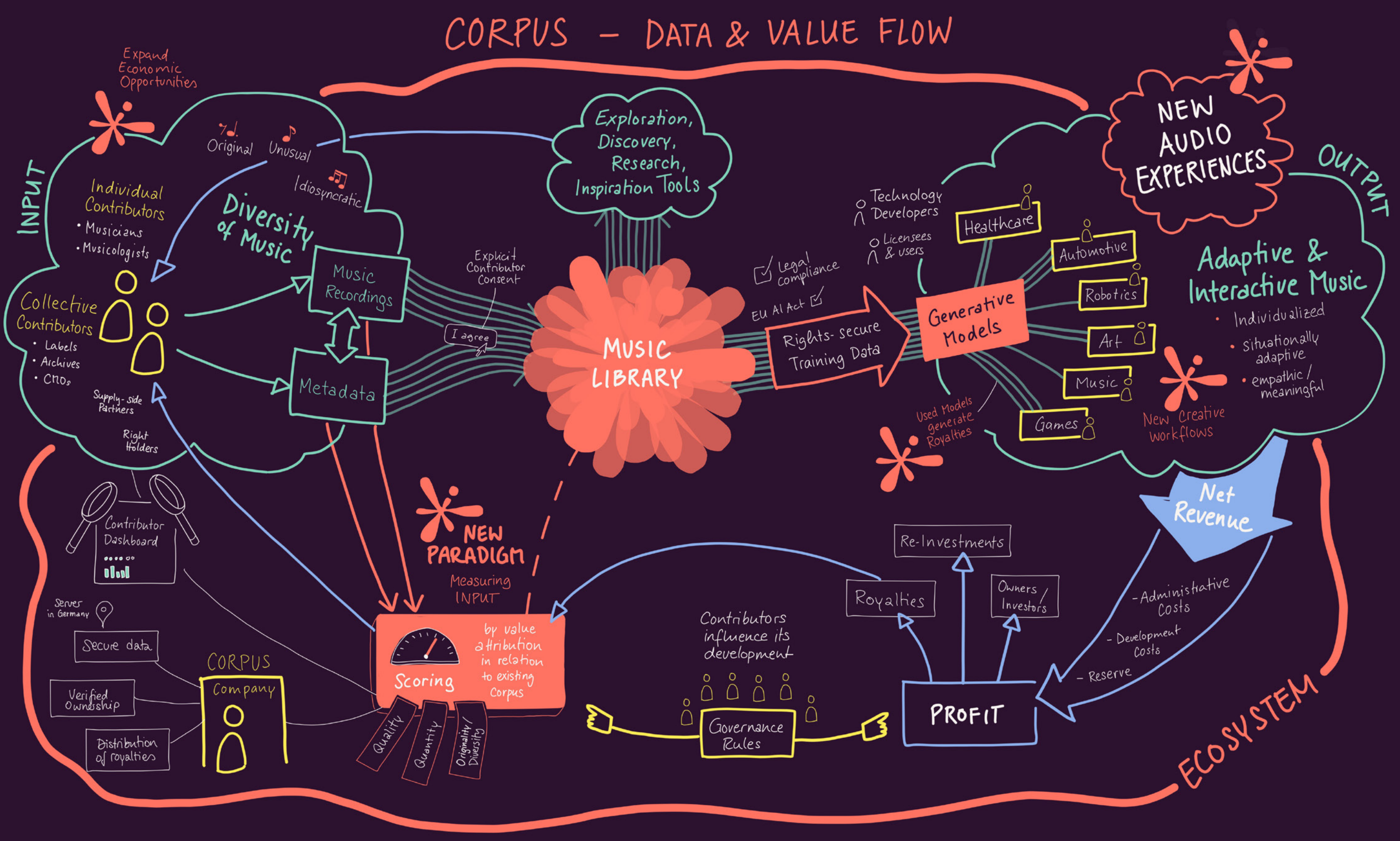
CORPUS

Building the Infrastructure for the Music-AI Economy



CORPUS Data and Value Flow: CORPUS links musical contribution, model training, and economic participation in a single system, shifting value attribution from copied outputs to licensed inputs and shared downstream use.

Executive Summary

From Music Licensing to Music Infrastructure

CORPUS is a platform and licensing protocol that enables music creators to participate in the current technological and cultural revolution. It combines a marketplace for musical works with a new licensing system that provides the legal and economic infrastructure for music to be integrated into intelligent systems.

Generative music AI today faces a deadlock: most models are trained either on unlicensed internet data, which cannot be commercialized, or on full catalogue buy-outs that are financially unsustainable and exclude artists from participation. CORPUS resolves this conflict by introducing a dynamic licensing and royalty system that connects rights holders, developers, and industry partners through a shared, rights-verified music library.

Contributors keep their rights and receive royalties from model revenues, based on a transparent evaluation of contribution quantity, quality, and originality, relative to the existing library. Within this system, originality becomes a source of value – not because it is rare or fashionable, but because musical diversity strengthens the models trained on it, making them more capable, adaptable, and expressive.

For companies, CORPUS provides a compliant, scalable, and cost-efficient alternative to both scraping and buy-outs.

But its role goes beyond licensing. CORPUS enables the music industry to claim the next market before it forms – a new layer of global music demand where billions of devices, applications, and environments will require lawful, adaptive sound models. By aligning artistic and industrial interests, CORPUS turns this transition into an opportunity, establishing the foundation for a sustainable and legally sound Music–AI economy.

Imprint

This white paper is a living document.
Current version: 21 December 2025

Written by
Mathis Nitschke (Founder and CEO, Sofilab and CORPUS)

Edited by
Virginie Berger (Business and brand consultant, artist advocate)
Maria Goeth (Communications director)

Proofreading by
Max Graf (AI Engineer, CORPUS)
Lars Ullrich (CTO, CORPUS)

Map by
Anja von Klitzing (www.dialogstifter.de)

Illustrations by
Jan Stoewe (www.jansteins.de)

Layout by
Anja Gerscher (www.anjagerscher.de)

Published by
Sofilab GmbH
Akademiestr. 3
80799 Munich
Germany

Represented by
Mathis Nitschke

Register court: Munich District Court
Commercial register: HRB233361

Contact
info@sofilab.art
www.sofilab.art

SOFILAB

CORPUS is a project by SOFILAB

Table of Contents

1 Introduction

1.1 From artistic Frustration to infrastructural Necessity

CORPUS is an initiative by Sofilab, a Munich-based sound design and innovation lab working at the intersection of music, technology, and human–machine interaction. Supported by EU funding and private investors, the project grew out of Sofilab's artistic and technological experimentation with generative music models in the late 2010s. Early experiments showed both the promise and the frustration of music AI: while the technology was advancing rapidly, the available training material was either low-quality, poorly annotated, or legally unusable. The absence of a structured, rights-compliant corpus (AI training dataset) made artistic exploration labor-intensive and yielded limited results.

Two insights crystallized from this period. First, the performance of generative models is determined at least as much by their training data as by their technical design. Without rich, diverse, and well-annotated datasets, the technology cannot reach its creative potential.

Second, every licensing system is also an incentive system. Traditional music licensing rewards works that maximize mass appeal and repeat consumption, because royalties are tied to copies and performances. CORPUS, by contrast, decouples value attribution from consumption: a work's weight in the revenue distribution depends on its usefulness for AI training, not on how often it is streamed or performed. Royalties are still paid from actual model revenues – but the internal logic that determines who earns what is based on contribution quality rather than audience reach. This makes unusual, original, and idiosyncratic material particularly valuable, since diversity strengthens models and expands their expressive range.

Parallel to this artistic and technical perspective, Sofilab's work as a service provider in the automotive, medical, and robotics industries revealed another dimension of the challenge. In user-experience sound design and non-verbal human–machine communication, industries increasingly seek ways to shape interaction in individualized, situationally adaptive, and even empathic ways. Drawing on its background in game sound and adaptive music, Sofilab developed early systems of this kind, but the limitation soon became clear: manual preproduction requires anticipating every case in advance, while real-world interactions are full of unforeseen situations. Each additional sensor or behavioral parameter multiplies the possible conditions, creating a combinatorial space that cannot be exhaustively designed by hand. This is not a challenge unique to Sofilab, but one faced broadly across industries. It is precisely here that AI systems reveal their potential – not as a gimmick, but as a way to make sound interactions flexible, context-sensitive, and meaningful even in the unexpected.

This also points to the economic stakes. While music AI may challenge or displace some existing markets for recorded music, it simultaneously opens new ones. Adaptive music and interactive sound environments already exist in games, but they remain constrained by pre-produced assets and finite design choices. With AI, adaptivity can move beyond these boundaries, creating systems that respond fluidly to unforeseen situations. The same principle extends to vehicles, healthcare, robotics, and other domains where sound becomes part of individualized, context-aware interaction. These markets do not exist in the current music industry, yet they could ultimately surpass it in size. If every car, appliance, or robot runs an AI model licensed on CORPUS, each device deployment generates ongoing royalties for contributors – royalties tied to the use of training data and licensed models, rather than to individual output plays.

These combined motivations define the rationale for CORPUS: to create a licensing and royalty system that makes music AI both artistically fertile and economically sustainable.

Figure 2: Independent musicians contribute original recordings to CORPUS

1.2 Who this Protocol is for and how the Ecosystem fits together

CORPUS is conceived as more than a licensing protocol: it is the foundation of a complete ecosystem that connects creators, rights holders, and technology developers. This includes a contributor dashboard, scoring methods, and governance rules alongside the licensing system itself. The system only works if these roles interact transparently and sustainably, aligning incentives across the entire value chain.

Within this ecosystem three primary groups can be distinguished:

- **Individual contributors (e.g., independent musicians)**
They supply the creative material. Their main concerns are: how their works are licensed, how royalties are calculated, and what rights and controls they retain. Transparency and a clearly defined opt-in structure are essential – including the ability to withdraw works from future use while maintaining earned rights from past training cycles.

- **Supply-side partners (e.g., publishers, labels, libraries)**
They manage larger catalogues. Their focus is on integrating CORPUS into existing workflows, scaling value attribution across thousands of works, and ensuring that legal safeguards protect both rights holders and the licensees who rely on them. In practice, such partnerships are typically established through direct agreements rather than via the public contributor platform, which is why they are referenced in this white paper only in selected scenarios.

Figure 3: Publishers, labels, and libraries integrate larger catalogues into CORPUS

- **Licensees and users (e.g., AI developers, game studios, tech companies)**
They need rights-secure training data. Their questions center on reducing legal and reputational risk, tracking and verifying usage, comparing CORPUS with buy-out models, and establishing provenance and permission with confidence.

The white paper is therefore structured around the licensing and royalty infrastructure that connects these groups. Model architectures or training workflows are referenced only where they help illustrate how licensing, attribution, and governance are applied in practice.

Figure 4: AI developers and technology companies license rights-secure training data from CORPUS

2 Current Industry Practices

Despite rapid progress in model architectures and commercial interest, the foundations of generative music AI remain fragile. The datasets used for training are either legally uncertain or financially prohibitive, leaving no sustainable path for artists nor developers.

For example, research datasets such as MAESTRO are licensed for non-commercial use only, making them unsuitable for commercial training, while other large-scale datasets have been criticized for sweeping in copyrighted works without consent.

2.1 Why Scraping, Buy-Outs, and Legal Exceptions Fail

Current practices and legal doctrines all face critical barriers:

- **Scraping** of streaming platforms and archives provides cheap data, but exposes developers and downstream users to lawsuits. For creators, scraping means no consent, no attribution, and no compensation. Models trained this way are nearly impossible to commercialize safely. Downstream adopters – music tools, games, ad platforms – inherit this liability.

- **Buy-outs** offer legal certainty but at prohibitive cost. Even modest catalogues can cost hundreds of thousands of euros, while competitive models require hundreds of thousands of hours of music. Once acquired, catalogues are static, cannot adapt to evolving model needs, and exclude artists from any ongoing revenue.

Both approaches – scraping and buy-outs – create bottlenecks. One excludes rights holders, the other excludes independent developers, startups, and regional rights holders who cannot afford massive catalogue acquisitions. As model complexity and dataset requirements continue to grow, this polarization makes the field increasingly unsustainable.

- In the United States, some developers argue that training qualifies as F**air Use**. But the four-factor test weighs against this view. Commercial intent is high, the entire work is ingested, music is inherently creative, and the effect on the market is disruptive. Recent cases, such as Warhol v. Goldsmith and ongoing lawsuits led by the Authors Guild, have further narrowed the scope of transformative use in commercial contexts. Relying on Fair Use as a long-term foundation is increasingly precarious.

- Europe follows a different path, but the outcome is similar. Article 4 of the **DSM Directive allows** commercial text- and data-mining only if rights holders have not opted out. Several CMOs, including GEMA, have already exercised such opt-outs, explicitly blocking unlicensed training on their repertoires. At the same time, the EU AI Act requires providers of general-purpose AI models to disclose the provenance of their training data. This combination makes compliance significantly harder: unlicensed datasets are not only risky but structurally non-compliant with upcoming regulation.

Taken together, these factors create a patchwork of legal uncertainty and escalating disputes. Scraping cannot remain viable, buy-outs cannot scale, and exceptions cannot be relied on. The space for grey-area practices is shrinking fast, leaving both developers and creators without a sustainable path forward.

2.2 What a new Approach must Deliver

The current landscape leaves both creators and companies without a viable path forward. Scraping erodes trust and invites litigation; buy-outs concentrate access in the hands of a few; collective rights systems were never designed for training data.

A sustainable framework must meet three conditions at once:

- **Legal compliance** – music explicitly licensed for training, not swept in by default.
- **Fair compensation** – royalties reflect the contribution and influence of each work in AI training.
- **Economic scalability** – access affordable for startups, SMEs and cultural institutions, not just major players.

CORPUS is built to meet these conditions by shifting licensing to the input side. Contributions are licensed opt-in, evaluated for quantity, quality, and originality relative to the existing corpus, and rewarded accordingly. This creates an incentive system that strengthens the dataset while remaining legally defensible and economically inclusive.

Such a system cannot be realized within the frameworks that dominate today's music industry.

3 Why Existing Rights Systems struggle with AI Music

3.1 CMOs were built for Copies and Performances, not Training

Collective rights management organizations (CMOs) such as GEMA, SACEM, or PRS were created to solve the problem of licensing copyrighted works at scale in the era of mechanical reproduction, broadcasting, and public performance – a framework shaped between the late 19th and mid-20th century. Their model assumes that works exist as fixed units and that value arises when those units are copied or transmitted at scale.

AI music generation does not fit this framework. It does not copy or perform existing recordings for end-user listening, but produces new audio from learned patterns. This raises immediate legal questions: is training itself a licensable act? Do CMOs even hold the relevant rights to issue such licenses? For example, PRS's performance rights terms do not cover training. More broadly, most membership agreements were written long before AI was a factor, covering public performance and reproduction but not machine learning.

Some CMOs have already begun to explore AI-specific licensing, but in very uneven ways. In Sweden, STIM has launched a pilot collective license for AI training in partnership with startups such as Songfox and Sureel, where authors can explicitly opt in and receive attribution and royalties. In France, SACEM has taken an authorization-first stance, requiring explicit consent for AI data-mining of its repertoire. In Germany, GEMA has proposed a licensing model that would require AI developers to pay rights-holders not only for training but also for certain downstream uses of AI-generated music, including proposals for significant revenue shares (e.g. 30% of net income) and minimum royalties. At the same time, GEMA has floated extensions of collective licensing frameworks, including Extended Collective Licensing (ECL), to apply to AI training – an approach whose legal basis in this domain remains uncertain.

These examples show that CMO strategies are not coordinated and sometimes push beyond existing mandates. This inconsistency highlights unresolved questions: Who decides what can be licensed? Who gets paid? And do CMOs legally hold the rights they claim to exercise over AI training?

3.2 The Collapse of "Copy logic" in a World of Generative Outputs

At a deeper level, the problem is not just contractual – it is conceptual. The entire logic of collective rights management is built around the economic regulation of copying: the assumption that value is created when a work is duplicated and used at scale.

In the context of generative AI, there may be no copies at all. Music will often be generated on demand by models that sit directly between listener and machine. A therapeutic robot generating calming music for a patient, or a game engine producing a one-time soundtrack for a player, illustrates the point: the output exists only in that moment and may never be heard again. There is no original and no copy – only individualized, ephemeral experience, with no intermediary platforms, publishers, or playback channels.

This undermines the basis for royalty calculation as we know it. Concepts like public performance, mechanical reproduction, and broadcast no longer apply in a meaningful way. Attempts to retrofit existing frameworks treat AI as just another distribution channel. This misses the core shift: music is no longer distributed, it is generated. The industry has made similar mistakes before: downloads were initially licensed like physical sales, and streaming was first treated as a variant of downloads, leaving creators under-compensated and sparking years of dispute. Doing the same for AI risks repeating this structural error, leading to overreach, legal friction, and a further loss of trust.

3.3 CORPUS's input-side Paradigm as a Response

CORPUS takes the opposite approach. It assumes that the decisive moment is not the detection of copies, but the licensing of contributions for training. This does not replace usage-based royalties; it defines how usage is weighted. CORPUS attributes value when works enter the training corpus: each contribution is assessed for how it enriches the existing library with respect to the goals of quantity, quality, and diversity. This assessment results in an input weighting that remains fixed once assigned. On its own, the weighting does not generate payment. It determines how contributors participate in the royalty pool once models trained on their works are licensed and deployed.

By shifting attribution to the input side while keeping payouts tied to real economic use – model licensing, device deployments, API calls – the protocol establishes a framework that matches the realities of AI music. It rewards contributions according to their influence on trained models, not according to the reproduction of finished works.

Rather than preserving the logic of collective rights management, it replaces it with a paradigm built for personalized generation, AI-powered tools, and non-fixed outputs – contexts where conventional output-based counting no longer applies.

Where current collective rights systems measure outputs, CORPUS measures inputs, and links those inputs to usage-based revenues.

4 Applications: Where Licensed, Diverse Training Data is Essential

CORPUS is designed to serve domains where conventional approaches to music cannot meet the demands of scale, responsiveness, or diversity – and where AI can unlock new possibilities. These use cases show where generative music creates the highest leverage: expanding creative workflows for musicians, enabling adaptive sound in industrial contexts, and opening cultural and educational applications. Crucially, AI models trained on unlicensed data cannot be deployed ethically and commercially in regulated or brand-sensitive sectors such as automotive, healthcare, or advertising. A licensed and diverse corpus is therefore not just desirable, but the only viable foundation for lawful, scalable, and trustworthy adoption.

Figure 5: Generative music unfolds as an adaptive, responsive environment rather than a fixed asset.

Music production tools and Digital Audio Workstations

AI-assisted composition tools and web-based DAWs are proliferating, lowering barriers for musicians who lack access to education, budgets, or professional networks. This segment is expected to see strong growth, with new platforms appearing almost weekly. In these environments, users do not simply consume AI output but actively refine and correct it. That human-in-the-loop process is precisely what generates value: each interaction strengthens creative outcomes and, when built on CORPUS-licensed models, feeds back into a virtuous cycle where musicians/producers benefit both as contributors and as users of ever-improving tools.

Interactive Artistic Environments

Games, XR, and installation art increasingly demand music that behaves like a living environment rather than a fixed soundtrack. CORPUS-trained models allow scores to unfold in real time, responsive to player movement, narrative, or audience interaction. While adaptive music already exists, it is bounded by pre-produced assets. With CORPUS, these boundaries expand: developers could, for instance, integrate a Wwise plugin into Unity or Unreal Engine, enabling AI-trained scores that adapt continuously without exhausting variation. This creates artistic experiences that are not just reactive but genuinely co-evolving with their users.

Adaptive soundscapes in mobility

Vehicles process a constant stream of inputs: driving behavior, speed, location, weather, and even driver state. Designing sound responses for all these conditions manually is impossible. CORPUS-trained models enable adaptive soundscapes that respond intelligently – supporting driver concentration, comfort, and emotional balance while maintaining consistency with the vehicle's brand identity. Each vehicle can host its own licensed model, generating feedback and music that adjust in real time to context – creating a coherent yet individualized auditory experience across an entire fleet.

Healthcare robotics and therapy

Hospitals increasingly explore humanoid robots and assistive devices that provide guidance, companionship, and emotional support, while also helping to reduce staff workload. For dementia patients, for example, music and responsive listening are key: familiar songs, regional repertoires, or even improvised singing can create comfort and connection. Research shows that music therapy can reduce agitation, improve mood and memory, and strengthen communication and social bonds for people living with dementia. CORPUS emphasizes diverse, global contributions, making such culturally attuned applications possible – something standard datasets, dominated by Western commercial music, cannot provide. Beyond robotics, hospital devices can also draw on music-therapy principles to address alarm fatigue – replacing uniform, high-stress alerts with sound cues that differentiate urgency levels and convey calm where appropriate. This allows critical alarms to remain clear while reducing unnecessary stress for patients and staff.

Advertising and brand communication

Brands increasingly invest in distinctive sonic identities – from carefully composed sound logos to curated music palettes – but today these assets remain largely static. CORPUS-trained models make it possible to extend such brand references into adaptive experiences that stay recognizably on-brand while responding to context. A retail store, for example, could move beyond fixed playlists: background music might shift with the time of day, then adapt in real time to live sensor data such as visitor numbers, demographics, or overall atmosphere. At the same time, lawful machine-generated music enables rapid, low-cost production of branded material – useful for campaign prototyping, regional adaptations, and highly segmented user groups. This allows brands to maintain aesthetic coherence across thousands of micro-contexts without relying on unlicensed or generic content.

Exploratory Cultural Mediation and Education
Apart from highly interactive and personalized instrumental lessons, music education still relies largely on static examples and fixed recordings – limiting active engagement and experimentation. CORPUS-trained tools make music itself interactive: learners could trace the lineage of a folk melody across cultures, reshape motifs on the fly, or dialogue with a model trained to explain harmonic structures. In public institutions such as libraries or museums, this enables exploratory interfaces where visitors not only listen to regional repertoires but transform them – remixing, extending, or cross-pollinating across traditions. Such forms of cultural mediation depend on detailed metadata that connects works across styles, origins, and histories – another reason CORPUS's annotated diversity is essential. These tools make musical literacy embodied, playful, and deeply satisfying, opening new pathways for institutions to connect heritage with contemporary creativity.

Taken together, these applications demonstrate a decisive fact: in many domains – from music production to artistic interactive environments, from mobility to healthcare, advertising and cultural mediation and education – the scale and adaptivity required cannot be achieved with traditional approaches to music creation. AI expands creative workflows by lowering barriers and adding new possibilities. Only AI models trained on a rich, diverse, and deeply annotated music library can generate sound that adapts continuously to context, scale, and interaction – and only if that corpus is built on trust, licensing, and contributor participation.

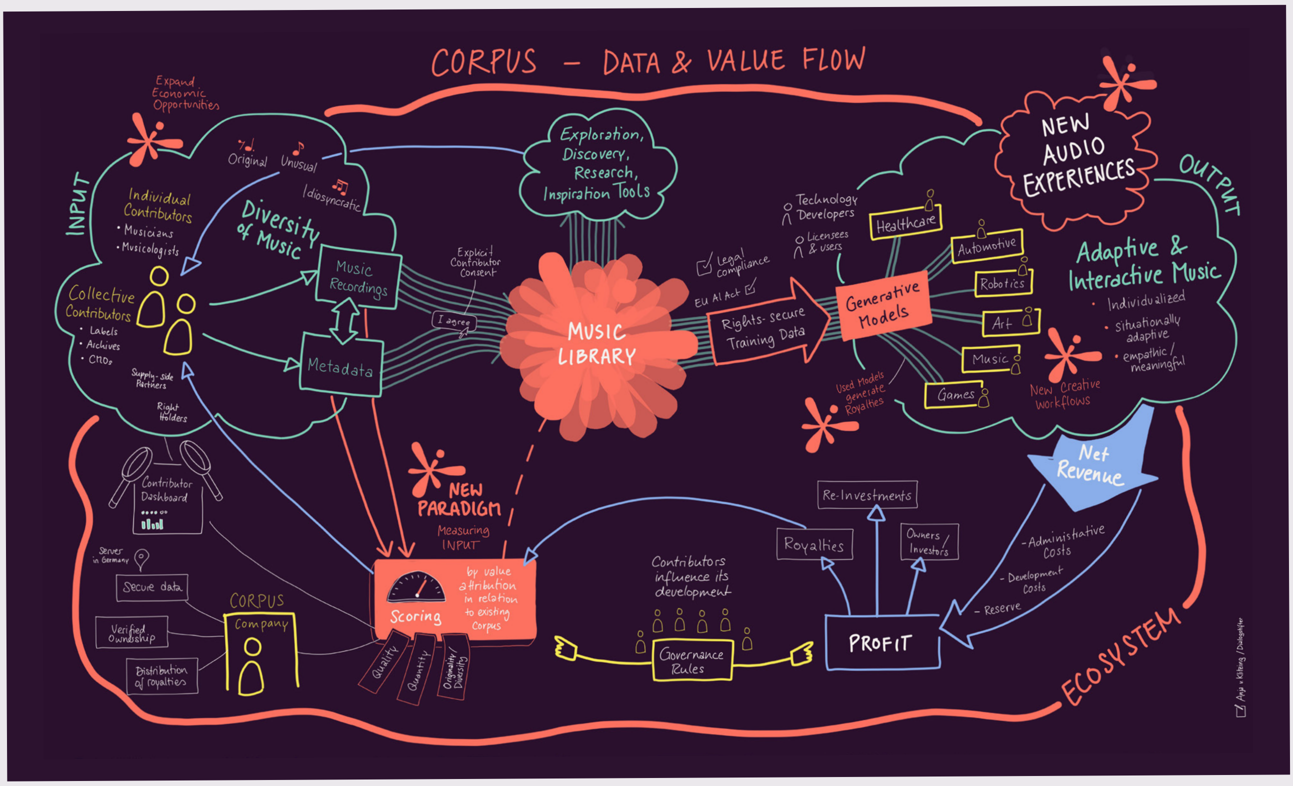


Figure 6 CORPUS – Data and Value Flow: Musical works enter CORPUS, are evaluated for their contribution, and serve as licensed training data. Revenues from model use return to contributors, linking creative input to shared economic value.

5 The CORPUS Royalty Protocol

CORPUS is in development. What follows is not a description of an existing system, but the design principles guiding its implementation. The protocol is conceived as a licensing and royalty infrastructure that incentivises the global musician community to contribute their music to a shared library, designed to serve as a training dataset for generative AI and other downstream AI applications.

Security and rights management are integral to this process: uploads will be verified for provenance, licensed works tracked across training and deployment, and royalty flows made auditable and transparent. The principle is clear: businesses and their users gain legal certainty, and music contributors maintain trust and long-term participation. Later chapters describe the mechanisms in detail – from verification procedures to auditing and governance.

In its initial phase, CORPUS will operate as a platform to establish the dataset and validate the licensing model. Over time, it is intended to evolve into an open protocol that can be embedded into existing services and creative tools, providing the foundation for an ecosystem that extends beyond any single marketplace or application.

Future integrations could make this process seamless. For example, musicians uploading tracks via distribution services such as MusicHub, TuneCore, or DistroKid might see an additional option: “also register with CORPUS.” The same could apply to community platforms like Bandcamp, where independent artists could license their works for AI training alongside sharing them with audiences. With a connected wallet, the contribution would be registered in one step. Depending on the integration, CORPUS could reuse existing rights metadata from the distributor or import structured data via standards such as DDEX, with only a lightweight additional declaration for AI-specific consent to ensure all parties agree. The same principle extends to creative tools: in DAWs or music apps, CORPUS could appear as a native publishing or export target.

The protocol borrows the structural logic of Web3 systems – automated and trustless execution of agreements – while remaining technology-neutral at this stage. Whether blockchain becomes part of the technical foundation will depend on scalability and adaptability; for now, CORPUS prioritises flexibility in the early stages of development.

Figure 7: By designing licensing, provenance, and auditability into the system from the outset, CORPUS makes regulatory compliance— including alignment with the EU AI Act—a structural outcome rather than an added constraint.

5.1 Core Design Principles: Dynamic Licensing, Traceability, Governance

CORPUS balances creator incentives with industry compliance, ensuring that contributions remain usable, secure, and commercially deployable in sensitive markets. These principles define the framework for licensing, attribution, and governance, making CORPUS not only fair for contributors but also the only viable path for lawful, scalable adoption in regulated industries.

- **Dynamic licensing and royalties:** Instead of buy-out catalog purchases, CORPUS licenses works under ongoing agreements. Contribution value is weighted on the input side, evaluated for quantity, quality, and originality relative to the existing library, so that rewards favor material that expands diversity and depth. Royalties follow the monetization logic of downstream applications – for example, models licensed per device in a car fleet, or API usage in a game engine – so that rewards scale with real economic activity.

- **Fully licensed, traceable, and opt-in:** All music in CORPUS enters through explicit contributor consent. Works remain owned by their creators and can be traced across models and applications through a central registry that links dataset usage back to individual contributions. In practice, contributors gain visibility through their dashboard: once a dataset has been compiled for a model, they can see where their music was included and how it generates value.

- **AI as a creative partner and market enabler:** Models trained on CORPUS are not designed to replace musicians but to expand both artistic workflows and economic opportunities. They can function as responsive collaborators in creative contexts, while also powering entirely new markets for adaptive sound in vehicles, healthcare, robotics, advertising, and other interactive domains. These emerging markets have the potential to outweigh the losses caused by generative AI in traditional music distribution by a large factor. Equitable structuring is essential; without it, the value created by these markets bypasses creators.

- **Community-governed evolution:** The protocol includes mechanisms for contributors to influence its development: auditing revenue flows, adjusting attribution rules, and participating in governance decisions. This can happen through contributor voting, advisory boards, or other representative structures. The principle is that those who provide the data also help shape the system’s direction and safeguards.

- **Open and engaging discovery:** CORPUS-native tools are designed to make both creation and discovery adaptive and alive. The library is not a closed archive but a living resource: contributors’ music remains accessible for exploration, learning, and interaction.

Discovery is intended to be playful and serendipitous rather than static or purely transactional.

Beyond search, CORPUS enables new creative workflows: musicians and producers can explore music by mood, story, or emotional arc, uncovering connections that spark ideas rather than reproducing clichés – and in doing so, reward the diverse contributions that feed these exploratory tools.

Tools like our Story2Music link narrative prompts directly to musical ideas, reviving collaborative workflows that stock music has displaced and opening new spaces for artistic and professional creation.

5.2 The Licensing Framework: Connecting Contributors and Licensees

The licensing framework connects three key actor groups:

- **Contributors:** Musicians provide recordings, and they can also add or refine annotations such as metadata or quality checks. While creators are encouraged to enrich their own works, CORPUS recognizes that only a minority will consistently do so. To ensure coverage and quality, other participants can contribute and edit annotations as well and receive points for this value addition. In this way, both music and metadata become part of the shared corpus and are rewarded through the same incentive system.

- **The CORPUS dataset**: Licensed works and annotations form a central dataset that supports multiple AI training pipelines. A single contribution can generate value across many contexts – for example, a track included in training a generative engine for cars may also be part of subsets later licensed for adaptive music in games, healthcare applications, cultural projects or mobility systems. In this model, contributors are not compensated only once, but share in multiple downstream revenue streams each time the dataset or its subsets are licensed for training. This makes every contribution a recurring source of income, aligned with the evolving applications of the corpus.

- **Model users, developers, and service providers:** AI developers license access to the CORPUS dataset, either the full library or specific subsets tailored to their application. Revenue from these deployments flows back into CORPUS and is redistributed to contributors according to the system’s dynamic value attribution.

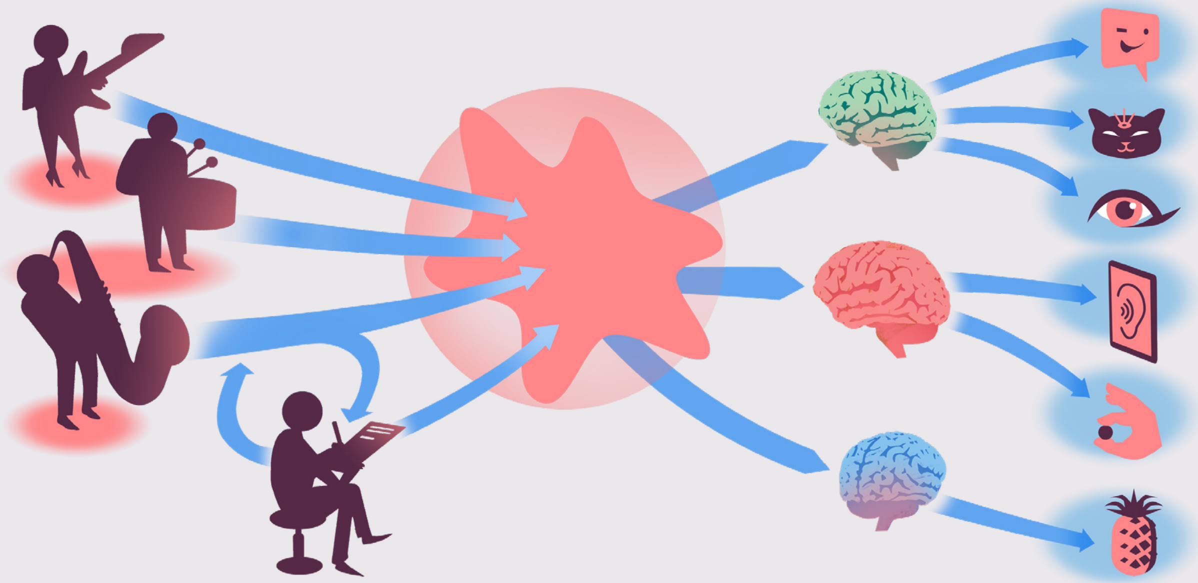


Figure 8: Musicians contribute recordings to a shared corpus. This corpus is used to train multiple AI models with different goals. Each model can then power a wide range of applications and behaviors.

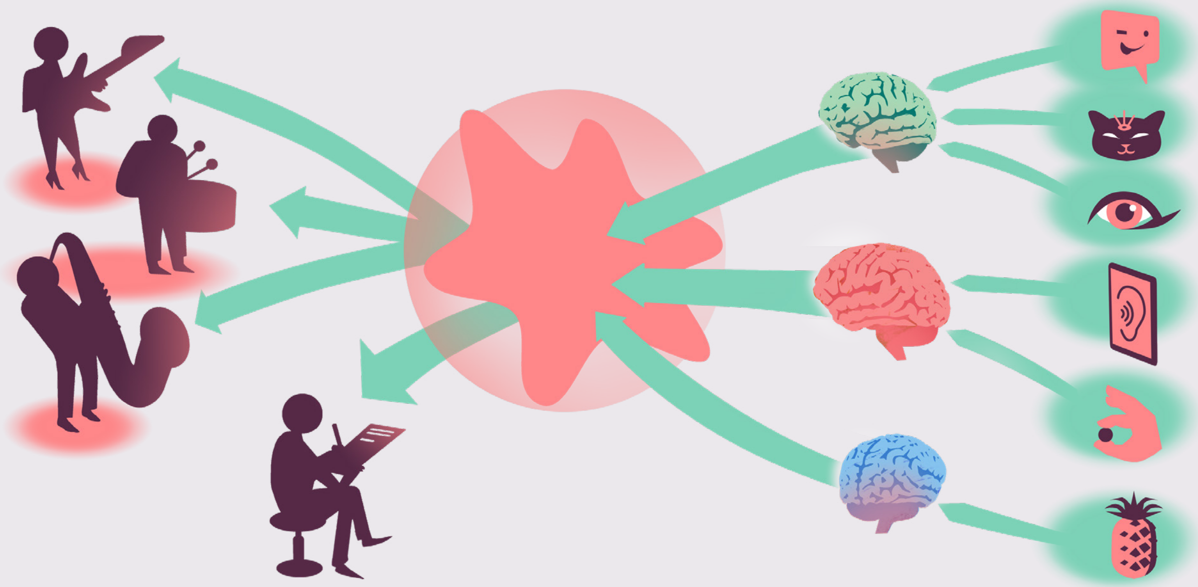


Figure 9: AI developers pay for model usage; downstream applications generate revenue; and royalties flow back to contributors based on their weighted participation in trained models.

6 Securing Contributions and Rights Integrity

The credibility of CORPUS depends on whether rights holders, industry partners, and regulators can trust that the system secures data, verifies ownership, and distributes royalties in a legally defensible way. Security and rights integrity are not add-ons but the backbone of the protocol. This chapter outlines how contributions are uploaded, verified, licensed, and traced – some measures already in place, others phased in as CORPUS scales. Together, they form the roadmap for a transparent, enforceable licensing infrastructure.

6.1 Protecting Uploads and Storage from the Start

All contributions enter CORPUS through encrypted upload channels and are stored on self-administered servers located in Germany under EU data protection law. This avoids reliance on opaque cloud hyperscalers and ensures data is handled under clear governance. As the corpus grows, this will evolve into a federated global server network with nodes in strategic locations, each operating under the same standards.

Infrastructure is aligned with recognized security frameworks, including ISO 27001 (risk and information security management) and SOC 2 (audited controls for secure operations). While formal certification will follow at commercial scale, alignment from the outset ensures compatibility with partner expectations and with upcoming regulatory obligations under the EU AI Act, such as Article 52 provenance requirements for foundation models.

For model training, CORPUS plans to integrate confidential computing technologies such as NVIDIA's secure enclaves, which ensure that even during active computation, training data cannot be inspected or exfiltrated by providers – closing one of the most sensitive attack vectors in AI workflows.

6.2 Verifying Rights and ensuring Provenance

Ensuring that contributions are legally usable is central to CORPUS's value. The system combines contributor consent, ownership safeguards, and transparent verification procedures.

Contributions are currently governed by provisional, non-commercial terms. These permit internal model training for R&D, prototypes, and partner demos, but not the public release of trained models. At no point does contributor data become platform-owned. Once commercial licenses are finalized, contributors will explicitly opt in, ensuring that all future use is revenue-sharing and transparent.

- **Contributor ownership and consent**

Contributors retain ownership, and works are licensed only through explicit opt-in. Withdrawal is possible at any time, stopping future use of a work in new training runs. However, if a work has already contributed to model training, its influence remains embedded in those weights. CORPUS therefore continues to treat the work as royalty-eligible, ensuring that contributors are compensated for past contributions even after withdrawal.

This approach respects withdrawal rights for future uses while recognizing the technical impossibility of “untraining” existing models. It also sets CORPUS apart from today’s industry norm, where once a work is ingested, contributors lose both control and economic participation. CORPUS instead guarantees residual value – a unique mechanism that keeps compensation flowing even when rights holders later change their participation status.

- **Split-contribution attribution**

CORPUS supports split attribution for works with multiple rights holders (e.g., composer, performer, producer). Each contributor is acknowledged, and commercial licensing requires explicit confirmation or indemnification to manage disputes transparently.

- **Peer review and metadata validation**

Metadata and quality checks are partly community-driven. To prevent bias or neglect, initial deployments will rely on invite-only expert reviewers. From 2026 onward, CORPUS will introduce structured systems – for example, reviewer tiers or a strike system – supported by audit logs and escalation paths to ensure transparency and accountability.

- **Detection of AI-generated uploads**

As of now, CORPUS actively screens uploads for synthetic origin to prevent unauthorized AI-generated material from entering the corpus. Submissions are checked using dedicated detection systems as part of the ingestion pipeline, and flagged material is reviewed before acceptance. Because detection methods and generation techniques continue to evolve, this remains an adversarial and moving target.

Looking ahead, CORPUS plans to complement detection with additional safeguards, such as watermarking or hash-based provenance tools where they prove reliable and practical. In parallel, the protocol is being developed with emerging regulatory frameworks in mind. Several jurisdictions, including the EU under the AI Act, are moving toward mandatory provenance disclosure, and CORPUS is preparing to adapt as such requirements become concrete and enforceable.

- **Handling infringements and disputes**
If a contribution is challenged, CORPUS is providing an appeal and mediation process. Works may be quarantined during review, but contributors will not be penalized without due process.

6.3 Building Trust through Auditability and Transparency

Auditability is both a governance tool and a legal necessity. Every contribution, license, and royalty flow in CORPUS is logged in an append-only, tamper-evident registry, ensuring past records cannot be altered without detection. These logs provide the evidentiary basis for compliance with EU law – for example, DSM Directive Article 4 on text- and data-mining opt-outs and AI Act Article 52 transparency obligations – while also giving contributors confidence in the transparency of corpus management.

Looking ahead, CORPUS will also support embedding attribution metadata into model outputs – for example, through ISCC codes or provenance hashes. This ensures links between training data and generated material remain visible across the ecosystem. A game studio or mobility partner, for instance, could verify the provenance of a licensed model before integration into its engine or device.

In sum, CORPUS treats security and rights management not as features but as the foundation of its licensing protocol. With append-only audit logs, explicit contributor consent, and defensible verification procedures, it builds an infrastructure that can withstand regulatory scrutiny, protect creators, and give industry partners confidence in the corpus's legitimacy.

7 From contributions to Royalties: How Value flows in CORPUS

Unlike traditional licensing, where royalties are tied to the reproduction of finished works, CORPUS allocates relative value on the input side – assigning each contribution a weighting that reflects its usefulness for AI training. These weightings determine how future revenues from licensed model use are distributed, creating a continuous link between artistic contribution and economic participation. The result is a system where compensation follows actual model impact, not consumption metrics.

7.1 Incentive logic: rewarding Quantity, Quality, and Diversity

The attribution system translates contributions into stable scores that determine long-term royalty flows. Scoring happens once – at the moment a contribution enters CORPUS – and remains fixed except for metadata improvements or fraud/error corrections. This design creates clear and predictable incentives.

Three dimensions of value

- **Quantity**: every contribution expands the breadth of the corpus. A baseline score ensures participation itself is rewarded.
- **Quality**: works must meet technical and descriptive standards to be usable. Clean recordings, accurate metadata, and consistent annotation raise weighting.
- **Diversity**: contributions that expand the corpus into underrepresented areas earn extra weight. Originality is measured through relational analysis, not by surface-level tags: the system evaluates how a work positions itself relative to existing data points, capturing structural, stylistic, and cultural distinctiveness. In this way, CORPUS rewards genuine expansion rather than saturation. The method also helps counter historical biases in training datasets – such as the overrepresentation of Western commercial music – ensuring that the corpus grows as a resource that is both economically fair and inclusively designed.

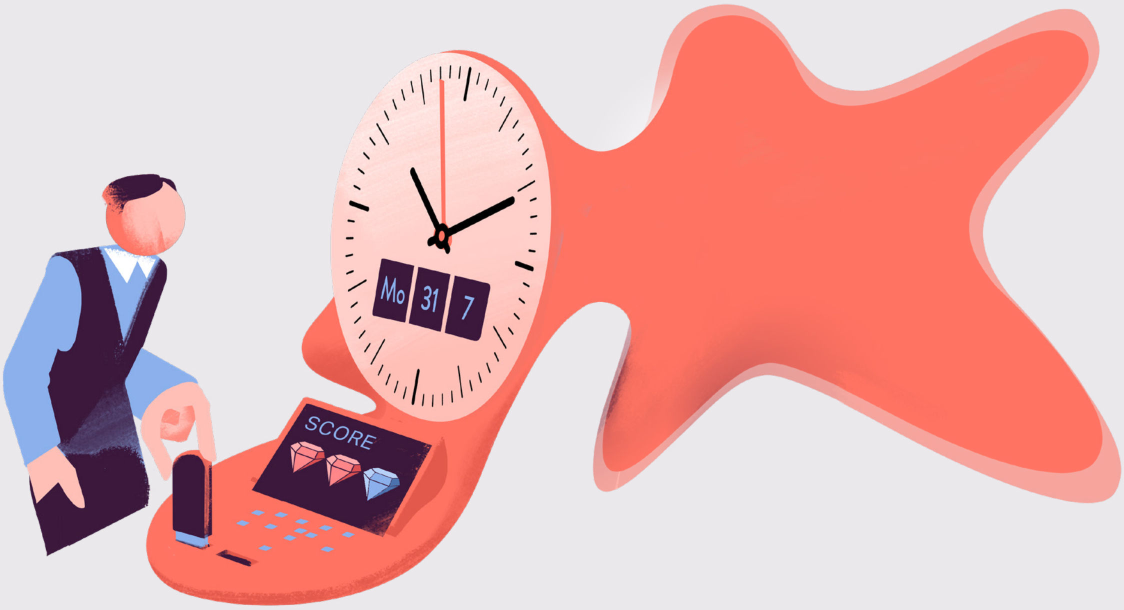


Figure 10: The contribution of each work is assessed at the moment it enters the corpus, and an input weighting is assigned based on how it enriches the library in terms of quantity, quality, and diversity.

Technical attribution structure

Figure 11: Baseline Contribution Scoring

- **Quantity: Base points for contribution**

Each work receives a baseline score depending on input format and completeness. For example, an uncompressed WAV may earn 100 points, an MP3 50, with additional points for stems, MIDI files, or detailed metadata.

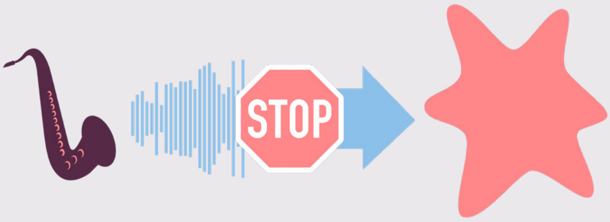


Figure 12: Contribution Integrity Checks

- **Quality: Automated filters and edge cases**

Non-musical content, illegal samples, or corrupted files are excluded. Filters may include spectral anomaly detection, provenance hashes, similarity searches against databases of copyrighted songs, and AI-detection tools. Flagged works earn zero points until reviewed, with appeal options for contributors.

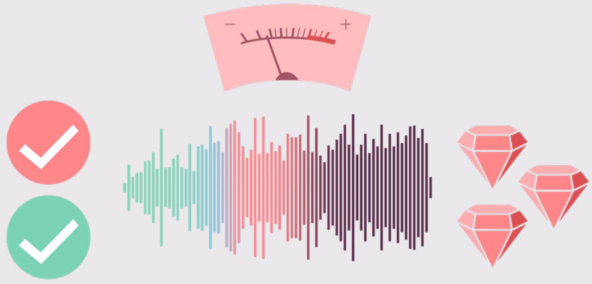


Figure 13: Production Quality Assessment

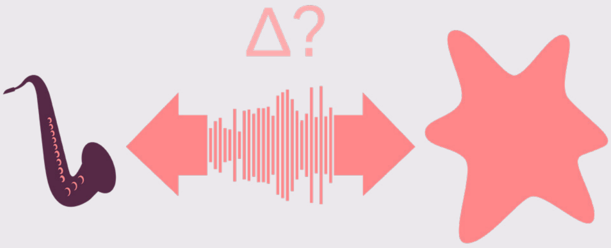


Figure 14: Relational Originality Scoring

Figure 15: Metadata as First-Class Contribution

- **Quality: Production quality**
An evaluation system using Music Information Retrieval (MIR) benchmarks assesses production quality (spectral balance, dynamic range, noise levels, stereo field). High-quality recordings can earn bonus points; flawed material may earn none or be excluded. To avoid frustration, contributors can revise or resubmit flawed works, supported by automated feedback that highlights specific issues.

- **Originality: Diversity and Novelty**
Contributions are compared against the existing music library at the moment of ingest. Works that fill gaps or introduce new material earn additional points, while oversaturated areas are weighted lower. Originality is never judged in absolute terms but always in relation to what already exists. Crucially, this extends beyond genre tags: CORPUS applies musical analysis that detects distinctiveness in structure, timbre, and style across genres. Originality can thus arise from unusual harmonic language, experimental production, or cultural specificity just as much as from genre novelty.

- **Annotation and metadata contributions**
Annotations – such as genre, mood, instrumentation, or cultural context – are a critical part of corpus value. Musicians can add or refine these for their own works, but others may also contribute, either voluntarily or incentivized. All verified annotation work is rewarded with points, recognizing that metadata quality is as vital to training as the recordings themselves.

Together, these mechanisms discourage mass-uploading of poor or redundant material and reward contributions that expand the corpus in quantity, quality, and diversity. Originality is easier to achieve in the early stages and becomes progressively more challenging as the dataset grows – a dynamic that incentivizes contributors to keep pushing into new territory rather than repeating what is already abundant.

7.2 Translating Points into Revenue shares

Contributor points translate into revenue shares through a transparent two-step process.

- **Revenue pool and costs**
  All revenue from sublicensing and from CORPUS-owned products and tools flows into a central pool. A defined share covers infrastructure, legal, staff, R&D, and reserves for growth. The remaining net revenue forms the royalty base – ensuring contributors are paid transparently while CORPUS retains resources to scale and create long-term value.
- **Model-level attribution**
  Royalties are tied to the training stage, not to individual outputs. Contributors are paid when their works form part of the dataset for a monetized model. For each distribution round:
  - Points per contributor are summed across all works included in that model.
  - A global point total is calculated across all contributors in the model.
  - Revenue is allocated proportionally.
- **Example**
  If Model A generates €1M net revenue, and Contributor X holds 0.5% of the model's total points, they receive €5,000.

Works that score highly but are not yet included in a monetized model remain eligible for future rounds. As new models are trained, these contributions may begin generating revenue, ensuring that value can be realized over time rather than only at the moment of upload.

7.3 Oversight and Dispute handling

Trust in the protocol requires more than formulas. CORPUS builds governance and transparency into payout operations so contributors can verify how their works generate value.

- **Contributor dashboard**: each contributor has a dashboard showing which of their works are included in monetized models and which remain pending for future eligibility. It also displays point balances, royalty shares, and payout history, making attribution and revenue flows visible in real time.

- **Distribution round**: initial payouts may be annually or quartely, with frequency increasing as revenue grows. Small balances roll over to reduce transaction costs.
- **Payout processing**: contributors complete Know-Your-Customer procedures before payments; each payout is accompanied by an invoice or credit note that the contributor can use in their own tax declaration.
- Auditability: all financial flows are logged in tamper-evident records. Contributor-elected boards will review reports and audit trails, with the option for external audits as the system matures.
- **Dispute resolution**: contributors can appeal point allocations, fraud flags, or payout calculations. Governance mechanisms define escalation steps and mediation procedures, ensuring disputes are handled fairly and transparently.

Beyond formal audits, CORPUS will establish participatory feedback loops. These may include contributor surveys, score simulations, and public comment periods, giving musicians and partners a direct role in testing attribution rules and refining payout practices. This ensures that oversight is not only top-down but actively shaped by the community that powers the corpus.

7.4 Balancing Complexity with Usability

Any system that balances fairness with resistance to abuse will sound complex. The danger is opacity: rules pile up until even insiders no longer understand them, as seen in tax codes or legacy royalty schemes.

CORPUS addresses this tension in two ways. First, complexity exists only where it increases fairness or prevents abuse – not where it obscures flows. Second, interfaces absorb the complexity. Contributors don’t need to understand formulas (but they can if they want to); they see a clear breakdown in their dashboards: which works, which models, which revenues, which payouts.

Community input will remain central. Contributors will test, debate, and refine attribution rules to keep the balance between fairness and usability.

7.5 Beyond Royalties: Ownership and Long-term Participation

One of CORPUS's founding inspirations comes from equity participation models: systems where contributors, like members and employees in cooperative or startup settings, share in the long-term growth of the network itself rather than only receiving transactional payouts.

Streaming platforms like Spotify concentrate profits with shareholders, while musicians see only fractions of revenue. CORPUS explores alternative models where consistent, high-value contributors share in system-level appreciation – whether through revenue-linked bonuses, governance rights, or future equity-like schemes.

This area remains under research, since conventional mechanisms can create tax or compliance challenges. But the principle is clear: contributors should not only power CORPUS with their works, but also participate in the upside of its growth – under the strict condition that such mechanisms remain legally sound and free of hidden liabilities for contributors.

8 Organizational Structure and Legal Evolution

8.1 Why structure matters for Trust and Scalability

CORPUS is not only a technical protocol but also an institutional arrangement. How rights, revenues, and responsibilities are organized will determine whether the system can scale while maintaining contributor trust. For artists and rights holders, long-term safeguards and visibility into how their works are treated are essential. For funders and investors, legal clarity and enforceable rights are the precondition for scalable growth. For regulators, alignment with existing frameworks is critical to ensure defensibility.

In practice, this means resolving questions such as: who owns the protocol's IP, how contributor payouts are guaranteed, and what mechanisms (such as exit veto rights) protect against misuse. The organizational design must therefore balance speed and innovation with protection against misuse, ensuring that growth strengthens trust rather than erodes it.

8.2 Current setup: from Sofilab Project to dedicated Entity

CORPUS is currently developed and operated by Sofilab GmbH, with project funding from the European Union and private investment from the founder. The next step is to spin CORPUS out into a dedicated GmbH under German law. This structure is easy to establish, provides legal clarity, and makes the project investable at scale, enabling additional participation and growth beyond what is possible within the current Sofilab framework. Over the longer term, further safeguards will be required to ensure contributor trust is protected even as the company scales – a need that the following section explores in more depth.

8.3 Future Pathways: balancing Speed, Trust, and Investor viability

The long-term structure must balance two forces: the speed demanded by today's AI industry, and the trust contributors need to commit their works.

In conventional startup culture, the endgame is often an exit – the sale of the company to a larger player. Legally, any buyer would remain bound by contributor agreements. Yet experience in the music-tech sector shows that this is not enough to guarantee trust. When Bandcamp was acquired, for example, artists quickly voiced fears that its pro-artist revenue model could be diluted under new ownership. Even without proven breaches, the perception of misalignment was enough to create uncertainty. CORPUS faces the same challenge: contributors may worry that an exit could shift priorities toward investor returns, narrow the interpretation of agreements, or pressure renegotiation.

To prevent this, CORPUS is currently evaluating governance models that combine investor viability with long-term rights protection:

- **Cooperatives:** These guarantee contributor control, since major decisions require community consensus. But they are often too slow and rigid to compete in the fast-moving AI sector.

- **Foundation-based safeguards:** A dedicated foundation could hold a Golden Share in the CORPUS IP. This special share grants veto rights over strategic decisions, preventing misuse of the protocol's IP without foundation approval. Such a foundation would have a board of trustees (Stiftungsrat or Kuratorium) responsible for oversight (e.g. approving financial reports, supervising management), and this board could include members elected from the contributor community as per its statutes – giving contributors real voice in governance.

In the early phase, before formal governance mechanisms are established, contributors retain the unilateral right to opt out if they disagree with the project's direction. Once governance is enacted, decisions will be shaped collectively through those structures, and withdrawal will follow the agreed contributor terms.

8.4 Ongoing Evaluation of Governance Models

Just as CORPUS balances openness, quality, and consent at the content level, the legal structure must do the same at the organizational level. No single structure has yet been fixed. The project team is currently testing how best to balance three priorities:

- **Contributor protection** – ensuring that artists retain visibility, influence, and long-term security of their rights.

- **Funder and Investor alignment** – building a scalable company that attracts capital while safeguarding the asset value of the corpus.

- **Governance efficiency** – avoiding overly complex systems that would slow adaptation in a fast-moving sector.

The final design will be developed in dialogue with contributors, investors, and legal experts. What is clear already is that CORPUS cannot rely on legacy models alone. The same values that guide the protocol must guide the institution: transparency, resilience, profitability, and fair alignment of incentives – especially in a sector where default structures too often benefit platforms, not contributors.

9 From Artistic Desire to Legal Protocol

CORPUS began not as a rights-management concept, but from a musical desire: at its core, it aims to enable experiences that traditional methods could never realize – and this creative ambition demanded a new kind of infrastructure. The protocol described here is the legal and technical foundation for that vision.

By shifting licensing and attribution to the input side, CORPUS creates a framework that is fair, transparent, and scalable. Contributors are rewarded for what they add to the corpus, with authenticity and originality at the center. Value is not tied to mass-market replication but to distinctive voices, unique styles, and cultural diversity – and crucially, these qualities translate directly into long-term royalty participation.

At the same time, the protocol opens entirely new markets where conventional music production cannot scale: in-vehicle sound personalization, therapeutic music in dementia care, interactive soundscapes in games and XR, or exploratory learning tools in education. These domains are larger than the markets being disrupted today, and they require exactly what CORPUS is designed to provide: licensed, diverse, and context-responsive training data.

This dual motivation – artistic and industrial – is what drives CORPUS. It is both a practical response to legal and economic realities and an offensive move: a chance to build infrastructure that not only withstands disruption but creates new cultural and commercial value. The question at stake is not just whether music survives AI, but how we choose to shape its future.

The path ahead is complex. Many features described here – from automated quality evaluation to community governance – will take years to develop, test, and refine. But the direction is clear. CORPUS will advance step by step: assembling the dataset under provisional licenses, piloting real-world applications with industry partners, and refining the attribution system with community input.

During an industry event, a lawyer from Buma/Stemra listened to a short explanation of CORPUS. His reaction was immediate: “The logic is compelling,” he said with a smile, “but I wouldn’t want to be the one implementing it.” The remark captures the core challenge: entrenched systems resist transformation, even when the legal and economic rationale is clear. CORPUS is designed precisely for this task – to turn compelling logic into workable infrastructure.

Our goal is to bring together allies – contributors, investors, industry partners and institutions – who share this conviction: that music AI should be built not on exploitation or shortcuts, but on trust, consent, and artistic integrity. If we succeed, CORPUS will not only set a new licensing standard but also create a community where originality is rewarded, where exchange is authentic, and where technology helps us rediscover why we started making music in the first place.

9.1 Next Steps

CORPUS is currently moving from its alpha stage into a closed beta phase. Starting in December 2025, selected contributors will be onboarded on an invite-only basis to test the platform's submission and attribution workflows. A broader rollout is planned for early 2026, following evaluation of the beta phase.

In parallel, Sofilab is developing REEF, CORPUS's real-time adaptive model for interactive sound applications, and is establishing partnerships for industrial pilot projects in sectors such as automotive, healthcare, and music technology. To accelerate corpus growth, collaborations with existing music libraries, labels, and archives are being prepared.

CORPUS is now opening its network to musicians, rights holders, and industrial partners who wish to contribute to and shape the foundation of a lawful, scalable Music–AI ecosystem.