

A Framework for Cleaning and Integrating Multimodal Data in Digital Cultural Experience Research

Andrei DANILA

Beia Consult International, Bucharest, Romania
andrei.danila@beia.ro

George SUCIU

Beia Consult International, Bucharest, Romania
george@beia.eu

Cristina DOBRE*

Beia Consult International, Bucharest, Romania
**Corresponding author, cristina.balaceanu@beia.ro*

Nicoleta CAPBUN

Beia Consult International, Bucharest, Romania
nicoleta.capbun@beia.ro

Abstract. *The rapid digital transformation of cultural institutions has generated increasing volumes of complex data related to visitor behaviour, emotional engagement, and interaction with cultural content. Museums and cultural organisations increasingly use digital technologies, sensors, and interactive platforms to collect data about visitor experiences. However, these datasets are often heterogeneous and affected by issues such as missing values, measurement errors, inconsistent formats, and fragmented data structures, which limit their analytical usability and reliability. This paper proposes a structured methodological framework for cleaning and integrating multimodal datasets generated in digital cultural experience research. The framework is discussed with relevance to project contexts such as META-MUSEUM and unLoc and focuses on transforming heterogeneous data into more reliable, validated, and interoperable forms that can support future analytical processes and artificial intelligence applications. The proposed approach includes procedures for data validation, error detection, artefact removal, treatment of missing data, and standardisation of variable formats. In addition, the paper outlines a data integration process based on metadata harmonisation, format alignment, and the creation of consistent data structures that enable the combination of multiple data sources. The study highlights the importance of robust data governance practices for improving dataset quality, reproducibility, and analytical value. Its contribution is primarily methodological, offering a literature-informed framework for approaching multimodal data preparation in a more structured and reproducible way.*

Keywords: multimodal data integration, data cleaning, digital cultural experience research, cultural analytics, data governance, artificial intelligence applications.

Introduction

The increasing use of digital technologies in museums and other cultural institutions has changed the way researchers study visitor experience. Interactions that were once explored mainly through interviews, direct observation, or questionnaires can now be examined with the help of digital platforms, interactive installations, and computational tools. These technologies provide a richer

picture of how visitors engage with cultural heritage, participate in collaborative activities, and respond to different forms of cultural content.

At the same time, research carried out in digital cultural environments generates a growing variety of data. Depending on the design of the study, this information may include behavioural traces, interaction logs, survey responses, and other indicators that help researchers better understand visitor engagement. Working with such diverse data sources creates new opportunities, but it also introduces practical difficulties related to organisation, consistency, and interpretation. The experience of projects such as META-MUSEUM, which explores empathy, co-creation, inclusion, and interdisciplinary digital approaches in the cultural heritage domain, illustrates both the potential and the methodological complexity of this field (META-MUSEUM, 2024).

The growing use of digital technologies has also influenced the methodological approaches used to study cultural experience. In the past, visitor engagement was typically examined through separate methods such as interviews, direct observation, or questionnaires. Today, these traditional sources are often complemented by digital records, interaction logs, and data automatically generated by online platforms and interactive systems. This combination of evidence allows researchers to explore visitor behaviour in greater depth and from several perspectives. Digital cultural research increasingly relies on datasets that vary in format, level of detail, and interpretive value. When organised effectively, such multimodal data can provide a more nuanced understanding of visitor experience.

As digital cultural environments generate larger and more diverse volumes of information, the main difficulty is no longer limited to data collection itself. An equally important task is preparing these datasets so that they can be used reliably in research and decision-making. In practice, data gathered in cultural studies often come from different platforms, devices, and observation methods. As a result, researchers frequently encounter incompatible formats, incomplete metadata, missing values, and variations in the level of detail at which information has been recorded. If these issues are not addressed early in the research process, the usefulness of the data is significantly reduced. Inconsistent or poorly documented datasets are harder to combine and may lead to results that are difficult to interpret or reproduce. For this reason, systematic preprocessing has become an essential part of multimodal research. This need is particularly evident in projects that rely on formal data management procedures and shared research infrastructures. The unLoc project, for example, places strong emphasis on data management protocols, storage solutions, GDPR compliance, and the integration of information across multiple case studies, highlighting the practical importance of well-organised and interoperable datasets (unLoc, 2024).

Data cleaning and integration are therefore more than routine preprocessing steps. They represent core methodological activities that make multimodal datasets suitable for meaningful analysis. Before such data can be used for advanced analytics, artificial intelligence applications, or comparative studies, they need to be carefully checked, standardised, documented, and arranged in a consistent structure. This involves not only correcting errors, but also aligning metadata, harmonising formats, and maintaining clear records of how the data have been processed. These requirements are particularly important in projects where information is collected from several systems and may later need to support both immediate analysis and long-term reuse.

Starting from these considerations, this paper proposes a methodological framework for cleaning and integrating multimodal data in digital cultural experience research. The framework is especially relevant for initiatives such as META-MUSEUM and other research infrastructures that generate diverse datasets through digital and interdisciplinary methods. The study addresses a practical question: how can heterogeneous data collected in digital cultural environments be

organised so that they become more coherent, interoperable, and useful for future research? In response, the paper offers a literature-based methodological approach intended to support data preparation in a systematic and reproducible manner, while also highlighting its value for subsequent analytical applications.

Literature review

The growing use of digital technologies in museums and other cultural institutions has changed the way visitor experience is investigated in cultural heritage and digital humanities research. Interactive installations, online platforms, and data-driven systems now allow researchers to observe in greater detail how visitors engage with exhibitions, cultural narratives, and participatory activities. As a result, research in this field is concerned not only with interpreting cultural content, but also with understanding how the underlying data are collected, organised, and used in analysis (Giannini & Bowen, 2019; Poole, 2022).

An important development in this area is the increasing use of multimodal data. Studies carried out in digital cultural environments often combine behavioural observations, interaction logs, survey responses, and other indicators related to visitor experience. By bringing together these different sources of information, researchers can obtain a more complete picture of human behaviour than would be possible through a single method alone. Baltrušaitis et al. (2022) emphasise that multimodal approaches are particularly valuable because they integrate complementary forms of data within a shared analytical framework. This is especially relevant in cultural research, where visitor experience is shaped by multiple contextual and behavioural factors.

At the same time, working with heterogeneous datasets introduces a number of methodological challenges. Data collected through different devices and platforms frequently vary in structure, format, timing, and level of documentation. In practice, researchers often encounter missing values, duplicated records, inconsistent coding schemes, incomplete metadata, and measurement artefacts. These issues can reduce the reliability of subsequent analyses and limit the usefulness of the data. For this reason, data cleaning should be viewed as an essential stage of the research process rather than as a purely technical task. Studies on data wrangling and data preparation have shown that identifying and correcting inconsistencies is necessary for producing coherent datasets that are suitable for analysis (Kandel et al., 2011).

Preparing individual datasets, however, is only part of the process. Multimodal research also depends on effective data integration. In this context, integration involves more than merging files from several sources. Formats, metadata, and semantic structures must be aligned so that the information can be interpreted consistently across modalities. Rekatsinas (2022) notes that successful data preparation relies on standardisation, interoperability, and schema alignment. These requirements are particularly important in digital cultural experience research, where different datasets may describe related aspects of the same visitor interaction but remain difficult to combine unless a clear integration strategy is established.

The importance of careful data preparation is also emphasised in research on artificial intelligence and machine learning. Sambasivan et al. (2021) show that weaknesses introduced during data preparation can affect the reliability of all subsequent analytical stages. This observation is especially relevant for digital cultural research, where increasing interest in AI-supported analysis depends on datasets that are accurate, well documented, and suitable for reuse. In this context, multimodal cultural datasets require not only technical preprocessing, but also procedures that support consistency, transparency, and reproducibility.

These considerations draw attention to data governance and research infrastructures. Borgman (2020) highlights the importance of responsible stewardship, while Kitchin (2021) points out that data are shaped by the technical and organisational practices through which they are produced and managed. Candela et al. (2020) likewise underline the role of digital infrastructures in supporting the integration and long-term reuse of research data across disciplines. The FAIR principles reinforce this perspective by recommending that datasets should be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). In cultural research, these principles suggest that high-quality datasets depend not only on cleaning operations, but also on clear documentation and consistent metadata practices.

Digital cultural heritage data present additional interoperability challenges because they often combine metadata schemes, vocabularies, and technical standards originating from different institutional and disciplinary contexts. As a result, integration requires both technical alignment and semantic consistency. Approaches that focus only on correcting errors within individual files may improve local data quality, but they do not fully address the broader challenge of making multimodal datasets comparable and reusable across projects and platforms.

Taken together, the literature shows that digital cultural research increasingly relies on heterogeneous multimodal data, and that such data require careful cleaning, integration, and governance in order to become analytically useful. However, these dimensions are often treated separately in existing research, with limited attention to frameworks that connect them specifically in digital cultural experience research. This gap is important because the value of multimodal cultural data depends not only on their collection, but also on the extent to which they can be organised into reliable, interoperable, and reusable forms. It is this gap that motivates the present paper, which proposes a literature-informed methodological framework for cleaning and integrating multimodal datasets in a more structured and reproducible way.

Methodology

The present study proposes a methodological framework for cleaning and integrating multimodal data used in digital cultural experience research. The framework was developed based on insights from the literature on multimodal data analysis, data preparation, interoperability, and data governance. Its purpose is to address a set of recurring difficulties that researchers encounter when working with heterogeneous datasets collected through different digital platforms and research infrastructures.

Rather than testing the framework on a specific dataset, this study takes a conceptual and literature-based approach. The objective is to bring together the main challenges identified in previous research and organise them into a coherent methodological process. This approach is particularly relevant for digital cultural environments, where data are often generated by multiple systems, stored in different formats, and documented with varying levels of detail. The emphasis of the study is therefore on structuring the preparation process in a way that is transparent, systematic, and reproducible.

The framework was developed in three successive steps. First, the literature was reviewed to identify common issues affecting multimodal datasets, including missing values, duplicated records, inconsistent coding schemes, incomplete metadata, and difficulties related to interoperability. Second, these challenges were organised into a logical sequence reflecting the stages that data typically pass through before they can be analysed effectively. Third, this sequence was adapted to the specific context of digital cultural experience research, where information from several platforms, devices, and institutional settings often needs to be aligned and interpreted together.

The resulting framework includes three main stages: data preparation, data cleaning, and data integration. During the data preparation stage, available datasets are identified and organised within an initial structure that supports further processing. The data cleaning stage focuses on improving internal consistency by detecting errors, addressing missing values, removing duplicate records, and standardising variables and formats (Kandel et al., 2011). The final stage, data integration, brings together the cleaned datasets by aligning formats, harmonising metadata, and creating a coherent structure that can be reused in subsequent analyses (Rekatsinas, 2022).

These three stages were selected because they reflect the essential sequence required to transform heterogeneous datasets into resources that are suitable for analysis. Data must first be identified and organised, then checked and corrected, and finally integrated into a structure that supports interoperability and reuse. By following this sequence, the framework aims to help researchers build datasets that are more consistent, better documented, and easier to use in future analytical applications. More broadly, the framework offers a practical methodological basis for preparing multimodal data in research settings where transparency, consistency, and reusability are important objectives.

Results and discussions

The main result of this study is the development of a methodological framework designed to support the cleaning and integration of multimodal data in digital cultural experience research. The framework responds to practical challenges that frequently arise when researchers work with heterogeneous datasets collected through different digital platforms, devices, and research procedures. It combines three closely related stages—data preparation, data cleaning, and data integration—into a single workflow. Together, these stages address recurring issues discussed in the literature, such as fragmented data structures, inconsistent formats, incomplete metadata, missing values, and limited interoperability. Figure 1 presents the overall structure of the framework and illustrates how diverse data sources can be transformed into a more coherent and reusable dataset.

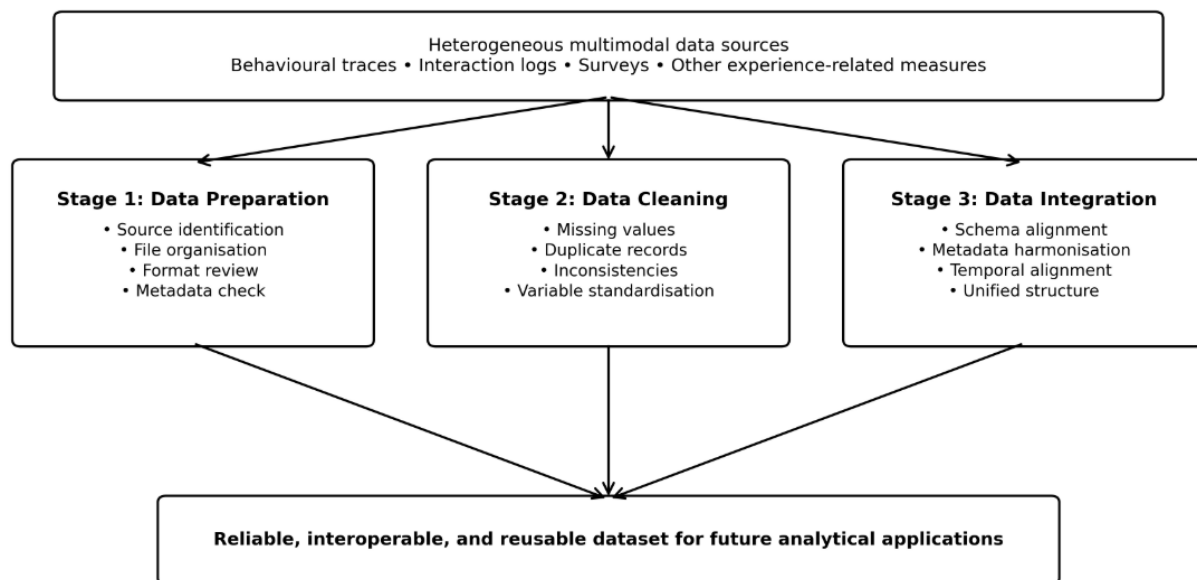


Figure 1. Methodological framework for cleaning and integrating multimodal data

Source: Authors' contribution.

The first stage, data preparation, involves identifying and organising the datasets that will be included in the workflow. In digital cultural studies, information is often collected through multiple platforms, devices, and observation methods, so an initial review of the available data is essential. During this stage, researchers examine file types, variable formats, naming conventions, and the extent to which metadata are available. This preliminary assessment provides a clearer understanding of the data and helps prevent inconsistencies from being carried into later stages. In practical terms, data preparation serves as a diagnostic step that clarifies what information is available and how ready it is for further processing.

The second stage, data cleaning, focuses on improving the internal quality of the datasets. Typical problems include missing values, duplicated records, formatting inconsistencies, coding differences, and other anomalies that may affect interpretation and analysis. Cleaning procedures are used to make the data more consistent and easier to work with. Rather than being viewed simply as error correction, this stage should be understood as an essential step in producing datasets that are sufficiently stable for subsequent analytical use. In multimodal research, this is particularly important because problems in one dataset may influence the interpretation of relationships across several data sources. As noted by Kandel et al. (2011), systematic data cleaning plays a key role in creating datasets that are suitable for reliable analysis.

The third stage, data integration, brings the cleaned datasets together within a common structure. This involves aligning formats, harmonising metadata, and establishing consistent relationships among data sources so that different modalities can be analysed in combination. In digital cultural research, behavioural, interactional, and other experience-related data are often collected separately, even though they refer to related aspects of the same research setting. A well-defined integration stage makes it easier to compare datasets, reuse them in future studies, and support more advanced analytical applications. It also ensures that the final result is more than a set of corrected files and instead functions as a unified analytical resource (Rekatsinas, 2022; Wilkinson et al., 2016).

When considered together, these three stages form a practical workflow for transforming heterogeneous multimodal data into well-structured datasets that are easier to analyse and reuse. The framework is particularly relevant for projects such as META-MUSEUM, where diverse sources of information need to be organised in a transparent and reproducible manner before analysis can begin (META-MUSEUM, 2024). It is equally applicable to initiatives such as unLoc, which involve the management and integration of different types of project and case-study data (unLoc, 2024).

One of the main contributions of the framework is that it treats data preparation, cleaning, and integration as interconnected parts of the same methodological process. In many studies, these activities are discussed separately. The approach proposed here emphasises that, in digital cultural research, the usefulness of data depends not only on technical corrections but also on metadata consistency, traceability, and sound documentation practices. From this perspective, multimodal datasets become analytically meaningful only when all stages of preparation are approached in a coordinated way.

The framework also highlights the fact that data-related challenges are not exclusively technical. Many difficulties arise because information is generated through different platforms, institutional procedures, and observational settings. As a result, successful interoperability depends on shared formats, clear metadata practices, and consistent documentation rather than solely on local adjustments within individual files. In this respect, the framework connects technical

preprocessing with broader concerns related to stewardship, reproducibility, and long-term reuse (Borgman, 2020; Kitchin, 2021).

Another implication of the framework is that it encourages researchers to view data preparation as a continuous process rather than as a series of isolated tasks. In practice, cleaning, formatting, and integration are often carried out at different moments without a common methodological structure. The proposed framework underlines the importance of treating these activities as mutually dependent. This is particularly relevant in digital cultural research, where datasets are frequently generated across multiple experimental and institutional contexts. Under such conditions, the analytical value of the data depends on the transparency and consistency of the entire preparation process.

The framework also has certain limitations. It was developed from the existing literature and remains conceptual in nature. At this stage, it has not been tested empirically on a specific multimodal dataset. Consequently, the study does not claim measurable improvements in data quality or analytical performance. Its contribution lies in providing a structured methodological foundation that can be applied and refined in future research. Further studies may implement the framework in concrete digital cultural case studies and evaluate its effectiveness in improving dataset quality, comparability, and reusability across different research settings.

Conclusion

This study introduces a methodological framework for cleaning and integrating multimodal data in digital cultural experience research. As museums and other cultural institutions increasingly rely on digital technologies, researchers are working with datasets that are more diverse and structurally complex than in the past. Under these conditions, preparing data in a systematic way has become an essential part of the research process. The framework proposed in this paper responds to this need by bringing together data preparation, data cleaning, and data integration within a single methodological workflow.

One of the main contributions of the study is that it connects several issues that are often discussed separately, including data quality, interoperability, and governance-oriented data organisation. By linking these elements, the paper offers a structured and literature-based approach to managing heterogeneous datasets generated in projects such as META-MUSEUM and other digital cultural research infrastructures. The framework is intended to help researchers organise multimodal data in a way that supports transparency, consistency, and future reuse.

The framework remains conceptual and has not yet been tested on a specific multimodal dataset. For this reason, its practical effectiveness still needs to be examined through empirical applications. Future research could implement the proposed approach in concrete case studies, evaluate how well it performs in real research settings, and further refine its stages. Such work would also help determine the broader usefulness of the framework for digital cultural analytics and related areas of interdisciplinary research.

Acknowledgments: This work was supported by the Horizon Europe project META-MUSEUM (Grant Agreement No. 101132488) and by the unLoc project, Exploring the Synergy of Human and Machine Creativity in Architecture. Redesigning Urban Space through Machine Learning, Artistic Expression, and Community Collaboration, funded through the Joint Call Building Transformation Capacity through Arts and Design: Unlocking the Full Potential for Urban Transitions within the European Union's Horizon 2020 research and innovation programme ERA-NET Cofund Urban Transformation Capacities (ENUTC) under Grant Agreement No. 101003758.

The authors used generative artificial intelligence tools solely to improve the linguistic clarity and readability of the manuscript. The conceptual framework, literature review, methodological design, analysis, interpretation of results, and conclusions were developed entirely by the authors, who assume full responsibility for the content of this article.

References

- Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2022). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 9027–9051.
- Borgman, C. L. (2020). Open data, grey data, and stewardship: Universities at the privacy frontier. *Berkeley Technology Law Journal*, 35(2), 365–412.
- Candela, G., Castelli, D., & Pagano, P. (2020). Data infrastructures for digital humanities research. *Future Generation Computer Systems*, 112, 137–147.
- Giannini, T., & Bowen, J. P. (Eds.). (2019). *Museums and digital culture: New perspectives and research*. Springer.
- Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4), 271–288.
- Kitchin, R. (2021). *Data lives: How data are made and shape our world*. Bristol University Press.
- META-MUSEUM. (2024, December 4). META-MUSEUM kick off meeting, 14–15 October 2024, Turin. <https://meta-museum-project.eu/meta-museum-kick-off-meeting-14-15-october-2024-turin/>
- Poole, A. H. (2022). The conceptual landscape of digital cultural heritage. *Journal on Computing and Cultural Heritage*, 15(3), 1–19.
- Rekatsinas, T. (2022). Data integration and data preparation for data science. *Communications of the ACM*, 65(8), 46–53.
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., & Aroyo, L. M. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery.
- unLoc. (2024). unLoc – H2020, Horizon Europe. BEIA Consult International. <https://beiaro.eu/unloc/>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, Article 160018.