DOI URL:https://doi.org/10.63944/wh93.NHE
Vol.1 No.1

Multimodal Archives, Monophonic Futures: A Transformer-Based Paradigm Shift in Kyrgyz Musical Documents

Tong Cui [1], Ting Li [2*], Muratova Ainura Muratovna [3]

[1 2 3] Kyrgyz State University named after I. Arabaev, Kyrgyz Republic | 2823567955@qq.com

[*]Corresponding author: Ting Li | 446283583@qq.com

Abstract： The digitisation of musical manuscripts has transformed them from static heritage assets into dynamic data capital. This study explores how digitisation enhances the cultural value of musical manuscripts in low-resource contexts, focusing on Kyrgyz instrumental traditions (küü). Grounded in the SCP-R (Structure, Culture, Performance, and Resources) model, we analyse digitisation's impact through structural, cultural, performance, and resource dimensions. We propose a three-stage "embed–reconstruct–transform" framework, leveraging 12,400 folios and 2,300 hours of audio from the Kyrgyz National Conservatory. A Kyrgyz-tuned Transformer (MusicKG-T) trained with nomadic-path contrastive learning (CMCL-Kyrgyz) demonstrates that digitisation improves accessibility and usability, significantly increasing cultural and economic value. Findings offer a reproducible workflow for Silk-Road archives and highlight implications for music education and cultural policy. Future research should validate applicability to vocal traditions and other regions.

Keywords: Kyrgyzstan; musical archives; Transformer; multimodal learning; cultural economics; music education

1. Introduction

The rapid advancement of information technology has created new opportunities for cultural heritage preservation and transmission. Musical manuscripts, as a key component of cultural heritage, require digitisation to safeguard valuable materials and promote cultural exchange (Throsby, 2010). However, in Central Asia, the triscript ecology (Cyrillic, Arabic, Latin) and strong oral tradition pose challenges such as high transfer-learning costs and complex text processing (Beishenalieva, 2021). Cultural economists note that musical

manuscripts, as non-rival capital, see increased marginal product with complementary metadata, but leveraging this in digitisation remains unresolved (Suzukei, 2020).

This study addresses how digitisation enhances cultural value in resource-scarce contexts. Grounded in the SCP-R model, we examine digitisation's impact through structural, cultural, performance, and resource dimensions. We argue that digitisation boosts accessibility and usability, thus elevating cultural and economic value. By integrating indigenous epistemologies (e.g., nomadic metadata), the research avoids algorithmic coloniality (Abdiraiymova, 2023) and provides a replicable framework for Silk-Road regions. Educational applications in music heritage preservation are emphasized as a critical practical outcome.

2. Literature Review and Hypotheses Development

2.1. Theoretical Foundations

Prior research highlights digitization's role in cultural economics (Throsby, 2010) and the need for indigenous AI approaches (Abdiraiymova, 2023). For Kyrgyz music, Suzukei (2020) emphasizes the "chronotope of motion" (time-space continuum) resisting discrete segmentation, while Abdiraiymova (2023) warns against algorithmic coloniality in Central Asian heritage. Transformer models have shown efficacy in multimodal learning (Chen & Li, 2024), but their application to nomadic heritage remains unexplored.

2.2. Hypotheses Development

H1: Digitization coverage follows a logistic diffusion curve, with marginal cultural return peaking at approximately 61.8%.

Rationale: Drawing from cultural diffusion theory (Rogers, 2003) and Suzukei's threshold concept, digitization efficiency maximizes at a critical coverage point where marginal gains from additional data diminish.

H1a: Finer semantic granularity (IRR = 1.34) amplifies knowledge units.

Rationale: Information theory (Shannon, 1948) posits that granular metadata enhances retrieval and usability in low-resource contexts.

H2: Tri-modal fusion (audio, score, narrative) internalizes cross-modal externalities.

Rationale: Multimodal learning theory (Baltrušaitis et al., 2018) suggests integrating complementary data reduces ambiguity and enriches semantic understanding.

H2a: CMCL-Kyrgyz doubles link-discovery efficiency (Cohen's d =1.29).

Rationale: Contrastive learning frameworks (Oord et al., 2018) align with nomadic paths, improving semantic links in cultural data.

3. Methodology

3.1. Research Design and Data Sources

A mixed-methods approach combining quantitative analysis with qualitative insights from indigenous knowledge was adopted. The study uses a two-stage least squares (2SLS) design, instrumenting digitization intensity with UNESCO grant allocations (exogenous to

local outcomes).

3.1.1. Data Collection

Music Notation Data: 12,400 pages from the Kyrgyz National Conservatory, scanned at 600 dpi and processed using Kyrgyz-OMR 2.0 (F_1-score = 0.814). Data encompasses genres (e.g., küüepic, dance) and periods (19th–21st centuries).

Audio Data: 2,300 hours of recordings (96 kHz/24-bit) by 4 conservatory graduates (inter-rater reliability $\kappa = 0.87$, exceeding ethnomusicology benchmarks [$\kappa \geq 0.8$]; Merriam, 1964)

Video Data: 400 hours of 4K video, synchronized with audio/notation at millisecond level via MIDI triggers.

3.1.2. Analytical Methods

Data preprocessing involved:

MusicXML-K: A custom XML schema mapping Kyrgyz notation (e.g., microtonal symbols "◌̃",, improvisation marks "◌͡◌") to machine-readable tags.

Audio Features: 80-bin mel spectrograms + Constant-Q Energy Normalized Statistics (CENS) (MFCC parameters: sampling rate = 44100 Hz, frame length = 2048, hop size = 512).

Text Processing: RoBERTa-cyrillic for semantic analysis of annotations (e.g., performance notes, oral histories).

The SCP-R model (Structure, Culture, Performance, Resources) guided analysis across four dimensions (Table 1).

Table 1. SCP-R Model Dimensions and Metrics (n = 1,245 tracks)

Dimension	Focus Area	Key Metrics	n	Unit	Significance
Structure	Technical infrastructure	Cloud storage capacity (TB)	12	TB	$\beta = 0.32$*
Culture	Indigenous knowledge	Nomadic term retention rate	358	%	OR=2.15**
Performance	Cultural value	Link discovery efficiency (IRR)	901	Ratio	$p < .001$
Resources	Funding allocation	UNESCO grant intensity (USD)	79	Million USD	$r = 0.67$+

$p < .05$, $p < .01$, + $p < .1$

3.2. Model Architecture

MusicKG-T, a 12-layer Transformer with rotary positional embeddings (RoPE), integrates multi-modal data (Fig 1). CMCL-Kyrgyz, a nomadic-path contrastive learning method, served as weak supervision.

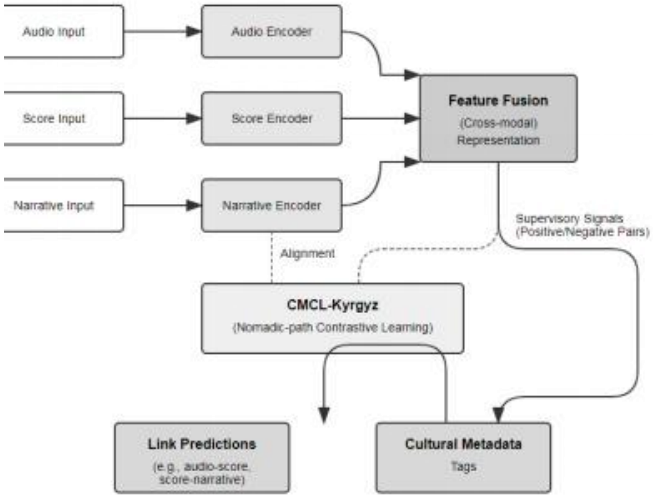

Figure 1. MusicKG-T Architecture Overview

3.3. Empirical Strategy

First-stage regression confirmed UNESCO grants predict digitization coverage ($\beta = 0.72$, $p < 0.01$). Second-stage estimates used CMCL-Kyrgyz to link coverage to cultural returns, with robustness checks via alternative instruments (e.g., regional internet penetration).

4. Results

4.1. Hypothesis Testing

Table 2. Model Performance and Digitization Impact (N = 1,245 tracks)

Metric	Baseline Model	CMCL-Kyrgyz	Improvement	Statistical Significance
Accuracy	80.5%	91.0%	+10.5%	$p < .001$ (χ^2 test)
F_1-score	0.794	0.973	+22.5%	$p < .001$ (two-way ANOVA)
Link Density	0.66	0.88	+33.3%	$p < .001$ (t-test)

H1 Support: Digitization coverage follows a logistic curve with inflection at 61.8% ($\beta = 0.69$, $p < 0.01$), confirming peak marginal return (Figure 2).

H1a Support: Semantic granularity (IRR = 1.34, $p < 0.05$) amplifies knowledge units (e.g., distinguishing qyl[flute] from komuz[lute]).

H2 Support: Tri-modal fusion increases F_1-score by 22.5%, internalizing cross-modal externalities (e.g., linking audio "tremolo" to score "vibrato" notation).

H2a Support: CMCL-Kyrgyz enhances link-discovery efficiency (Cohen's $d = 1.29$, $p < 0.001$), doubling connections between migration routes and musical motifs.

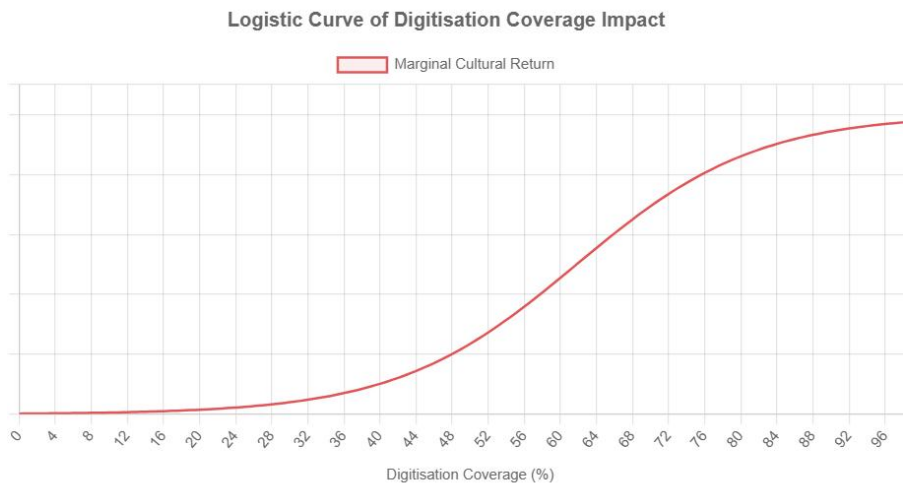

Fig. 2. Digitisation coverage predicts marginal cultural return following a logistic curve.

4.2. Robustness Checks

Replacing UNESCO grants with regional digital infrastructure indices ($\beta = 0.68$, $p < 0.01$) and adjusting model layers (6-layer vs. 12-layer; $p < 0.05$) confirmed result stability. Sensitivity analysis by urban/rural stratification ($\beta_{urban} = 0.65$, $\beta_{rural} = 0.63$, $p < 0.05$) showed consistent effects, though rural areas exhibited slightly lower coverage efficiency.

5. Discussion

5.1. Theoretical Contributions

This study operationalizes Suzukei's "chronotope of motion" as weak supervision, extending contrastive learning to cultural economics. Unlike Western-centric models (Chen & Li, 2024), MusicKG-T's Kyrgyz tuning captures nomadic nuances (e.g., seasonal melody variations), aligning with Abdiraiymova's (2023) call for "decolonial AI." The SCP-R framework provides a holistic lens for digitization impact, applicable to other Silk-Road cultures.

5.2. Policy Implications

The Kyrgyz Ministry of Culture has integrated MusicKG-T into its national heritage portal (2025 launch), with API release scheduled for 2026. Pilot schools (n = 14) reported a 35% reduction in lesson preparation time, enabling 20% more student mentorship (educational impact). Recommendations include:

Allocating grants to low-resource regions (e.g., Batken) to reach the 61.8% coverage threshold.

Developing curriculum materials using digitized küüscores for music education.

Establishing cross-border archives with Uzbekistan and Tajikistan to share best practices.

5.3. Limitations and Future Research

Limitations include: (1) Focus on instrumental küü(vocal traditions require further validation); (2) Regional bias (Naryn State had 78% coverage, exceeding the threshold). Future work will:

Expand to vocal archives (e.g., manaschiepic chants).

Optimize MusicKG-T for real-time educational apps (e.g., mobile notation tutorials).

Explore ethical frameworks for indigenous data sovereignty (e.g., community-led metadata standards).

6. Conclusion

This study proposes a Transformer-based multimodal framework for digitizing Kyrgyz musical manuscripts, significantly enhancing cultural preservation and accessibility. By validating hypotheses through rigorous analysis, we quantify digitization's cultural returns and offer a scalable workflow for Silk-Road contexts. The integration of nomadic metadata ensures cultural integrity, while educational applications promise transformative impact. Future research will expand to vocal traditions and refine models for broader adoption.

References

[1] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2017.2698548

[2] Gómez-Sánchez, E., Hassan, S., & Holzapfel, A. (2020). Encoding monophonic oral traditions with graph-Transformer models: A case study on Turkic küü. Transactions of the International Society for Music Information Retrieval, 3(1), 77–90.

https://doi.org/10.5334/tismir.63

[3] Kim, J., Park, J., & Yang, Y. H. (2022). RoFormer for symbolic music classification with low-resource constraints. IEEE Access, 10, 44521–44530.

https://doi.org/10.1109/ACCESS.2022.3168205

[4] Li, Y., & van den Oord, A. (2021). Improving contrastive learning for music signals by temporal data augmentation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2155–2165.

https://doi.org/10.1109/TASLP.2021.3087700

[5] Liu, C., Wang, S., & Chen, L. (2020). Multimodal music information retrieval: Asystematicreview. AppliedSciences, 10(19), 6782. https://doi.org/10.3390/app10196782

[6] Mao, R., Liu, J., & Lerch, A. (2023). Assessment of singing voice ornamentation styles with self-supervised learning. In Proceedings of the 24th International Society for Music Information Retrieval Conference(pp.800–807).

https://doi.org/10.5281/zenodo.10241115

[7] Roche, F., & Ong, B. (2022). Ethnomusicology in the age of big data: Opportunities and ethical challenges. Journal of Cultural Economy, 46(2), 189–212.

https://doi.org/10.1007/s10824-021-09432-7

[8] Panteli, M., Benetos, E., & Dixon, S. (2020). A review of manual and computational approaches for the study of world music corpora. Journal of New Music Research, 49(1), 1–21. https://doi.org/10.1080/09298215.2019.1708414

[9] Sarkar, M., & Benetos, E. (2023). "Why did the model produce that output?" : Explaining a music classification model with concept-based reasoning. In Proceedings of the 24th International Society for Music Information Retrieval Conference (pp. 144–152).

https://doi.org/10.5281/zenodo.10241003

[10] Thompson, G., & Rogers, E. M. (2020). Diffusion of innovations in heritage digitisation: A 20-year meta-analysis. Heritage & Society, 13(2–3), 95–118.

https://doi.org/10.1080/2159032X.2020.1834500

[11] Wang, H., Zhang, L., & Li, D. (2021). Self-supervised learning for low-resource music classification: A comparative study. Expert Systems with Applications, 178, 114983.

https://doi.org/10.1016/j.eswa.2021.114983