CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21876

An IMDb_id Centric Framework for Building Multi-Source Film Datasets for Audiovisual Market Analysis

Carlos A. del Álamo-Alonso[1], Antonio M. Durán-Rosal[1], Javier Pérez-Rodríguez[2]

[1]Department of Quantitative Methods, Universidad Loyola Andalucía, Spain, [2]Department of Computer Science and Artificial Intelligence, University of Córdoba, Spain.

How to cite: Álamo-Alonso, C.A. del; Durán-Rosal, A.M.; Pérez-Rodríguez, J. (2026). An IMDb_id Centric Framework for Building Multi-Source Film Datasets for Audiovisual Market Analysis. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 26-28 June 2026. https://doi.org/10.4995/CARMA2026.2026.21876

Abstract

The growing availability of online film databases has enabled new forms of empirical research in audiovisual economics and media analytics. However, integrating heterogeneous data sources remains a major methodological challenge due to inconsistent identifiers, release definitions, and data quality issues. This paper presents an IMDb_id centric framework for constructing a large-scale film dataset by integrating data from IMDb, OMDb, TMDb, and The Numbers. The IMDb_id is used as the unique and persistent identifier across most of these sources, ensuring deterministic entity resolution and full traceability. The data collection process, identifier-based integration strategy, harmonization of release windows, and validation procedures are documented. The contribution is primarily methodological, offering a reproducible process for researchers working with Internet-based audiovisual data and highlighting best practices for multisource dataset construction. The resulting dataset integrates 498,296 titles, with 20,684 records successfully linked to box office performance.

Keywords: *Internet Open Data; Web scraping; Audiovisual industry; Data integration; Reproducibility.*

1. Introduction

Empirical research on the audiovisual industry has increasingly leveraged large-scale Internet data sources to analyze production characteristics, release strategies, and market performance. The growing availability of online platforms such as IMDb1, box office tracking websites, and social data repositories has enabled predictive modeling and large sample empirical analyses

[1] https://imdb.com/

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

that were previously infeasible using traditional industry reports alone (Eliashberg et al., 2006; Lazer et al., 2009).

This data-driven turn has been particularly relevant in the study of box office performance and audience demand, where machine learning and computational approaches have been successfully applied to predict commercial outcomes using heterogeneous online signals, including metadata, user activity, and early demand (Mestyán et al., 2013; Sharda & Delen, 2006).

This paper addresses this gap by documenting the construction of a multi-source film dataset explicitly designed for audiovisual market analysis. The central methodological decision is the use of the IMDb identifier (IMDb_id) as the unique and canonical key across most of the data sources. The core of this paper will rely on records that are linked deterministically through this identifier and for those where this identifier is unavailable, film year and title matching will be our best effort.

Despite their analytical potential, Internet-based film data sources are not designed for research purposes. They differ substantially in identifiers, temporal definitions, territorial coverage, and variable semantics, which introduce significant challenges for data integration and reproducibility. Prior work in computational social science and data-intensive economics has emphasized that these challenges, if not explicitly addressed, can compromise measurement validity and comparability across studies (Einav & Levin, 2014; Peng, 2011).

The main contribution of this work lies in its methodological approach to enabling robust empirical research. Through describing the full data collection and integration pipeline, we aim to improve transparency, reproducibility, and comparability in Internet-based research on audiovisual markets.

2. Data Sources

2.1. IMDb and OMDb

IMDb constitutes the most comprehensive publicly accessible catalogue of films, covering production metadata, genres, cast and crew information, and user-generated indicators. Part of IMDb's content can be downloaded and used for academic research purposes in accordance with its data usage policies. In this study, however, structured film metadata are primarily obtained through the OMDb API[2], which provides programmatic access to film records in a standardized format.

[2] https://www.omdbapi.com/

Article in Press

OMDb serves as the backbone of the dataset by exposing the IMDb_id, which is adopted as the unique primary key throughout the integration process. Variables collected via OMDb include production year, runtime, country of origin, original language, genre classifications, and basic popularity indicators.

IMDb has become a de facto reference database in empirical film research, frequently used either as a primary data source or as a backbone for integrating complementary datasets. Several studies have relied on IMDb metadata and user-generated indicators to analyze demand formation and box office outcomes, highlighting both its richness and its central role in audiovisual analytics (Cheng & Yang, 2022).

2.2. TMDb

TMDb[3] is used to complement IMDb/OMDb metadata with detailed release information. TMDb provides structured release calendars by country and release type, including theatrical, digital, physical, and television releases. Films are matched to TMDb entries using the IMDb_id through TMDb's external identifier mappings. This enables the extraction of multiple release dates per title, which are essential for reconstructing release window strategies.

2.3. Box Office Data: The Numbers

Box office performance data are primarily obtained from The Numbers[4], which is adopted as the main reference source for theatrical revenue information. Detailed box office data are often subject to commercial access restrictions; however, The Numbers makes a substantial subset of this information publicly available, allowing data collection for research purposes through web scraping techniques.

Access to detailed box office data remains constrained by commercial restrictions in many professional databases. As a result, publicly accessible platforms such as The Numbers have become an important alternative for academic research, despite requiring additional data engineering effort. The use of web scraping techniques to collect economic and market data from publicly available online sources is well established in applied economics and media studies, if data collection procedures and limitations are transparently documented (Einav & Levin, 2014).

Figure 1 provides an overview of the data construction pipeline adopted in this study, illustrating the integration of multiple heterogeneous sources into a unified film dataset.

Article in Press

[3] https://www.themoviedb.org/

[4] https://www.the-numbers.com/

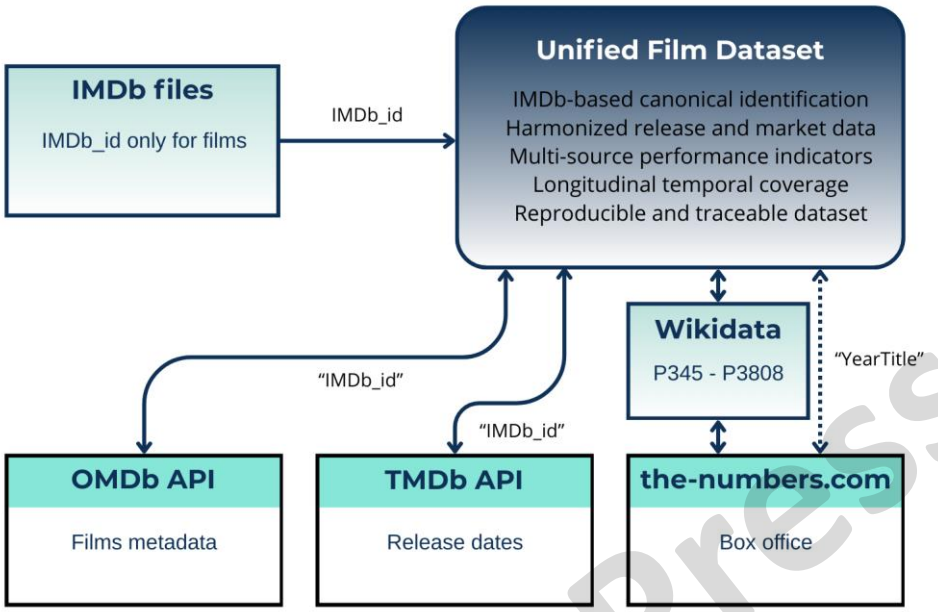


Figure 1. IMDb-centric data integration framework for constructing a unified film dataset.

3. IMDb_id Centric Data Integration Framework

3.1. Canonical Identifier Strategy

The entire dataset is structured around the IMDb_id as the unique and persistent identifier. Each observation corresponds to a single IMDb title and is ultimately anchored to a single IMDb_id. Even in the heuristic stage, matches are anchored back to the IMDb_id.

The decision to adopt the IMDb_id as the canonical identifier follows best practices in large-scale data integration, where stable and globally unique identifiers are essential to ensure deterministic entity resolution. Prior research in data management and knowledge graph construction has emphasized that identifier-based integration significantly reduces ambiguity and cumulative matching error compared to probabilistic or fuzzy approaches (Hogan et al., 2021).

3.2. Identifier-Based Matching Across Sources

IMDb provides the initial list of titles and serves as the anchor point for data collection across all integrated sources. OMDb and TMDb records are matched using the shared IMDb_id ensuring deterministic linkage whenever possible. Film titles, localized names, and release

Article in Press

years are never used as primary join variables in this stage, as these fields frequently generate ambiguity due to remakes, international retitling, and staggered release schedules. By relying on a single canonical identifier, the integration process minimizes cumulative linkage error and preserves internal consistency across heterogeneous databases. As a result, uncertainty is concentrated exclusively in the box office component, where deterministic identifiers are not consistently available and heuristic matching becomes necessary.

As shown in Figure 1, metadata and release information are integrated using IMDb_id matching whenever available, while box office data are linked through the hybrid methodology explained in the next section.

3.3. Integration with Economic Performance Data: A Hybrid Linking Strategy

While IMDb and TMDb provide extensive metadata, reliable financial data (production budgets and worldwide box office gross) is crucial for market analysis. To retrieve this information, we integrated data from The Numbers (the-numbers.com). Unlike OMDb and TMDb APIs, The Numbers does not natively support the IMDb_id query parameter, presenting a significant entity resolution challenge.

To address this effectively and ensure data integrity, we implemented a hybrid two-stage record linkage strategy that combines semantic web technologies with heuristic matching:

Stage 1: Deterministic Linking via Linked Open Data (LOD)

Wikidata was consulted as an auxiliary semantic resource to retrieve deterministic correspondences between IMDb identifiers (Property P345) and The Numbers film identifiers (Property P3808). Coverage was limited, as only 11,599 of 68,659 films from The Numbers were included in Wikidata[5], yielding a narrow but high-confidence set of exact matches. With this method we were able to get box office data from 8,017 films.

Stage 2: Heuristic Fallback with Validation

For records lacking a direct Wikidata entry, we applied a heuristic matching algorithm against data scraped from The Numbers Report Builder[6], adding box office data for a further 12,667 films. The procedure was implemented using Power Query's fuzzy matching engine, which applies an n-gram Jaccard similarity measure over character sequences. Prior to matching, both title sets were normalized through case-folding, punctuation removal, and whitespace standardization, and a release year tolerance of ±1 was enforced to account for international

[5] https://www.wikidata.org/wiki/Property_talk:P3808

[6] https://www.the-numbers.com/movies/report-builder

Article in Press

distribution discrepancies. A similarity threshold of 0.80 was selected to balance precision and recall.

The performance of the heuristic algorithm was evaluated against the 8,017 deterministically linked titles obtained in Stage 1, which served as ground truth. At this threshold, the algorithm successfully matched 7,009 cases, yielding a precision of 95.2%, a recall of 91.4%, and an F1-score of 0.933. The remaining error was largely driven by missed matches (false negatives, n=654) rather than erroneous assignments (false positives, n=354). A complementary manual inspection of 100 borderline matches (0.80–0.85 similarity range) confirmed 91% as correct associations. Most linkage errors involved international title variations ("The Attack" vs. "L'attentat") or semantically similar titles ("Infidel" vs. "The Insider"), rather than systematic failures of the linkage strategy.

4. Harmonization of Release Windows

One of the main methodological challenges concerns the heterogeneity of release definitions across platforms. To address this issue, release dates obtained from TMDb are standardized into six mutually exclusive categories: premiere, theatrical (limited), theatrical (wide), digital, physical, and television.

Release window strategies have been widely studied as a central dimension of audiovisual distribution, particularly in relation to theatrical performance and downstream revenues. Recent work has shown that the sequencing and timing of release windows can have measurable effects on theatrical demand and market dynamics, underscoring the need for precise and comparable temporal definitions when constructing empirical datasets (Brennan et al., 2025).

For each category, the earliest observed release date is retained. This approach enables consistent reconstruction of release windows and the computation of temporal distances between distribution channels. All release window variables are linked to films exclusively through the IMDb_id, ensuring full alignment with production and performance attributes obtained from other sources.

5. Resulting Dataset Structure

The final dataset contains 498,296 records from 1950 to 2024, with several variables organized into four main blocks:

- Production attributes, including genre indicators, runtime, and budget proxies.
- Release strategy variables, capturing the timing and sequencing of distribution windows (399,458 valid records).

Article in Press

- Market performance indicators, such as domestic, international and worldwide box office revenues (20,684 valid records).
- Contextual controls, including release year, country, and distributor.

The dataset is designed to support predictive modeling, counterfactual simulations of release strategies, and longitudinal analyses of audiovisual markets.

6. Discussion

This study contributes to the growing literature calling for greater methodological transparency in Internet-based audiovisual research. While predictive and causal models often receive primary attention, the reliability of empirical findings ultimately depends on the quality and traceability of the underlying data construction process. By explicitly documenting identifier choices, linkage strategies, validation metrics, and release harmonization rules, this framework makes the integration logic transparent and reproducible. The use of a canonical identifier combined with a validated hybrid linkage strategy reduces cumulative matching error and limits ambiguity across heterogeneous sources. This shifts the methodological focus from model selection to data integrity, a foundational but often underreported aspect of computational social science.

The IMDb_id-centric architecture also provides a scalable blueprint for future extensions. Because deterministic linkage is prioritized whenever possible and heuristic matching is formally evaluated against a ground truth benchmark, the integration process remains both flexible and auditable. This structure facilitates dataset updates, incremental expansions, and replication across research teams without redefining entity resolution rules. The framework shows that robust audiovisual analytics require disciplined, reproducible data engineering.

A structural limitation concerns the dependence on scraped box office data from The Numbers, which lacks programmatic access and may change its website structure or access policies. In addition, box office coverage is uneven within the final dataset and is likely biased toward commercially visible titles with wider theatrical circulation, limiting the representativeness of the financial subset.

7. Conclusions

This paper presents a reproducible framework for constructing a multi-source film dataset anchored on the IMDb_id as a canonical identifier. By combining deterministic identifier-based integration with a validated heuristic fallback strategy, the proposed architecture addresses the central challenge of entity resolution in heterogeneous Internet data environments. The empirical validation results demonstrate high linkage precision and recall, supporting the

Article in Press

reliability of the resulting dataset for downstream econometric and machine learning applications. Through harmonization of release windows and systematic documentation of integration choices, the framework ensures temporal consistency and full traceability of attributes across sources. These design principles collectively strengthen the methodological foundations of large-scale audiovisual market research.

Future research will leverage this consolidated dataset to conduct predictive modeling and causal analyses of release strategies and market outcomes. Beyond film, the same identifier-centric architecture can be extended to serialized and episodic content, subject to differences in release structures and identifier coverage. As Internet-based data continues to reshape empirical research in cultural and creative industries, formalized integration pipelines such as the one presented here will be essential for ensuring comparability, transparency, and long-term reproducibility.

Replication materials are available in a public repository at: https://github.com/caaa81-research/unified-film-dataset

References

Brennan, D., Maclean, K. D. S., & Sánchez, M. (2025). Beyond the big screen: Secondary channel releases and their impact on the theatrical market. *Journal of Cultural Economics*, *49*(2), 257–280. https://doi.org/10.1007/s10824-024-09522-0

Cheng, L.-C., & Yang, Y. (2022). The Effect of Online Reviews on Movie Box Office Sales: *Journal of Global Information Management*, *30*(1). https://doi.org/10.4018/JGIM.298652

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1–24. https://doi.org/10.1086/674019

Eliashberg, J., Elberse, A., & Leenders, M. A. A. M. (2006). The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Science*, *25*(6), 638–661. https://doi.org/10.1287/mksc.1050.0177

Hogan, A., Blomqvist, E., Cochez, M., & others. (2021). Knowledge graphs. *ACM Computing Surveys*, *54*(4), 71. https://doi.org/10.1145/3447772

Lazer, D., Pentland, A., Adamic, L., & others. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity. *PLOS ONE*, *8*(8), e71226. https://doi.org/10.1371/journal.pone.0071226

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, *30*(2), 243–254. https://doi.org/10.1016/j.eswa.2005.07.018

Article in Press