.170	—	—	Protected	Protected
Sonatina, Op.205	—	—	Protected	Protected
Igor Stravinsky (1882–1971), Russian				
4 Etudes, K009	1910	Public domain	Public domain	Public domain
Pastorale, K006	1910	Public domain	Public domain	Public domain
The Firebird (ballet), K010	1910	Public domain	Public domain	Public domain
Petrushka, K012	1912	Public domain	Public domain	Public domain
3 Poems to Japanese Lyrics, K016	1913	Public domain	Public domain	Public domain
Le rossignol, K018	1914	Public domain	Public domain	Public domain
Symphony in E-flat major, K003	1914	Public domain	Public domain	Public domain
Souvenir d'une marche boche, KN14	1916	Public domain	Public domain	Public domain
3 Easy Pieces, K021	1917	Public domain	Public domain	Public domain
5 Easy Pieces, K025	1917	Public domain	Public domain	Public domain
Renard, K023	1917	Public domain	Public domain	Public domain
Piano-Rag-Music, K032	1920	Public domain	Public domain	Public domain
Le chant du rossignol, K026	1921	Public domain	Public domain	Public domain

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
3 Pieces for String Quartet, K019	1922	Public domain	Public domain	Public domain
Les cinq doigts, K037	1922	Public domain	Public domain	Public domain
Les noces, K040	1922	Public domain	Public domain	Public domain
Concertino, K035	1923	Public domain	Public domain	Public domain
Concerto for Piano and Wind Instruments, K042	1924	Public domain	Public domain	Public domain
Octet, K041	1924	Public domain	Public domain	Public domain
Pulcinella (suite), K034	1924	Public domain	Public domain	Public domain
Mavra, K039	1925	Public domain	Public domain	Public domain
Piano Sonata, K043	1925	Public domain	Public domain	Public domain
Serenade in A, K044	1926	Public domain	Public domain	Public domain
Suite No.1, K045	1926	Public domain	Public domain	Public domain
Oedipus Rex, K047	1927	Public domain	Public domain	Public domain
4 Russian Peasant Songs, K028	1930	Protected	Protected	Protected
Violin Concerto, K053	1931	Protected	Protected	Protected
Concerto for 2 Pianos, K058	1936	Protected	Protected	Protected
Danses concertantes, K063	1942	Protected	Protected	Protected
Circus Polka, K064	1943	Protected	Protected	Protected
Élegie, K072	1945	Protected	Protected	Protected
Ebony Concerto, K074	1946	Protected	Protected	Protected
Mass, K077	1948	Protected	Protected	Protected
Septet, K080	1953	Protected	Protected	Protected
In memoriam Dylan Thomas, K084	1954	Protected	Protected	Protected
Choral-Variationen, K087	1956	Protected	Protected	Protected
Monumentum pro Gesualdo di Venosa, K094	1960	Protected	Protected	Protected
Variations, K103	1965	Protected	Protected	Protected
The Owl and the Pussy-Cat, K107	1967	Protected	Protected	Protected
Piano Sonata, K001	1974	Protected	—	Protected
Apollon musagète, K048	—	—	Public domain	Public domain
Histoire du soldat, K029	—	—	Public domain	Public domain
Pater noster, K046	—	—	Public domain	Public domain
Suite No.2, K038	—	—	Public domain	Public domain
Symphonies of Wind Instruments, K036	—	—	Public domain	Protected
The Faun and the Shepherdess, K002	—	—	Public domain	Public domain
The Firebird (suite), K010	—	—	Public domain	Protected

continued on next page

(continued)

Composition	Pub. year	Year	IMSLP	GPT-5
Marcel Dupré (1886–1971), French				
3 Preludes and Fugues, Op.7	1920	Public domain	Public domain	Public domain
Variations sur un Noël, Op.20	1923	Public domain	Public domain	Public domain
Symphonie-Passion, Op.23	1925	Public domain	Public domain	Public domain
Lamento, Op.24	1928	Public domain	—	Public domain
7 Pieces, Op.27	1931	Protected	—	Protected
Le Chemin de la Croix, Op.29	1932	Protected	Protected	Protected
Offrande à la Vierge, Op.40	1945	Protected	Protected	Protected
8 Short Preludes on Gregorian Themes, Op.45	1948	Protected	Protected	Protected
Choral et fugue, Op.57	1962	Protected	Protected	Protected
Ernst Krenek (1900–1991), Austrian				
Kleine Suite for Piano, Op.13a	1922	Public domain	Public domain	Public domain
String Quartet No.3, Op.20	1924	Public domain	Public domain	Public domain
Orpheus und Eurydike, Op.21	1925	Public domain	Public domain	Public domain
5 Piano Pieces, Op.39	1926	Public domain	Public domain	Public domain
O Lacrymosa, Op.48	1926	Public domain	Public domain	Public domain
Das geheime Königreich, Op.50	1928	Public domain	Public domain	Public domain
Der Diktator, Op.49	1928	Public domain	Public domain	Public domain
Piano Sonata No.2, Op.59	1928	Public domain	Public domain	Public domain
Reisebuch aus den österreichischen Alpen, Op.62	1929	Public domain	Public domain	Public domain
Leo Ornstein (1893–2002), American (Russian-born)				
Cello Sonata No.1, SO 612	1918	Public domain	Public domain	Public domain
Piano Sonata No.4, SO 360	1924	Public domain	Public domain	Public domain
6 Preludes for Cello and Piano, SO 611	1975	Protected	—	Protected
A Long Remembered Sorrow, SO 102a	1990	Protected	—	Protected
Suicide in an Airplane, SO 6	1990	Protected	—	Protected

G A simple model of platform incentives

As a guiding framework for platform incentives and their welfare implications, we set up an intentionally stylized model. It abstracts from many details of copyright law and Content ID implementation and keeps three ingredients that map directly to our empirical setting. First, the platform earns advertising revenue from user-generated content. Second, rights holders obtain automated enforcement only if their works are represented in the platform's rights database. Third, representation in that database is tiered: large rights holders may obtain direct access, while smaller rights holders reach the database indirectly, through distributors, publishing administrators, multi-channel networks, or specialized rights-management firms.

The intermediary channel matters because non-participation by smaller rights holders need not mean that they were unaware of Content ID or formally excluded from it. It can instead be an equilibrium choice not to pay the fixed costs, revenue shares, and administrative costs of intermediated enforcement. The existence of intermediaries does not make access frictions disappear: intermediation may be costly, selective, or an imperfect substitute for direct access.

The question we ask is whether the resulting allocation of enforcement is socially efficient. We show that even when the matching technology is accurate, a platform choosing access and verification policies to maximize its own payoff can generate both under-enforcement and over-enforcement. Under-enforcement arises when valid copyrighted works are not effectively represented in the database. Over-enforcement arises when stale, invalid, or overbroad references generate claims against works that are in the public domain or otherwise not controlled by the claimant.

G.0.1 Setup

Consider a digital platform P that hosts user-generated content uploaded by users U. Uploaded videos may incorporate works owned by rights holders. There is a continuum of rights holders $i \in [0, 1]$ with heterogeneous view potential v_i, drawn from a distribution $F(v)$ with density f on $[0, \bar{v}]$. Let

$$\bar{M} = \int_0^{\bar{v}} v \, dF(v)$$

denote total view potential on the relevant scale. A fraction $\pi \in (0,1)$ of relevant uses involve copyrighted works; the remaining share $1-\pi$ involves works in the public domain or otherwise not subject to a valid claim by the party asserting rights.

The platform operates an automated matching system. Motivated by the evidence above, and to isolate the institutional source of errors, we assume the matching technology is accurate conditional on the relevant work being represented by a correct reference file. Enforcement errors therefore originate in access rules and database quality, not in acoustic matching.

Direct and intermediated access. The platform grants direct Content ID access only to rights holders above an access threshold θ, reflecting the practice of reserving direct access for rights holders with complex copyright management needs, such as major labels, large studios, collecting societies, and rights-management service providers. Rights holders with

$$v_i \geq \theta$$

obtain direct access and register reference files with the platform.

Rights holders below θ can still reach the database through an intermediary, which administers references, claims, monetization, whitelisting, reporting, and disputes on their behalf. This channel is costly. Let F_I denote the fixed cost of intermediated participation and $\rho \in [0,1)$ the intermediary's revenue share. A below-threshold rights holder participates through an intermediary when the expected enforcement revenue covers this cost,

$$(1-\rho)\,\alpha r\, v_i - F_I \geq 0,$$

where r is advertising revenue per view and α is the share of claimed revenue paid to the rights-holder side. This defines an intermediary participation threshold

$$z_I \equiv \frac{F_I}{(1-\rho)\alpha r}.$$

The expected revenue of a valid copyrighted enrollment enters this condition at its full value, because the database imperfections introduced below concern claims on public-domain material rather than the enforceability of a genuine copyrighted work.

We distinguish three sets:

$$\mathcal{A}_D(\theta) = \{i : v_i \geq \theta\}, \qquad \mathcal{A}_I(\theta, z_I) = \{i : z_I \leq v_i < \theta\},$$

$$\mathcal{A}_N(\theta, z_I) = \{i : v_i < \theta \text{ and } v_i < z_I\}.$$

The first group has direct access, the second is eligible for intermediated access, and the third has no effective automated access.

Intermediated access is not necessarily a perfect substitute for direct access. Let $\lambda \in [0, 1]$ denote the effectiveness of the intermediary channel: the fraction of eligible below-threshold rights holders who obtain effective coverage, or, equivalently, a quality-adjusted measure of intermediated access. At $\lambda = 1$ intermediaries fully cover rights holders between z_I and θ; at $\lambda < 1$ coverage remains partial.

Writing

$$M_D(\theta) = \int_{\mathcal{A}_D(\theta)} v \, dF(v), \qquad M_I(\theta, z_I) = \int_{\mathcal{A}_I(\theta, z_I)} v \, dF(v),$$

effective coverage of copyrighted works is

$$M_C(\theta, z_I, \lambda) = M_D(\theta) + \lambda M_I(\theta, z_I),$$

the copyrighted view potential left without automated enforcement is

$$M_N(\theta, z_I, \lambda) = \bar{M} - M_C(\theta, z_I, \lambda),$$

and the under-enforcement rate is

$$\psi(\theta, z_I, \lambda) = \frac{M_N(\theta, z_I, \lambda)}{\bar{M}}.$$

The direct-access threshold shapes coverage differently on the two sides of z_I. Away from the point $\theta = z_I$,

$$\frac{\partial M_C}{\partial \theta} = \begin{cases} -\theta f(\theta), & \theta < z_I, \\ -(1-\lambda)\,\theta f(\theta), & \theta > z_I. \end{cases}$$

When $\theta < z_I$, a marginal rise in θ pushes rights holders who cannot profitably use intermediaries out of coverage altogether. When $\theta > z_I$, it shifts marginal rights holders from direct to intermediated access, so coverage falls only through the incompleteness $1 - \lambda$ of that channel; if $\lambda = 1$ the shift is seamless and coverage is unchanged.

Public-domain exposure and database errors. Automated claims can also reach material in the public domain. Two conditions must line up: the public-domain material must be exposed to the database through some legitimate participant's reference data, and the resulting claim must be invalid. Staleness covers cases where a work remains claimed even though the relevant rights have expired in a jurisdiction, were never valid for that territory, or are not held by the claimant. Overbroad or opportunistic assertions of rights over uncontrolled material also fall here, as in the LatinAutorPerf pattern we document for classical public-domain recordings.

To connect the model to this empirical pattern, we track the public-domain view potential exposed to claims through the database,

$$B_E(\theta, z_I, \lambda) = \beta_D\, M_D(\theta) + \beta_I\, \lambda\, M_I(\theta, z_I),$$

where $\beta_D, \beta_I \geq 0$ convert the size of directly and intermediately represented catalogues into public-domain view potential exposed to possible claims, with $B_E \leq (1-\pi)\bar{M}$. The two channels need not expose the same amount of public-domain material: directly represented catalogues may create more opportunities for stale territorial claims, while intermediated catalogues may create more or fewer such opportunities depending on the intermediary's verification practices, so β_D and β_I can differ.

Let $x \geq 0$ denote the platform's verification effort. Let $\phi(x)$ be the probability that an erroneous reference survives verification, with

$$\phi'(x) < 0, \qquad \phi''(x) > 0,$$

and let verification cost $K(x)$ satisfy

$$K'(x) > 0, \qquad K''(x) > 0.$$

Let $s \geq 0$ denote the probability that an exposed public-domain item carries a stale or legally invalid claim, with $s + \phi(x) \leq 1$. The view potential of public-domain works subject to erroneous automated claims is

$$\Omega(\theta, z_I, \lambda, x) = B_E(\theta, z_I, \lambda)\,[\,s + \phi(x)\,], \qquad \Omega \leq (1-\pi)\bar{M},$$

and the over-enforcement rate is

$$\omega(\theta, z_I, \lambda, x) = \frac{\Omega(\theta, z_I, \lambda, x)}{\bar{M}}.$$

Legitimate enforcement thus operates on the copyrighted coverage M_C, while public-domain over-enforcement operates on the exposure B_E, with s and x governing how much of that exposure turns into invalid claims.

Monetization. Each view generates advertising revenue r. When a copyrighted work is correctly matched and enforced, the rights-holder side receives αr and the platform keeps $(1-\alpha)r$. When no match is made, the uploader keeps γr and the platform keeps $(1-\gamma)r$.

For public-domain works, an unclaimed use yields the uploader γr and the platform $(1-\gamma)r$. When a public-domain work is erroneously claimed, the uploader receives nothing from that use, the claimant receives αr, and the platform keeps $(1-\alpha)r$. Erroneous claims may also reduce viewer surplus when content is blocked, muted, demonetized, or disputed. Let η denote the probability that an erroneous claim causes such an availability loss, h_V the viewer-surplus loss per unit of affected public-domain view potential, and CS_0 baseline viewer surplus absent erroneous removals.

G.0.2 Payoffs

Rights holders. A rights holder with direct access receives

$$\Pi_i^D = \alpha r v_i,$$

one with effective intermediated access receives

$$\Pi_i^I = (1-\rho)\alpha r v_i - F_I,$$

and one without effective automated access receives $\Pi_i^N \approx 0$. The gap between Π_i^D and Π_i^I is the price of intermediation: it opens access to smaller rights holders, but only where expected enforcement revenue exceeds the fixed cost and revenue share. Observed non-participation can therefore be an equilibrium outcome even when the access regime is not socially efficient.

Uploaders and viewers. Uploaders gain from missed claims on copyrighted works, since revenue is not diverted to rights holders; they earn on unclaimed public-domain works; and they lose that revenue when a public-domain work is erroneously claimed. Viewers lose surplus when erroneous claims reduce availability. Uploader and viewer surplus is

$$\begin{aligned} S_{UV}(\theta, z_I, \lambda, x) = & \underbrace{\gamma r\, \pi M_N(\theta, z_I, \lambda)}_{\text{uploader revenue on missed copyrighted claims}} \\ & + \underbrace{\gamma r\big[(1-\pi)\bar{M} - \Omega(\theta, z_I, \lambda, x)\big]}_{\text{uploader revenue on unclaimed public-domain works}} \\ & + \underbrace{CS_0 - \eta h_V\, \Omega(\theta, z_I, \lambda, x)}_{\text{viewer surplus net of erroneous availability losses}} . \end{aligned}$$

The uploader loses the γr it would have earned on each erroneously claimed unit; the claimant's receipt αr is recorded separately below.

Intermediaries. Intermediary charges may be transfers or real costs, so we split them. Write $F_I = f_I + a_I$, where f_I is a fixed fee paid to the intermediary and a_I is a real participation or administrative cost borne by the rights holder, and let $k_I(v)$ be the intermediary's real servicing cost for a rights holder

of view potential v. Intermediary payoff is

$$\Pi_{\mathcal{I}}(\theta, z_I, \lambda) = \lambda\pi \int_{\mathcal{A}_I(\theta, z_I)} \big[\rho\alpha r v + f_I - k_I(v)\big]\, dF(v).$$

Treating all of F_I as a real cost sets $f_I = 0,\ a_I = F_I$; treating it all as a fee sets $a_I = 0,\ f_I = F_I$. The participation threshold depends on the rights holder's total private cost F_I in either case.

Illegitimate claimants. Erroneous public-domain claims generate receipts for parties who do not hold the relevant rights,

$$\Pi_Q(\theta, z_I, \lambda, x) = \alpha r\, \Omega(\theta, z_I, \lambda, x).$$

Whether these receipts enter welfare is a normative choice, so we attach a weight $\chi_Q \in [0, 1]$ to them: $\chi_Q = 0$ treats receipts from invalid claims as unentitled rents, $\chi_Q = 1$ treats them as ordinary private income, and intermediate values allow distributional weights.

Platform. The platform earns on correctly enforced copyrighted works, on copyrighted works that remain unmatched, on unclaimed public-domain works, and on erroneously claimed public-domain works, net of the costs of erroneous claims, verification, and access administration. Let D_P be the platform's private cost per unit of erroneous public-domain claims (disputes, reputation, compliance). Its payoff is

$$\begin{aligned}
\Pi_P(\theta, z_I, \lambda, x) = {} & \underbrace{(1-\alpha) r\, \pi M_C(\theta, z_I, \lambda)}_{\text{share on correctly enforced copyrighted works}} \\
& + \underbrace{(1-\gamma) r\, \pi M_N(\theta, z_I, \lambda)}_{\text{share on missed copyrighted claims}} \\
& + \underbrace{(1-\gamma) r \big[(1-\pi)\bar{M} - \Omega(\theta, z_I, \lambda, x)\big]}_{\text{revenue on unclaimed public-domain works}} \\
& + \underbrace{(1-\alpha) r\, \Omega(\theta, z_I, \lambda, x)}_{\text{revenue on erroneously claimed public-domain works}} \\
& - \underbrace{D_P\, \Omega(\theta, z_I, \lambda, x)}_{\text{private cost of erroneous claims}} - \underbrace{K(x)}_{\text{verification}} - \underbrace{C_D(\theta)}_{\text{direct-access administration}}.
\end{aligned}$$

The direct-access cost falls as the threshold rises, since admitting fewer rights holders reduces reference-file processing, claim management, and dispute handling:

$$C_D'(\theta) < 0.$$

Social welfare. Welfare sums the platform payoff, legitimate rights-holder payoffs, intermediary payoffs, uploader and viewer surplus, and the weighted receipts of illegitimate claimants. We also allow valid enforcement to carry an additional per-view social benefit $\delta \geq 0$, capturing deterrence, dynamic creator incentives, and legal compliance beyond the advertising transfers:

$$W(\theta, z_I, \lambda, x) = \Pi_P(\theta, z_I, \lambda, x) + \pi \int_{\mathcal{A}_D(\theta)} \Pi_i^D \, dF(v) + \lambda\pi \int_{\mathcal{A}_I(\theta, z_I)} \Pi_i^I \, dF(v)$$
$$+ \Pi_{\mathcal{I}}(\theta, z_I, \lambda) + S_{UV}(\theta, z_I, \lambda, x) + \chi_Q \, \Pi_Q(\theta, z_I, \lambda, x) + \delta\pi M_C(\theta, z_I, \lambda).$$

Advertising revenue is a transfer among the platform, uploaders, and rights holders, so those flows cancel in W; what remains are the real costs (verification, access administration, intermediary servicing, and real participation costs), the deterrence benefit δ, and the welfare treatment of claimant receipts through χ_Q.

A social planner chooses access and verification to maximize W. The planner can widen access by lowering θ, cutting the cost F_I of intermediated access, capping the revenue share ρ, raising intermediary effectiveness λ, or creating collective or certified access mechanisms. The platform instead chooses θ and x to maximize Π_P, taking most rights-holder, uploader, viewer, and intermediary effects as external to its own payoff.

G.0.3 Platform Choices and Comparative Statics

Proposition 1 (Platform vs. social-planner incentives)**.** *Suppose intermediation is not a perfect substitute for direct access at the relevant margin, so that either $\theta < z_I$ or $\lambda < 1$. Suppose also that the social marginal benefit of valid enforcement exceeds the platform's private marginal benefit at that margin,*

$$\delta v > r(\gamma - \alpha)v,$$

net of any common marginal administration costs. Then the platform chooses lower effective coverage of valid copyrighted works than the planner.

For verification, let the platform's private benefit from removing one unit of erroneous public-domain claims be

$$L_P = D_P - r(\gamma - \alpha),$$

and the corresponding social benefit be

$$L_W = D_P + \eta h_V + (1 - \chi_Q)\alpha r.$$

If $L_W > L_P$, the platform chooses lower verification effort than the planner.

Sketch. Access. Moving a valid copyrighted work of view potential v from the unmatched state to the claimed state changes the platform's advertising revenue by

$$(1 - \alpha)rv - (1 - \gamma)rv = r(\gamma - \alpha)v,$$

before administration costs. Its sign depends on α versus γ: if $\alpha > \gamma$ the platform earns less from a claimed than from an unmatched copyrighted work, while if $\gamma > \alpha$ it may gain from enforcement. The under-enforcement result therefore cannot rest on the transfer αr alone.

The planner's marginal benefit additionally includes the enforcement value δv. Because the advertising transfers cancel in W, the gap between platform and planner at the access margin is the comparison of $r(\gamma - \alpha)v$ with δv, net of common marginal costs. The role of the threshold follows the coverage derivative above: raising θ removes marginal rights holders from coverage when $\theta < z_I$, and shifts them to intermediated access, losing a fraction $1 - \lambda$ of coverage, when $\theta > z_I$. Hence, except in the full-substitution case $\lambda = 1$, $z_I < \theta$, if $\delta v > r(\gamma - \alpha)v$ net of common costs, the platform's access policy delivers lower effective coverage than the planner's.

Verification. Since $\Omega = B_E[s + \phi(x)]$ and $\phi'(x) < 0$, higher x reduces erroneous claims. Up to terms independent of x, the platform's payoff from public-domain claims is

$$\Omega(\theta, z_I, \lambda, x)\,[\,r(\gamma - \alpha) - D_P\,] - K(x),$$

so its private benefit from removing one unit of erroneous claims is $L_P = D_P - r(\gamma - \alpha)$. The social benefit adds the avoided viewer loss ηh_V and the value $(1-\chi_Q)\alpha r$ of not transferring αr to an illegitimate claimant, giving $L_W = D_P + \eta h_V + (1 - \chi_Q)\alpha r$. Whenever $L_W > L_P$, the marginal social return to verification exceeds the private return, so the planner sets higher effort. □

Corollary 1 (Equilibrium selection)**.** *Observed non-enforcement of smaller rights holders' works need not mean they were unaware of Content ID or formally denied access. It can arise because the expected private benefit of intermediated participation falls short of its cost,*

$$(1 - \rho)\alpha r v_i < F_I,$$

and such non-enforcement can still be inefficient when the social benefit of valid enforcement exceeds the private benefit that the rights holder and platform jointly internalize.

Corollary 2 (Incumbent advantage)**.** *When direct access tracks expected view potential, large rights holders lie above θ, while smaller ones participate through intermediaries or fall outside effective enforcement. If $\lambda < 1$ or F_I is high, intermediation does not erase this asymmetry, so the model predicts higher effective enforcement for major labels and large catalogues than for smaller or independent rights holders.*

Corollary 3 (Policy levers)**.** *Welfare improves under policies that reduce staleness s, reduce surviving errors $\phi(x)$, lower the intermediary threshold z_I, or raise intermediary effectiveness λ. These need not grant unrestricted Content ID access. They may instead lower F_I, cap excessive revenue shares ρ, require transparent and objective access criteria, certify low-cost rights-management providers, separate matching and reporting from blocking authority, or mandate periodic re-attestation of rights and jurisdiction-specific metadata. Exposure itself can be reduced by lowering β_D or β_I through better metadata, stronger proof-of-ownership requirements, and jurisdiction-specific validity checks.*

G.1 Model and Empirical Findings

The model offers a compact reading of our empirical results. First, the high matching accuracy we document matches the assumption that, given correct reference data, fingerprinting works well; enforcement errors therefore sit in access and database quality rather than in acoustic matching.

Second, the gap in enforcement between major and non-major labels corresponds to incomplete effective coverage among smaller rights holders,

$$M_C(\theta, z_I, \lambda) = M_D(\theta) + \lambda M_I(\theta, z_I).$$

Major labels tend to hold direct access above θ. Smaller rights holders participate only when expected enforcement revenue clears intermediary fees, revenue shares, and administrative burdens, and even then coverage is partial when $\lambda < 1$. Our estimate of $P(\text{flag} \mid V = 1, \text{type})$ therefore measures effective enforcement coverage, not merely formal inclusion in Content ID.

Third, the large share of non-claimable uploads that receive claims corresponds to a high value of

$$\Omega(\theta, z_I, \lambda, x) = B_E(\theta, z_I, \lambda)\,[s + \phi(x)],$$

which tracks public-domain exposure to stale, invalid, or overbroad database entries. The LatinAutorPerf claims in the *Classical Old* sample fit this mechanism: the system identifies the work correctly, but the database misencodes who owns what, where, and for how long.

Fourth, the difficulty of separating compositions still copyrighted in the EU from those in the US public domain reflects low investment in granular, jurisdiction-specific data, corresponding to high s and low x. Periodic re-attestation of rights and territory-specific metadata would lower s, while stronger ex ante verification and dispute-sensitive database maintenance would lower $\phi(x)$.

Taken together, the model shows that automated copyright enforcement can be technologically sophisticated and still institutionally incomplete. Content ID implements copyright through a tiered access regime: large rights holders participate directly, smaller ones through intermediaries, and some works remain outside effective enforcement, while stale and overbroad entries generate claims against public-

domain material. The resulting allocation can be privately rational for the platform, rights holders, intermediaries, and uploaders without being socially efficient.