DIJON Project Status Report
Leveraging Structured Repertoire for Scalable Harmonic and Formal Analysis

Peter Mynett

February 20, 2026

Abstract

This project investigates how harmony and musical form can be analyzed, compared, and visualized across performances using a structured Music Information Retrieval (MIR) pipeline. By leveraging genres with standard repertoire and readily available lead sheets, the project aims to enable scalable harmonic and formal comparison across performances. The system aims to integrate canonical data layers, external annotation tools, and computational feature extraction to support exploratory and comparative computational musicology.

1 Project Overview and Central Thesis

Leveraging genres with shared standard repertoire and accessible symbolic representations enables scalable harmonic and formal comparison across performances. Jazz and bluegrass provide an advantageous research setting: they combine repeated repertoire, interpretive variation, and lead sheets. This allows harmonic and formal information to be inferred even when audio recordings are weakly labeled or unlabeled.

The long-term objective is to design a reproducible framework that supports comparative analysis, clustering, and visualization of harmony and form across multiple performances.

2 Data Model and Multi-Tier Truth Structure

A core design principle is a three-tier data model:

- **Gold**: A curated set (10–15 songs) with dense manual annotation including formal markers, precise chord progressions, tempo information, and structural segmentation. This tier supports controlled experimentation and algorithm refinement.

- **Silver**: A larger set (50–100 songs) where recordings are associated with known lead sheets, and

have a high volume of material available on youtube. Although temporally unaligned, these provide harmonic and formal priors at scale, enabling broader comparative analysis.

- **Bronze**: Unlabeled recordings collected using the custom yt-dlp wrapper. These enable large-scale exploratory visualization once representations have been refined using Gold and Silver tiers.

- **Leading**: A temporary small curated set of songs to lead the proof-of-concept.

The gold data supports pipeline development and experimentation. Silver data leverages shared repertoire and symbolic priors to scale harmonic inference. Bronze data allows higher volume visualization and exploratory clustering.

3 System Architecture

The system is CLI-driven with notebook-based exploratory layers. Canonical data is separated from exploratory artifacts:

- **Acquisition and Raw Layer**: External audio and symbolic representations are ingested into a canonical raw layer.

- **Annotation Layer**: Manual annotations (e.g., formal markers from Reaper, lead sheets from MuseScore).

- **Derived Layer**: Computed features such as novelty functions, beat maps, chromagrams, and other harmonic are formal representations.

- **Notebook Layer**: Jupyter notebooks to support exploratory analysis and demonstration.

External tools are integrated strategically. Reaper enables rapid formal annotation. MuseScore supports creation and inspection of MusicXML lead sheets. A custom wrapper around yt-dlp allows efficient repertoire-linked recording acquisition.

4 Work Completed

Current progress includes:

- Partial implementation of a CLI system for acquisition and ingestion pipeline.

- Initial novelty and temporal representations.

- Annotation ingestion workflows for external tools (Reaper and MuseScore).

- Scaffolding for reproducible filesystem structure distinguishing raw, annotated, and derived data.

- Initial development of Gold and Leading dataset subsets.

The project currently provides a basic pipeline for acquiring and ingesting data, and an initial demonstration of derived data.

5 Planned Development

5.1 Representation Refinement

- Experimenting with and strengthening temporal feature extraction processes.

- Developing harmonic and formal representations.

- Linking derived data to lead sheets.

5.2 Comparative Analysis

- Designing similarity measures for harmonic progression and form.

- Exploring dimensionality reduction techniques for structural comparison.

- Investigating clustering across performances of same or similar repertoire.

5.3 Visualization Layer

A primary research goal is to explore new visualization approaches for harmonic and formal relationships.

- Harmonic progression heatmaps across performances.

- Form-aligned structural comparison.

- Low-dimensional embeddings of chord progression or form vectors.

The final visualization design remains unknown and will be explored as representations are developed.

6 Evaluation Strategy

Evaluation remains an open research question. Potential evaluations include:

- Coherence between audio-derived structure and symbolic priors.

- Clustering across shared repertoire.

- Usefulness of generated visualizations.

- Insights that are not readily apparent through traditional inspection.

The goal is to move from infrastructure development toward computational musicology: using structured data to reveal relationships between performances and songs that may not be immediately obvious.

7 Expected Contributions

The project aims to contribute:

- A reproducible MIR framework tailored to repertoire-driven analysis.

- A multi-tier annotated dataset leveraging shared standard repertoire.

- Novel visualization approaches for harmonic and formal comparison.