March 2026

ISRC Clustering: Achieving Metadata Integrity at Scale

Music Story White Paper

Executive Summary

In a global ecosystem where over **200 million tracks** are available on streaming services and approximately **100,000 new ISRCs are delivered to platforms every day**, metadata fragmentation remains a primary obstacle for music platforms and rights holders.

The Solution: Music Story's **ISRC Clustering** solution provides a definitive recording reference by aggregating, matching, and deduplicating data across heterogeneous sources. By resolving multiple ISRCs into unified clusters, we enable partners to eliminate catalog fragmentation, ensure accurate rights attribution, and optimize revenue collection.

---

1. The Challenge: The Cost of Fragmentation

Managing a modern music catalog is increasingly complex, with an average of **6–7 ISRCs per recording**.

- **Data Silos:** Recording data is often scattered across various partner catalogs and internal repositories.
- **Attribution Gaps:** Discrepancies in metadata lead to missed royalties for Collective Management Organizations (CMOs) and Performing Rights Organizations (PROs).
- **User Experience:** Fragmented catalogs result in duplicate entries for users on Streaming Platforms (DSPs).

2. Strategic Benefits: Why Clustering Matters

We solve specific pain points across three core industry segments:

2.1. For CMOs & PROs: Revenue Maximization

- **Reporting Accuracy:** Cross-reference ISRCs reported by DSPs with internal databases to ensure accurate royalty attribution for members.
- **Enhanced Collection:** Identify "hidden" ISRCs via recording/publishing matches to improve the efficiency of collection cycles.

2.2. For Publishers: Rights & Claims

- **Ownership Tracking:** Detect work ownership across multiple recording versions to effectively assert claims.
- **Verification:** Utilize Music Story as a source of reliable and comprehensive information to validate claims and ownership positions.

2.3. For DSPs & Platforms: Catalog Management

- **Catalog Deduplication:** Consolidate fragmented recordings into a single "Canonical Recording" to improve search and user navigation.
- **Active Management:** Simplify the grouping of back-catalog and new releases for cleaner metadata.

3. The Music Story Methodology: Data Enrichment & Integrity

Our process transforms raw metadata into a high-integrity repository through a multi-layered approach.

3.1. Continuous Ingestion & Scalability

Music Story performs continuous ingestion of recording catalogs from over a dozen contracted partners and clients. This dual-purpose workflow cleans and enriches client catalogs while reinforcing the Music Story reference database.

- **Volume:** Our repository currently processes **250 million ISRCs**.
- **Velocity:** The system scales by approximately **20,000 new ISRCs every day**.

3.2. The Structural Matching Backbone

This continuous ingestion is supported by a matching pipeline that reconciles disparate catalog inputs into a unified reference model across Recording, Artist, and Album levels. Our **composite scoring architecture** utilizes automated canonical resolution (including multi-language support for Japanese and Korean), fuzzy string similarity, and context-aware signals - such as ISNI, UPC, and discographic consistency - to verify data proximity.

For complex cases where automated logic meets ambiguity, our **expert editorial team** provides manual curation, ensuring the high-fidelity integrity required for partners' metadata.

3.3. The Clustering Process

Clustering occurs in two distinct stages to move from raw data to a "ground truth".

1. **Iterative Aggregation:** We aggregate all ISRCs for a single recording using semantic keys such as duration, title/subtitle, and release membership.
2. **Assignment & Weighting:** ISRCs are assigned to specific clusters based on source (audio fingerprints, lyrics, publishing data) and internal semantic rules.

To learn more about the clustering process, please see appendix 1.

4. ISRC Cluster Taxonomy

Music Story provides three primary cluster types to serve different business needs:

Cluster Type	Definition	Primary Use Case
Phonogram Cluster	Identical or nearly identical audio recordings, based on similar audio fingerprints	Master-level deduplication (e.g. Remasters)

Artist-Title Cluster	Grouping recordings of the same title by the same artist across releases	Discography organization (e.g. Radio Edit, Live)
Work Cluster	Grouping recordings of the same musical work	Publishing (e.g. reinterpretation, covers and remixes)

5. Technical Integration & Delivery

Our clusters are designed for seamless integration into existing tech stacks.

- Immediate Access: Available via the **Metadata Monitor dashboard** and our dedicated **API endpoint**.
- **Data Schema:** Each delivery includes a unique Cluster ID, the Cluster Type, the Music Story Recording ID, and the associated ISRCs.

See the API detailed fields in the appendix 2.

6. Roadmap: The Future of ISRC Intelligence

Music Story is committed to continuous improvement of data integrity.

- **Hybrid Verification:** Moving toward a model where automatic clustering is supplemented by manual, verified clusters for high-priority projects.
- **Extended Coverage:** Increasing ISRC mapping for recordings outside our primary repository.
- **AI Integration:** Developing a proprietary **AI Audio model** to refine clustering accuracy via advanced audio analysis.

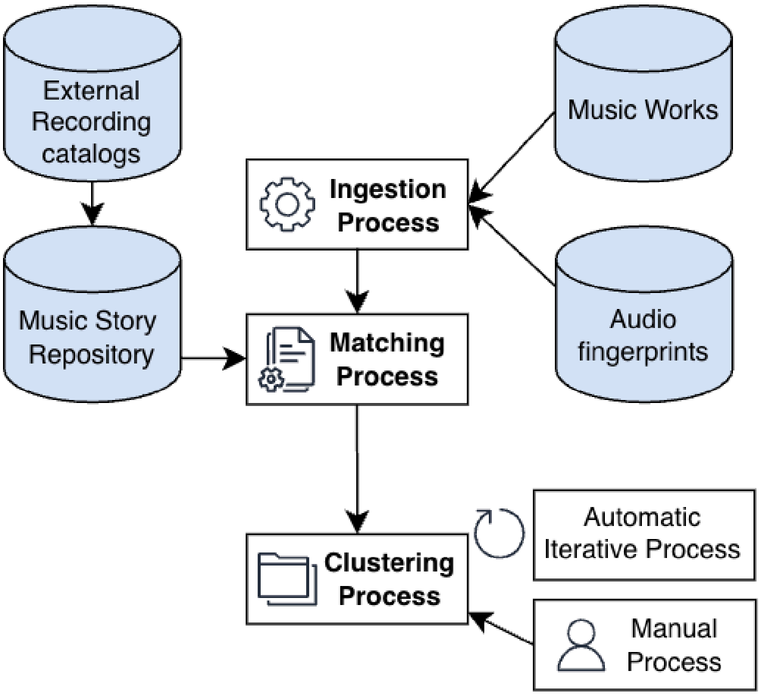
APPENDIX 1

Technical Deep Dive – The Clustering Process

A. Strategic Objective & Multi-Source Intelligence

The clustering framework groups ISRC entries referring to the same recording to consolidate fragmented representations. Unlike standard ingestion, this process leverages external reference sources, including **musical works** and **audio fingerprint databases**, to identify related recordings beyond what is visible in isolated catalogs.

Overview of the Ingestion and Clustering Workflow



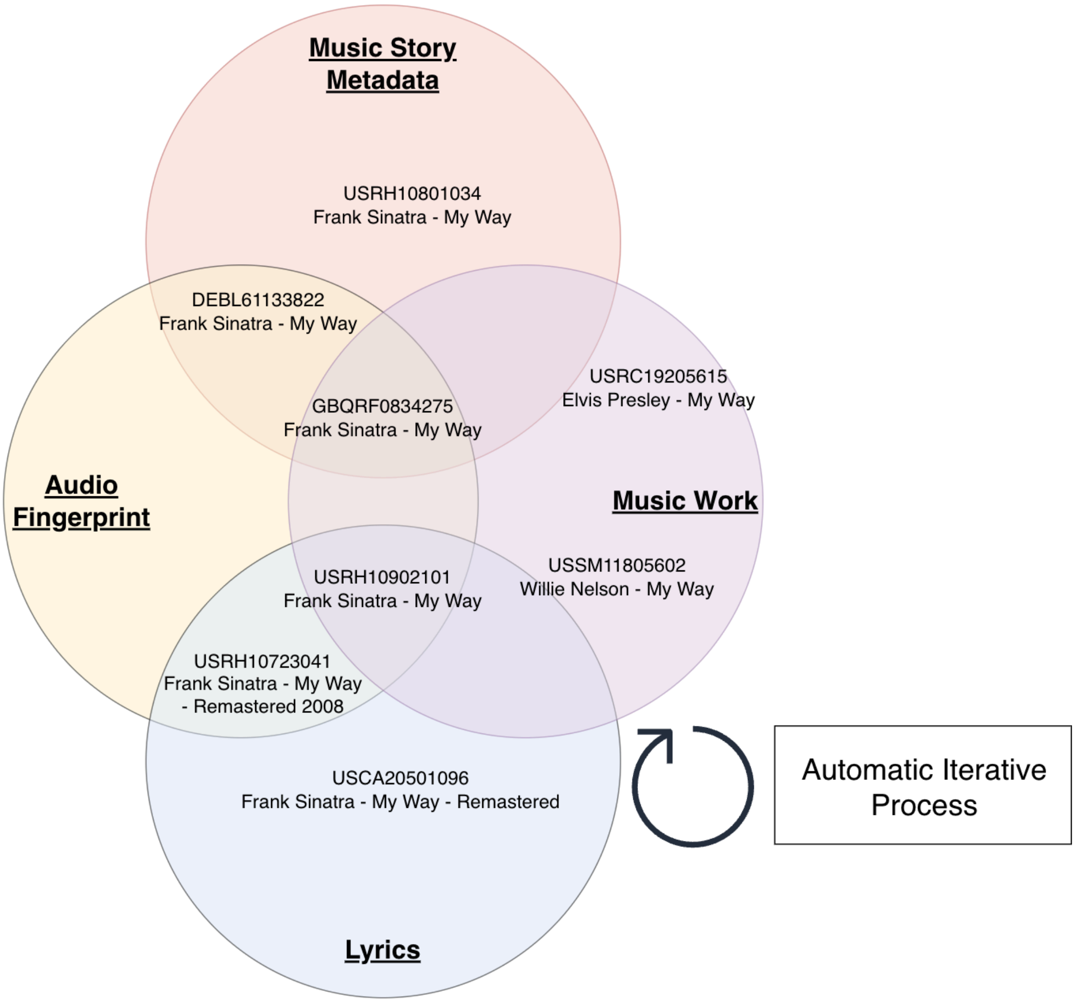
B. The Clustering Workflow

The workflow operates through a multi-stage aggregation and validation process:

1. **Multi-Criteria Clustering:** Recordings are initially grouped based on proximity criteria (such as shared musical works or audio fingerprint similarity).

2. **Quality Control & Cluster Partitioning:** When contradictions are detected (for example, recordings with identical fingerprints but differing durations), the cluster is split to maintain data integrity.
3. **Iterative Propagation:** The framework merges all clusters sharing a common ISRC, reconciling disparate metadata signals into a unique Cluster ID.

Overview of Clustering process with ISRC Sources



APPENDIX 2

The Clustering Dataset API Endpoints

Dataset	Field Name	Description
cluster	id	ID of the cluster
cluster	type	Type of the cluster : ● *work*: Similar works ● *artist_title*: Similar artists and titles ● *phonogram*: Similar phonograms ● *candidate*: Recordings outside of the Music Story repository with a proximity score to the cluster
cluster	recordings	Array of object contains all recordings of the cluster
cluster	recordings.id	ID of the recording
cluster	recordings.isrc	ISRC of the recording
cluster	recordings.title	Title of the recording
cluster	recordings.subtitle	Subtitle of the recording
cluster	recordings.duration	Duration of the recording
cluster	recordings.artists	Array of object contains all artists of the recording
cluster	recordings.artists.id	ID of the artist
cluster	recordings.artists.name	Name of the artist
cluster	recordings.artists.type	Type of the artist