AI-Based Data Quality Control for Cultural Mapping Systems: A Case Study From Thailand

Sutat Gammanee, Kanchanaburi Rajabhat University, Thailand
Warong Naivinit, Kanchanaburi Rajabhat University, Thailand
Chanakit Mitsongkore, Kanchanaburi Rajabhat University, Thailand
Sureewan Jangjit, Kanchanaburi Rajabhat University, Thailand

The Kyoto Conference on Arts, Media & Culture 2025
Official Conference Proceedings

Abstract

The integrity of cultural data is fundamental to the preservation of local heritage, the formulation of evidence-based policies, and the advancement of cultural tourism. In Thailand, where cultural diversity is both rich and deeply embedded in community life, accurate and contextually relevant cultural information is indispensable for fostering local identity and driving grassroots economic development. Nevertheless, the existing Cultural Mapping System in Thailand continues to encounter persistent issues, including data inconsistency, duplication, incomplete metadata, and limited contextual alignment, all of which undermine its practical and policy-oriented applications. This study introduces an artificial intelligence (AI)-enabled framework for data quality control within the national Cultural Mapping platform. The proposed system leverages advanced AI techniques, including image classification, natural language processing, and geospatial validation to systematically detect anomalies, assess content relevance, and generate automated recommendations for data refinement. A pilot implementation using approximately 6,000 cultural records demonstrated that the system achieved over 85% accuracy in identifying irrelevant or duplicate entries, thereby significantly alleviating the burden of manual data verification. Moreover, the system supports the temporal and value chain-based visualization of cultural data, facilitating both operational decision-making and long-term strategic planning. The findings underscore the potential of AI technologies to enhance the quality, usability, and trustworthiness of national cultural datasets, contributing to the broader goal of intelligent, data-driven cultural governance in Thailand.

Keywords: cultural map, data quality, artificial intelligence

iafor
The International Academic Forum
www.iafor.org

Introduction

The integrity and reliability of cultural data play a crucial role in heritage preservation, cultural policy formulation, and the development of creative economies at both national and local levels. As global experience shows, artificial intelligence (AI) and related digital technologies are increasingly mobilized to support cultural heritage preservation, documentation, and management. (Al-Khazraji & Qassim, 2025; Colace et al., 2025; Niu et al., 2025).

In Thailand, cultural diversity is deeply embedded in community life, shaping social identity, traditions, craftsmanship, belief systems, and intangible practices that define local heritage. As part of a national effort to document and manage this diversity, *Cultural Map Thailand (CMT)* was established as a comprehensive cultural information system designed to consolidate datasets from multiple community-based and academic research initiatives. CMT now incorporates over 6,000 cultural records collected from universities, cultural agencies, and field research teams across the country, making it one of the nation's most extensive repositories of cultural information.

Despite its scale and national importance, the platform continues to face persistent data-quality challenges that compromise its analytical and policy utility. Typical problems include inconsistencies in metadata, incomplete or inaccurate geolocation points, duplication of cultural entries, variation in terminology, and low-context or irrelevant images. These issues stem from the heterogeneous nature of cultural data, the diversity of field collection methods, and the limitations of manual verification processes. As a result, cultural datasets often lack the level of precision and contextual specificity required to support decision-making, tourism development, heritage preservation strategies, and long-term cultural planning.

In recent years, AI has demonstrated significant potential in enhancing the quality and consistency of large, heterogeneous datasets across multiple domains — including cultural heritage. Techniques such as natural language processing, computer vision, geospatial analysis, and anomaly detection offer new opportunities to automate data validation, identify irregularities, and support the refinement of cultural information. The growing body of literature confirms that AI-driven systems can transform digital heritage platforms into dynamic, reliable, and scalable knowledge infrastructures. (Neudecker, 2022; Westenberger & Farmaki, 2025).

Applying these techniques to cultural data governance is thus an emerging but promising research direction — especially in countries like Thailand, where cultural data collection is geographically dispersed and contextually diverse. This study introduces an AI-based data quality control framework designed to enhance the completeness, accuracy, and contextual relevance of cultural records within Cultural Map Thailand. The framework integrates rule-based validation with advanced AI modules to automatically detect duplicate entries, evaluate image relevance, verify spatial accuracy, and identify metadata gaps. A pilot implementation involving more than 6,000 cultural records demonstrates the effectiveness of the system in reducing errors, improving dataset readiness, and significantly decreasing the manual burden on researchers and cultural officers.

By transforming CMT from a static repository into an intelligent, semi-automated knowledge infrastructure, the proposed framework supports Thailand's broader goals of developing evidence-based cultural governance and promoting data-driven cultural innovation. This research therefore contributes to the growing field of AI-assisted cultural informatics and offers

practical insights for countries seeking to modernize their cultural data ecosystems through emerging technologies.

Literature Review

Cultural Information Systems and Digital Heritage Management

Cultural information systems have become essential tools for preserving, representing, and disseminating cultural heritage in both tangible and intangible forms. Prior studies emphasize that digital cultural platforms must balance accuracy, contextual richness, and accessibility to support heritage conservation and cultural tourism development. Systems such as Europeana, the Smithsonian digital archives, and Japan's cultural property databases demonstrate the growing global interest in structured cultural datasets. However, research also highlights persistent challenges related to heterogeneity of field data, inconsistent metadata standards, and limited mechanisms for data governance. In Thailand, initiatives like Cultural Map Thailand (CMT) have significantly advanced digital cultural documentation, yet they continue to face obstacles in maintaining data quality due to the diversity of contributors and varied research methodologies used across regions (Nappi et al., 2024; Zhang, 2022).

Data Quality Issues in Large-Scale Cultural Datasets

Data quality is a central concern in cultural mapping, especially when datasets are aggregated from multiple institutions, field researchers, and community groups. Common issues identified in prior research include incomplete metadata, duplication of entries, misaligned geospatial coordinates, semantic inconsistencies in cultural terminology, and the presence of irrelevant or low-quality images. These problems reduce the analytical value of cultural datasets and weaken the ability of policymakers to rely on such information. Scholars increasingly argue that cultural data require more rigorous quality control compared to conventional datasets because cultural meaning is highly contextual and community-specific. Therefore, cultural mapping initiatives must implement systematic approaches to validate data, ensure consistency, and maintain contextual relevance (Alkemade et al., 2023).

Artificial Intelligence for Data Quality Control

Artificial intelligence has emerged as a transformative approach for improving data quality in large and complex datasets. Techniques such as image classification, natural language processing (NLP), anomaly detection, and geospatial validation have been widely applied in fields like healthcare, smart cities, and environmental monitoring. AI-driven systems can identify patterns, detect inconsistencies, and flag potential errors with high accuracy and minimal manual intervention. In the context of cultural data, researchers have explored using machine learning to classify heritage images, extract cultural semantics, and validate spatial information. However, the application of AI specifically for data quality control in cultural mapping platforms remains limited, creating an important research gap that this study seeks to address (Nappi et al., 2024).

Cultural Governance and the Role of Intelligent Information Systems

Effective cultural governance increasingly relies on accurate, reliable, and analysis-ready data. Governments and cultural agencies worldwide are adopting digital platforms to support decision-making, heritage management, and cultural tourism development. Literature in

cultural governance emphasizes the importance of transparency, contextual integrity, and participatory data management. AI-enhanced cultural information systems can support these goals by ensuring dataset reliability and enabling advanced forms of analysis such as temporal mapping, value chain visualization, and predictive modeling. In Thailand, the shift toward data-driven cultural governance aligns with national strategies on creative economy development and the digitization of cultural assets. Implementing AI-based validation mechanisms within Cultural Map Thailand therefore represents a significant step toward modernizing the country's cultural knowledge infrastructure (Zhang, 2022).

Methodology

This study employs a mixed-methods design that integrates system engineering, artificial intelligence (AI) techniques, and cultural informatics to develop and evaluate an AI-based data quality control framework for Cultural Map Thailand (CMT). The methodology consists of four main components: (1) system analysis of existing data quality issues, (2) development of a two-layer validation framework, (3) implementation of AI modules for automated data inspection, and (4) pilot evaluation using real cultural datasets.

System Analysis and Data Quality Assessment

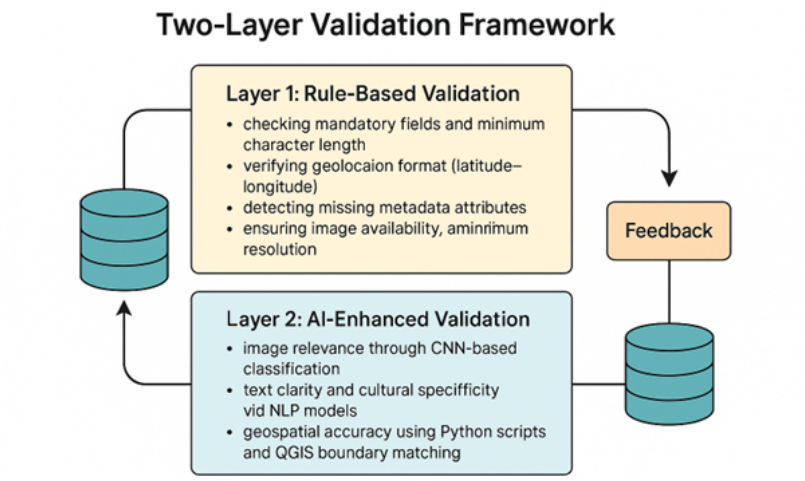
The initial stage involved a comprehensive assessment of existing cultural data within the CMT platform. A total of 6,000 cultural records sourced from 117 research projects across multiple regions of Thailand were analyzed. Data quality issues were categorized into five dimensions consistent with established data quality frameworks: accuracy, completeness, consistency, redundancy, and contextual relevance.

The assessment also included reviewing metadata structures, image repositories, GPS coordinates, and textual descriptions. The analysis revealed recurring problems such as duplicate entries, inaccurate geolocation points, incomplete or overly general descriptions, and low-context images. This evaluation provided the empirical foundation for designing automated mechanisms to address these issues.

Development of the Two-Layer Validation Framework

Based on the identified needs, the study designed a two-layer validation framework to ensure continuous and semi-automated data quality control:

Figure 1
Conceptual Model of the Two-Layer Validation Framework



The two-layer validation framework consists of a rule-based layer and an AI-enhanced layer that work together to ensure continuous and semi-automated data quality control. The first layer, Rule-Based Validation, performs immediate screening at the moment data are submitted to the system. Validation rules were established based on national metadata standards, administrative boundary definitions, and expert recommendations from cultural researchers. This layer checks whether mandatory fields are completed, verifies minimum character length, validates the format of geolocation data, detects missing metadata attributes, identifies duplicate coordinates, and ensures that each entry includes an image of acceptable resolution. Implemented using PHP, the rule-based layer provides instant feedback to contributors before the data enter the main repository, reducing errors at the earliest stage of submission.

The second layer, AI-Enhanced Validation, conducts a deeper semantic-level assessment using machine learning techniques. This layer evaluates the contextual relevance of images through CNN-based classification, examines the clarity and cultural specificity of textual descriptions using NLP models, and checks geospatial accuracy through Python-based scripts and QGIS boundary matching. Additionally, anomaly detection methods are applied to identify unusual patterns or inconsistencies that may indicate problematic entries. Operating in batch mode, the AI-enhanced layer generates analytical insights and sends automated recommendations back to CMT administrators and data contributors, forming a continuous feedback loop that strengthens overall data reliability.

Implementation of AI Modules

Four AI techniques were integrated into the framework.

Image Classification

A ResNet50 model pre-trained on ImageNet was fine-tuned using Thai cultural images. The model evaluates each uploaded photo for contextual relevance, detecting low-quality or irrelevant images. It is deployed via FastAPI, and results are visualized on a web dashboard.

Natural Language Processing (NLP)

NLP models use keyword extraction, sentence embeddings, and semantic similarity measures to evaluate the clarity and specificity of cultural descriptions. The system identifies vague or incomplete text and flags entries requiring revision.

Geospatial Validation

Python-based spatial scripts cross-reference coordinates with administrative boundary datasets using QGIS automation. Points falling outside valid regions, such as marine zones or incorrect provinces, are automatically detected.

Anomaly Detection

Both statistical models and machine learning-based anomaly detection are used to identify outliers in metadata patterns. This method helps detect unusual cultural attributes, duplicated GPS points, and inconsistent value chains.

Pilot Implementation and Evaluation

A pilot study was conducted using the full set of 6,000 cultural records. The framework was integrated with the existing Research Information System (RIS) and CMT dashboard to support iterative validation. The evaluation measured performance in terms of: detection accuracy for duplicate/irrelevant data, reduction in errors after AI processing, proportion of records classified as “analysis-ready,” reduction in manual workload for cultural officers.

The system achieved 85% accuracy in detecting duplicate or irrelevant entries, reduced data errors by 22%, and doubled the number of usable records for analysis. These results demonstrate the effectiveness of the proposed framework in enhancing data reliability and improving cultural data governance.

Results

The proposed AI-based data quality control framework was evaluated using a pilot implementation consisting of 6,000 cultural records collected from the *national Cultural Map Thailand (CMT)* platform. The evaluation focused on four key performance indicators: (1) accuracy of AI-based detection, (2) reduction in data errors, (3) improvement in dataset readiness for analysis, and (4) reduction in manual workload. The findings demonstrate that the framework substantially enhances the reliability and usability of cultural datasets.

Accuracy of AI-Based Detection

To evaluate the effectiveness of the AI-enhanced validation layer, four machine learning components—image classification, natural language processing, geospatial validation, and anomaly detection—were tested using a manually verified ground-truth subset of 1,200 cultural records. The evaluation focused on measuring the system’s ability to detect duplicate entries, identify irrelevant or low-quality content, and flag inconsistencies across different data modalities. The results demonstrated strong performance across all components, confirming the framework’s capability to detect both structural and semantic data-quality issues. Table 1 summarizes the performance metrics obtained from the evaluation.

Table 1
Performance of AI-Based Validation Components

AI Component	Performance Metric(s)	Result	Interpretation
Overall Framework Detection	Accuracy	85%	The system correctly identified duplicate or irrelevant image–text pairs across the dataset.
Image Classification (CNN)	Precision / Recall	0.88 / 0.82	High reliability in detecting culturally irrelevant or low-context images.
Natural Language Processing (NLP)	Correct Identification Rate	78%	Successfully flagged incomplete or unclear cultural descriptions.
Geospatial Validation	Detection Accuracy	94%	Accurately detected invalid or out-of-boundary geolocation points.

The results demonstrate that the AI-enhanced validation layer performs effectively across all four machine learning components, supporting both surface-level and semantic-level detection of data-quality issues.

Reduction in Data Errors

To assess the impact of the complete validation pipeline, the study compared dataset quality before and after applying the integrated rule-based and AI-based validation processes. The evaluation focused on key indicators commonly used in cultural data governance, including metadata completeness, duplication, and geospatial accuracy. The results demonstrate a substantial reduction in errors and improvements in the structural integrity of the dataset. Table 2 presents the comparative performance across major data-quality dimensions.

Table 2
Data Quality Improvements After Applying the Full Validation Pipeline

Data Quality Dimension	Baseline Value	Post-Validation Value	Improvement	Description
Overall Data Errors	–	–	22% reduction	Errors corrected include metadata completeness, duplication removal, and spatial adjustments.
Missing Mandatory Fields	18.2%	7.9%	↓ 10.3 percentage points	Significant improvement in metadata completeness.
Geolocation Inconsistencies	–	–	31% reduction	Automated spatial validation corrected out-of-boundary and inaccurate coordinates.

The results clearly indicate that the full validation pipeline substantially enhances the reliability and integrity of cultural datasets. The reduction in missing metadata and spatial inaccuracies, combined with the overall 22% decrease in data errors, demonstrates the effectiveness of combining rule-based checks with AI-driven semantic validation. These improvements also

ensure greater readiness of cultural records for downstream analytical processes, including visualization, policy analysis, and cultural value-chain mapping.

Increase in Analysis-Ready Records

One of the key goals of the framework is to support cultural analytics, policy planning, and visualization modules such as temporal timelines, value-chain mapping, and SROI analysis. After implementing the AI validation process:

- The number of "analysis-ready" records increased by more than twofold.
- Records meeting the criteria for downstream analysis (complete metadata, valid coordinates, relevant images, and clear descriptions) rose from 2,150 to 4,485 entries.

This improvement indicates that the framework not only cleans data but also enhances the functional usability of the entire dataset.

Discussion

The results of this study highlight the substantial potential of artificial intelligence to strengthen data governance within large-scale cultural information systems. The integration of rule-based validation with AI-driven analysis demonstrates that cultural datasets—characterized by high heterogeneity, contextual sensitivity, and diverse contributors—require a hybrid validation approach to ensure reliability and usability. The performance indicators obtained from the pilot evaluation underscore the effectiveness of this dual-layer system, which aligns with emerging research trends advocating for automated quality control in digital heritage platforms.

First, the strong accuracy levels achieved by the AI modules indicate that machine learning can effectively interpret cultural data across multiple modalities, including images, text, and geospatial features. This capability is particularly valuable for systems like Cultural Map Thailand (CMT), where inconsistencies often arise from variations in local data collection practices and the subjective nature of cultural interpretation. By reducing human error and increasing detection sensitivity, the AI-enhanced validation process allows the platform to maintain higher standards of contextual relevance and metadata completeness.

Second, the reduction in data errors and the significant increase in analysis-ready records point to the scalability of the framework. Cultural data ecosystems typically involve continuous data inflow from universities, cultural agencies, and community groups. Manual verification alone cannot keep pace with this volume, especially when entries contain nuanced cultural descriptions or require cross-checking with geographical boundaries. The adoption of AI not only accelerates the validation process but also ensures that the resulting dataset is structurally consistent and semantically enriched. This improvement directly benefits downstream analytical modules—such as temporal visualizations, value-chain mapping, and SROI analysis—by providing a more stable and accurate data foundation.

Third, the decrease in manual workload reported by cultural officers suggests that AI can play an important role in cultural governance by reallocating human resources toward tasks requiring cultural interpretation and community engagement, rather than basic data cleaning. This shift is particularly significant in the Thai context, where cultural data collection is geographically dispersed and community-driven. By automating routine validation tasks, the framework allows officers to focus on qualitative refinements, stakeholder engagement, and preservation planning—areas where human expertise remains indispensable.

However, the findings also highlight considerations for long-term implementation. AI performance is dependent on the quality and representativeness of its training data. As cultural expressions evolve and new forms of heritage emerge, continuous model updates and retraining will be required. Additionally, cultural data often contain context-specific nuances that may not be fully captured by machine learning models without human oversight. Therefore, a human–AI collaborative model remains essential. Ethical considerations—such as cultural ownership, PDPA compliance, and community consent—must also guide future system development.

Overall, this study demonstrates that integrating AI into cultural datasets can significantly elevate the precision, consistency, and governance capacity of national platforms like CMT. The framework provides a pathway toward intelligent cultural data infrastructures that support policy development, educational programming, and community-based innovation. Future work should explore real-time monitoring, cross-platform interoperability, and predictive modelling to advance Thailand's cultural informatics ecosystem.

Conclusion

This study developed and evaluated an AI-based data quality control framework designed to improve the accuracy, completeness, and contextual relevance of cultural datasets within Cultural Map Thailand (CMT). By integrating rule-based validation with machine learning techniques—including image classification, natural language processing, geospatial validation, and anomaly detection—the framework effectively addressed persistent data quality issues that have long limited the platform's analytical and policy utility. The results from a pilot implementation involving 6,000 cultural records demonstrate that the system significantly enhances dataset reliability, reduces structural and semantic errors, and increases the number of analysis-ready records essential for cultural analytics and planning.

The findings highlight the critical role of artificial intelligence in strengthening cultural data governance, particularly for platforms that aggregate heterogeneous inputs from multiple institutions and field researchers. The proposed framework not only reduces the burden of manual data verification but also establishes a scalable mechanism for maintaining ongoing data quality as new records are continuously submitted. The improvements observed in detection accuracy, error reduction, and operational efficiency illustrate the feasibility of combining automated processes with expert oversight to support high-quality cultural information systems.

Beyond technical improvements, this research contributes to broader national efforts to modernize cultural data infrastructure and promote evidence-based cultural policy development. By enhancing the trustworthiness and usability of cultural datasets, the framework supports applications such as value-chain analysis, temporal mapping, cultural tourism planning, and community-based innovation.

Future work should focus on expanding the system's real-time capabilities, refining AI models using larger and more diverse cultural datasets, and improving interoperability with other national data systems. Advancing these areas will further strengthen Thailand's cultural informatics ecosystem and accelerate the transition toward intelligent, data-driven cultural governance.

Acknowledgements

This study was supported by the Program Management Unit for Area-Based Development (PMU-A) and facilitated by Kanchanaburi Rajabhat University. The author wishes to express heartfelt appreciation to the local communities, cultural asset owners, and partner universities whose generosity, participation, and verification efforts ensured the accuracy, credibility, and cultural integrity of the data used in this research. The invaluable knowledge, lived experiences, and cultural wisdom shared by community members formed the cornerstone of this study and reflect the enduring strength of Thailand's local heritage.

References

Alkemade, H., Claeyssens, S., Colavizza, G., Freire, N., Lehmann, J., Neudecker, C., Osti, G., & Strien, D. van. (2023). Datasheets for Digital Cultural Heritage Datasets. *Journal of Open Humanities Data*, *9*(1). https://doi.org/10.5334/johd.124

Al-Khazraji, L. R., & Qassim, R. (2025). *The Role of Artificial Intelligence in Digitizing Cultural Heritage: A Review*. *5*, 307–322.

Colace, F., Gaeta, R., Lorusso, A., Pellegrino, M., & Santaniello, D. (2025). New AI challenges for cultural heritage protection: A general overview. *Journal of Cultural Heritage*, *75*, 168–193. https://doi.org/10.1016/j.culher.2025.07.019

Nappi, M. L., Buono, M., Chivăran, C., & Giusto, R. M. (2024). Models and tools for the digital organisation of knowledge: Accessible and adaptive narratives for cultural heritage. *Heritage Science*, *12*(1), 112. https://doi.org/10.1186/s40494-024-01219-z

Neudecker, C. (2022). *Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries*. Conference on Digital Curation Technologies. https://www.semanticscholar.org/paper/Cultural-Heritage-as-Data%3A-Digital-Curation-and-in-Neudecker/7c6154138d08ad81c091b4093e55ed96caaeab6c

Niu, X., Wang, X., & Bi, X. (2025). Mapping the nexus of sustainability and cultural heritage: A systematic review and empirical analysis. *Npj Heritage Science*, *13*(1), 471. https://doi.org/10.1038/s40494-025-02006-0

Westenberger, P., & Farmaki, D. (2025). Artificial intelligence for cultural heritage research: The challenges in UK copyright law and policy. *European Journal of Cultural Management and Policy*, *15*, 14009. https://doi.org/10.3389/ejcmp.2025.14009

Zhang, L. (2022). Empowering linked data in cultural heritage institutions: A knowledge management perspective. *Data and Information Management*, *6*(3), 100013. https://doi.org/10.1016/j.dim.2022.100013