IMDB4M: A Large-Scale Multi-Modal Knowledge Graph of Movies

No Author Given

No Institute Given

Abstract. The rapid proliferation of Large Language Models (LLMs) has fundamentally shifted how information is processed, creating an urgent demand for structured knowledge to mitigate hallucinations and ground reasoning. While Knowledge Graphs (KGs) serve as this factual backbone, current multimodal KGs are constrained by a bimodal bottleneck, typically limiting coverage to text and images while often neglecting other important modalities, such as video and audio. Furthermore, existing resources often lack well-established schemas or simply append modalities as flat attributes, limiting their effectiveness for advanced knowledge-based information retrieval and recommendation tasks. In this paper, we introduce IMDB4M, a large-scale, quad-modal knowledge graph of movies. IMDB4M comprehensively harmonises symbolic metadata of movies and actors and integrates them with four distinct modalities: text (plots, comments, reviews), images (posters, stills), video (trailers), and audio (soundtracks). Unlike prior resources often constructed with ad-hoc vocabularies, IMDB4M is engineered on schema.org to ensure semantic interoperability, discoverability and findability (in e.g. dataset search) and structural quality. Modalities are not simply appended as strings, but described in relation to their metadata (e.g., media type, source) and the entities involved (e.g., associated cast), while only external URIs pointing to their raw data are stored to prevent copyright infringement. Furthermore, we enrich the graph by explicitly linking core entities to Wikidata, bridging domain-specific knowledge with Web-scale open knowledge bases. We validate IMDB4M on a subset of movies and external links, and discuss its utility for downstream applications from movie recommendation and entity resolution to multimodal link prediction and multimodal KG embeddings.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Large Language Models (LLMs) have revolutionised information processing and generation, yet their applicability is constrained by inherent limitations such as factual inaccuracies ("hallucinations"), static knowledge bases, and opaque reasoning processes [1]. In this context, the demand for structured, verifiable, and explainable knowledge has become critical for trustworthy AI.

Knowledge Graphs (KGs) address these challenges by providing machine-readable representations of entities and relations [21]. They organise heterogeneous information into multi-relational graphs of facts, typically expressed as triples $(head, relation, tail)$. Integrating KGs with LLMs through Retrieval-Augmented Generation (RAG) mitigates hallucinations and enables dynamic updates without costly retraining [10, 15]. This synergy underpins Neuro-Symbolic AI systems [8] and supports transparent reasoning via auditable paths [13].

Human knowledge is inherently multimodal, combining text, images, audio, and video. Traditional KGs, focused on symbolic facts, fail to capture this perceptual richness, limiting holistic understanding. Multimodal Knowledge Graphs (MMKGs) bridge this gap by enriching entities with sensory data [21, 4]. For example, a conventional KG may store $(Titanic, hasSoundtrack, My\_Heart\_Will\_Go\_On)$, while an MMKG links *Titanic* to its poster (image), trailer (video), and soundtrack (audio), enabling cross-modal reasoning [20].

Despite their promise, existing MMKGs exhibit a "bimodal bottleneck" [21]. Most resources integrate only text and images, as seen in MMKG [11], Richpedia [18], and IMGpedia [6]. Conversely, video-centric datasets such as MovieNet [9] and LSMDC [16] lack semantic rigour and formal ontologies, functioning more as annotated corpora than interoperable KGs. This scarcity of quad-modal resources (Text, Image, Video, Audio) limits research on holistic scene understanding and multimodal recommendation systems.

Constructing a multimodal KG for cinema entails three key challenges. First, modelling heterogeneous modalities requires balancing ad-hoc vocabularies with standard schemas that ensure interoperability [2, 7]. Second, the absence of standardised schemas in many MMKGs creates data silos, hindering integration with the Semantic Web. Third, multimedia content is subject to intellectual property constraints, making hosting raw media legally problematic [**?**].

To address these gaps, we introduce IMDB4M, a large-scale, quad-modal knowledge graph for the movie domain that is unprecedented in multimodal cover, knowledge design, and links. IMDB4M is the first resource to combine a culturally rich domain with comprehensive coverage of four modalities: text (plots, reviews), images (posters, stills), video (trailers), and audio (soundtracks). Unlike prior resources, IMDB4M is engineered on the widely adopted `schema.org` vocabulary [7], ensuring semantic interoperability, discoverability, and structural quality [3]. This design choice positions IMDB4M as a resource that is not only domain-specific but also fully aligned with Web standards, enabling seamless integration with existing semantic infrastructures.

IMDB4M implements a strict "linking over hosting" policy, storing persistent URIs to legitimate platforms such as IMDb and YouTube rather than raw media files, thereby navigating legal and ethical constraints [**?**]. The graph is enriched with explicit links to Wikidata, bridging domain-specific knowledge with one of the largest open knowledge bases and enhancing connectivity across the Web of Data. In total, IMDB4M comprises over 1.8 million RDF triples, representing 376 fully annotated movies and 5,484 artists, with rich multimodal content including 6.9 images and 11.2 audio clips per movie. This scale and richness make

IMDB4M a unique benchmark for multimodal AI research, supporting tasks such as recommendation, entity alignment, link prediction, and KG embeddings.

We validate IMDB4M through SPARQL-based question answering and link verification, achieving high accuracy and coverage. Finally, we discuss its potential for downstream applications and its role in advancing multimodal reasoning within the Semantic Web.

2 Related Work

The development of MMKGs represents the convergence of Semantic Web technologies and Multimedia computing. This section surveys the evolution of MMKGs, highlighting the limitations of current resources and positioning IMDB4M within the landscape.

The first generation of MMKGs primarily aimed to augment standard link prediction benchmarks with visual features. These resources extended existing textual KGs by attaching images to entities, often as simple attributes rather than semantically described objects. **MMKG** [11] is a seminal collection comprising three datasets (FB15k-237-IMG, WN18-IMG, and DB15k-IMG) that link entities from Freebase, WordNet, and DBpedia to representative images. While MMKG established the benchmark for Multimodal Knowledge Graph Completion (MMKGC), its representation remains shallow, as images are typically treated as feature vectors for embedding models rather than first-class entities. Building on this idea, **Richpedia** [18] introduced a visual-relational KG constructed from Wikidata and Wikipedia. It applies diversity filtering to ensure representative image sets and mines visual relations from hyperlinks. Similarly, **IMGpedia** [6] links 15 million images from Wikimedia Commons to DBpedia resources, enabling "visuo-semantic" queries. Although these resources significantly enriched visual coverage, they remain constrained to static images and lack dynamic modalities such as audio and video.

Beyond general-purpose KGs, domain-specific efforts have explored richer modalities. In the music domain, WASABI [5] constructed a KG of over two million songs, integrating cultural metadata, lyrics, and audio analysis features. Likewise, **ChoCo** [**?**] focuses on chord progressions and symbolic music structures. These resources demonstrate that domain-specific KGs can achieve high modality density, but their scope is naturally limited to auditory content.

In the video domain, the computer vision community has produced large-scale datasets such as **MovieGraphs** [17], which provides graph-based annotations of social situations in movie clips, and **MovieNet** [9], which aggregates bounding boxes, cinematic styles, and trailers for over 1,100 films. While rich in perceptual detail, these datasets differ fundamentally from Semantic Web resources: they lack formal ontologies, global identifiers (URIs), and interoperability with the Linked Open Data cloud.

The integration of text, image, video, and audio into a single cohesive graph represents the current frontier in MMKG research. **TIVA-KG** [19] is the most significant predecessor to our work in terms of modality coverage. Constructed

Table 1. Statistical information of experimental datasets.

Dataset	Text	Image	Video	Audio	#Entity	#Relation
MKG-W	14 123	14 463	–	–	15 000	169
MKG-Y	12 305	14 244	–	–	15 000	28
TIVA	11 858	11 636	10 269	2441	11 858	16
KVC16K	14 822	14 822	14 822	14 822	16 015	4
IMDB4M	385 595	37 220	3 983	4 211	660 039	58

from ConceptNet, it introduces "triplet grounding," associating multimodal data with entire factual triples. However, TIVA-KG focuses on commonsense concepts (e.g., "dog", "run") rather than named entities in a culturally rich domain. Other recent efforts include **KVC16K** [12], derived from short instructional videos, and **VAT-KG** [14], a concept-centric resource designed for retrieval-augmented generation tasks.

Table 1 synthesises the position of IMDB4M within this landscape. While prior resources have addressed individual aspects of multimodality, IMDB4M is unique in combining *quad-modal coverage*, *domain specificity*, and *schema standardisation* within a Linked Data architecture. This combination enables interoperability, discoverability, and reuse at Web scale, positioning IMDB4M as a foundational resource for multimodal reasoning in the Semantic Web.

3 The IMDB4M Knowledge Graph

In this section, we detail the construction methodology of IMDB4M, a quad-modal knowledge graph designed to overcome the bimodal bottleneck of existing resources. Our approach is driven by the need to harmonise rigorous symbolic metadata with rich perceptual information—text, image, video, and audio—while adhering to Semantic Web standards. The construction process follows a four-stage pipeline: requirement analysis via competency questions, schema design based on `schema.org`, data population through seeded crawling, and enrichment via neuro-symbolic entity linking.

3.1 Requirements and Schema Design

To formalise the representational requirements of IMDB4M, we defined a set of Competency Questions (CQs) that reflect the information needs of contemporary movie-domain applications. These CQs span both *symbolic* metadata requirements—e.g., "Who directed this movie?", "Which actors played which characters?", "What is the film's box office budget?"—and *multimodal retrieval questions* such as "What is the trailer of the movie?", "Which promotional stills feature a given actor?", and "Which soundtrack corresponds to a particular musical theme or scene?". The resulting CQ set guided the schema toward a unified

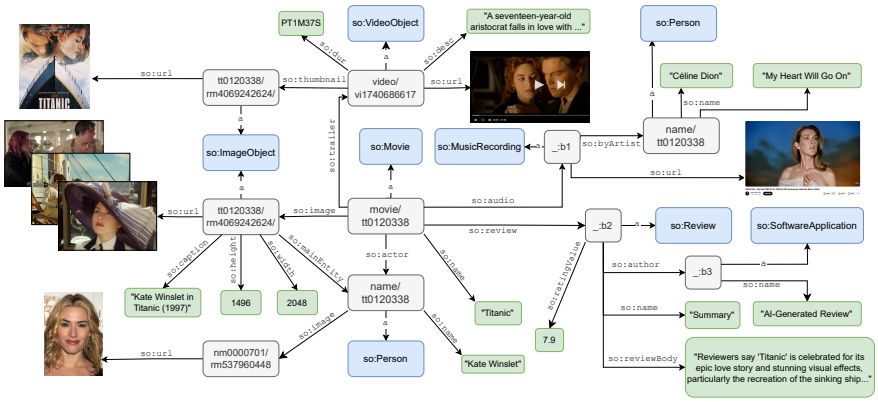

Fig. 1. Illustration of a sample movie KG in IMDB4M for the movie Titanic (`tt0120338`) focuse on showing the modality coverage and their representation via `schema.org`.

representation capable of supporting classic KGQA tasks as well as multimodal reasoning grounded in perceptual artefacts.

To satisfy these requirements, IMDB4M adopts the `schema.org` vocabulary as its primary ontology. This choice is motivated by three factors: (1) *coverage*, as `schema.org` provides modelling primitives for movies, creative works, media objects, ratings, monetary values, and reviews; (2) *expressiveness*, enabling rich structured representations of modalities through classes such as `schema:ImageObject`, `schema:VideoObject`, and `schema:MusicRecording`; and (3) *interoperability*, given the widespread adoption of `schema.org` across the Web of Data and its native use by platforms such as IMDb and YouTube. Unlike earlier resources such as TIVA-KG, which typically represent multimedia as untyped literals or ad-hoc relations, IMDB4M treats multimodal artefacts as first-class semantic objects suitable for complex queries.

Symbolic Modelling At the symbolic level, the schema is centred on the `schema:Movie` class, which serves as the hub for factual, contextual, and perceptual information. Each movie instance is described using properties for titles, release dates, plot summaries, production companies, filming locations, budgets and revenue (via `schema:MonetaryAmount`), aggregate ratings, and keyword metadata.

Roles and Creative Contributors. IMDB4M follows `schema.org` best practices by adopting the `schema:PerformanceRole` pattern to model actor participation. Rather than linking a movie and actor with a simple `schema:actor` triple, we instantiate a `schema:PerformanceRole` node that captures (i) the actor (`schema:actor`), (ii) the associated movie, and (iii) the `schema:characterName`.

This modelling choice preserves actor identity separately from the fictional roles they portray, a nuance commonly lost in less expressive movie knowledge graphs.

Other creative roles—directors, writers, producers, composers—are attached via their respective `schema.org` properties (e.g., `schema:director`, `schema:creator`, `schema:productionCompany`), providing a faithful representation of film-production metadata compatible with Linked Data practices.

Typed Attributes and N-ary Structures. Structured attributes and quantitative values are represented using typed blank nodes with appropriate datatypes, including `xsd:date`, `xsd:dateTime`, `xsd:duration`, `xsd:integer`, and `xsd:decimal`. Runtime, release dates, and monetary information are thus encoded in machine-interpretable form, enabling temporal reasoning, numerical comparison, and schema validation. When multimodal attributes require no internal metadata, they are linked as direct properties; when metadata is relevant to the CQs (e.g., image dimensions or captions), the artefacts are materialised as typed media objects.

Multimodal Modelling A key contribution of IMDB4M is its *quad-modal* representation of text, image, video, and audio. Across modalities, we follow a consistent design principle: whenever an artefact contains query-relevant metadata, it is represented as a typed `schema.org` resource rather than a plain literal.

Textual Modality. Textual content includes plot summaries (`schema:description`), genre and keyword metadata, and user reviews represented as `schema:Review` instances. Each review carries structured provenance information such as author, publication date, and review rating. This enables fine-grained querying over user-generated content, including filtering by sentiment, author, or recency.

Image Modality. Images—including posters, stills, and promotional photographs—are instantiated as `schema:ImageObject`. These objects record URLs, captions, image dimensions, and optionally `schema:mainEntity` links to cast members depicted in the image. This allows IMDB4M to support queries that reference visual content composition, such as retrieving all images in which a given actor appears.

Video Modality. Trailers are represented as `schema:VideoObject` instances. Metadata includes video URLs, thumbnail URLs, durations, and upload dates. This enables temporal and structural filtering over video-based artefacts, such as selecting movies with trailers longer than a specified duration or identifying those published within a certain time window.

Audio Modality. Soundtracks are modelled using a two-level structure: `schema:MusicRecording` for the performed audio artefact and `schema:MusicComposition` for the underlying musical work. Recordings link to performers via `schema:byArtist` and to creators via `schema:composer` or `schema:lyricist`. This design supports

complex audio-related queries such as identifying the composer of a theme or analysing performer contributions across multiple recordings.

In summary, the schema design of IMDB4M is driven by competency questions, grounded in `schema.org`'s expressive modelling framework, and realised in a fully multimodal, semantically coherent RDF dataset. The uniform representation of multimodal artefacts, creative roles, typed attributes, and symbolic metadata enables rich multimodal querying and supports a broad range of downstream applications.

3.2 Knowledge Graph Population

The population of IMDB4M followed a controlled seeding-and-expansion workflow designed to yield a representative and densely connected subgraph of the cinema domain. We adopted a balance between scale and depth, ensuring that multimodal completeness was prioritised over the indiscriminate inclusion of sparsely annotated entities. The process began with a curated seed set of movies selected to maximise temporal, stylistic, and production diversity within the corpus. Each seed movie was enriched with detailed textual, visual, audio, and video descriptors to guarantee high-quality multimodal coverage. We then expanded the graph by traversing from seed movies to their associated artists and subsequently incorporating additional works from these artists to capture meaningful neighbourhood structure. A pruning stage removed leaf movies with insufficient multimodal evidence to preserve the density and utility of the resulting graph. This iterative strategy ensured that IMDB4M remained both semantically coherent and practically useful for downstream multimodal and knowledge-intensive tasks.

Seeding Strategy We constructed the seed set by sampling $N = 100$ movies from each decade between 1980 and 2020 to ensure broad temporal coverage. This stratified approach captured shifts in cinematic style, technology, and production practices across analogue and digital eras. Owing to overlapping release windows in IMDb's search interface, some titles appeared in multiple decadal slices, resulting in a final seed set of 376 distinct movies. Each seed movie was enriched with all metadata required by our competency questions, including the top twenty cast members, the official trailer, associated promotional and still images, and a curated collection of user reviews and AI-generated summaries. This comprehensive enrichment ensured that seed entities served as high-quality anchors for subsequent expansion in the population pipeline.

Recursive Expansion and Pruning To strengthen the topological coherence of the graph, we performed a recursive expansion beginning with the artists identified in the seed set. For each of the 5,484 unique actors, directors, and composers extracted, we retrieved their complete filmography to extend the neighbourhood structure around the seed movies. These additional titles were

incorporated with lightweight metadata to establish valid relational links while maintaining a clear distinction from the fully annotated seed entities. This expansion enabled the graph to capture latent connections between seed movies through shared collaborators who may have worked together outside the initial sample. In a subsequent post-processing phase, we applied a pruning heuristic to remove "leaf node" movies that were connected to only a single artist and thus contributed little to the semantic density of the graph. This step eliminated approximately 69,000 sparsely connected nodes, yielding a refined and cohesive core comprising 660,040 movies and 5,484 artists linked by more than 1.8 million triples.

3.3 Enrichment via External Linking

To transform IMDB4M from a standalone dataset into an interoperable resource within the Linked Open Data ecosystem, we implemented a comprehensive external linking strategy. Our approach focused on integrating high-quality identifiers and multimedia references from Wikidata and YouTube to enhance semantic connectivity and multimodal richness. By aligning IMDB4M entities with established Linked Data hubs, we ensured that users could traverse seamlessly between our graph and broader knowledge networks. This enrichment phase strengthened the reusability, interoperability, and exploratory potential of IMDB4M for downstream semantic and multimodal applications.

Wikidata Alignment We aligned movies and artists in IMDB4M with Wikidata to integrate the graph into one of the Web's most comprehensive general-purpose knowledge bases. This alignment was performed by querying the Wikidata SPARQL endpoint for entities whose IMDb identifier matched the corresponding IMDB4M resource via property `P345`. The resulting identifier-based linking produced 4,284 alignments for artists, covering 78.1% of all actors in the graph, and achieved complete coverage for the seed movie set. These connections enable users to traverse seamlessly from IMDB4M's domain-specific multimodal representations to the wider encyclopaedic context provided by Wikidata.

Neuro-Symbolic Audio Linking A central challenge in constructing a quad-modal knowledge graph is the absence of explicit connections between movie soundtracks and their corresponding audio recordings on external platforms. To overcome this, we developed a Retrieval-Augmented Generation (RAG) pipeline to link internal `schema:MusicRecording` entities to their official or authoritative counterparts on YouTube. The pipeline operates in two stages to balance recall and precision while remaining scalable across large filmographies. In the first stage, we query the YouTube Data API v3 using soundtrack metadata—such as title, artist, and movie—employing a relaxation strategy that progressively broadens search parameters when exact matches are unavailable. In the second stage, we verify candidate videos using a Large Language Model (Gemini 2.5 Flash) acting as a neuro-symbolic reasoner. The model receives a

detailed prompt containing both the soundtrack's musicological metadata (composer, performer, arranger) and the retrieved video's metadata (channel name, description, view count). It returns a structured JSON evaluation that includes a confidence score, an identified best match, and a justification explaining the alignment decision. This two-step approach enables precise disambiguation between official releases, cover versions, and irrelevant uploads, thereby linking a substantial proportion of soundtracks to high-quality, playable audio sources.

3.4 Technical Implementation

The extraction pipeline leverages the structured data embedded within IMDb HTML documents through a multi-layered retrieval approach. Each page provides a JSON-LD block conforming to the schema.org vocabulary, which serves as the primary source for core metadata such as names, publication dates, descriptions, and aggregate ratings. In addition, IMDb employs Next.js for server-side rendering, embedding a comprehensive data payload that exposes deeply nested structures not available in the JSON-LD layer. This payload includes principal credits, filmographies with character roles, featured reviews, and detailed metadata describing associated images. When these structured sources fail to supply specific attributes, the system falls back to DOM traversal guided by test identifiers, CSS class patterns, and URL-based regular expressions. This fallback method enables extraction of properties such as alternate titles, production countries, estimated budgets, and media gallery references. The resulting knowledge graphs are serialised in Turtle format to ensure interoperability and readability within the Semantic Web ecosystem. IMDB4M adheres to a strict "linking over hosting" policy to maintain legal and structural integrity. Rather than minting new URIs for external entities when stable identifiers exist, we reuse canonical IMDb URLs (e.g., `imdb.com/title/tt0120338`) as subject URIs for movies and persons. The graph does not host raw multimedia content, aligning with copyright constraints and established Linked Data practices. Instances of `schema:ImageObject`, `schema:VideoObject`, and `schema:MusicRecording` contain external URIs that reference content hosted on legitimate platforms such as IMDb and YouTube. This design positions IMDB4M as a structural indexing layer that respects copyright while enabling advanced multimodal research and downstream semantic applications.

Resource Availability and Licensing The repository provides the complete IMDB4M dataset alongside the full extraction pipeline required to reproduce its construction. The knowledge graph and its accompanying tools are released under a Creative Commons Attribution–NonCommercial (CC-BY-NC) license to ensure alignment with IMDb's terms of use, which restrict data utilisation to academic and non-commercial settings. To comply with intellectual property regulations and IMDb's redistribution policies, IMDB4M functions strictly as a structural indexing layer rather than a media hosting platform. The resource does not store or redistribute any raw multimedia files, including poster images,

video trailers, or audio clips. Instead, the graph contains only external URIs that reference the original content hosted on IMDb and YouTube. This licensing and design strategy enables sophisticated multimodal research while upholding copyright requirements and ensuring that all media access is directed to the legitimate source platforms.

4 Validation

We conducted a systematic and reproducible validation of IMDB4M to assess its structural integrity, information completeness, and factual correctness. Our evaluation protocol follows established ISWC best practices for resource assessment, combining automated SPARQL-based consistency checks with human-in-the-loop verification procedures. The validation methodology integrates schema-oriented testing, competency-question coverage analysis, and accuracy benchmarking against a manually curated gold standard. Through this multifaceted approach, we address three core research questions that guide the quality assurance of the resource. (V1) Is the schema reused consistently and in accordance with the modelling principles of the graph? (V2) Is the dataset populated comprehensively relative to the intended scope and schema design? (V3) Does the extracted content faithfully reflect the factual information present in the underlying sources?

4.1 Triple Validation via Question Answering

To address V1, V2, and V3, we employed a structured Question Answering (QA) methodology grounded in the Competency Questions presented in Section 3.1. We operationalised each Competency Question as a SPARQL query, enabling systematic verification of schema usage, data population, and factual correctness. In total, we formalised 18 CQs into executable queries that target a broad range of information types represented in IMDB4M. These include core factual attributes such as directors, writers, and runtimes. They also cover numerical properties such as budgets, revenues, and aggregate review scores. Finally, they incorporate multi-valued and relational properties, including cast membership, performance roles, production companies, and genre assignments. This QA-driven strategy ensures that the validation directly reflects the intended functional capabilities of the knowledge graph.

Ground Truth Evaluation We selected a stratified random sample of 20 movies (approximately 5% of the seed set) to construct a gold-standard reference set. For each movie in this sample, human annotators manually extracted the correct answers to all 18 CQs directly from the corresponding source HTML pages. We then executed the SPARQL queries derived from these CQs against the generated knowledge graph and compared the retrieved answers with the gold-standard annotations. To accommodate minor surface-level variations in

textual representation (e.g., "The Untouchables" vs. "Untouchables"), we conducted both strict exact-match comparison and fuzzy matching using a normalised Levenshtein distance with a similarity threshold of $\tau = 0.8$. For questions requiring multi-valued answers, we computed set-based Precision, Recall, and F1 scores to capture partial correctness in both under- and over-generation.

The evaluation results indicate that IMDB4M exhibits high fidelity in its representation of factual information. Across 360 question–answer instances, the knowledge graph achieved an overall F1 score of 94.4% and an average Levenshtein similarity of 0.993. Core bibliographic attributes, including directors, writers, and release dates, achieved perfect F1 scores of 1.0. The most notable deviations occurred in genre assignments, where the Recall was 38.1%. This discrepancy is attributable to inconsistencies between the JSON-LD embedded in the HTML pages, which our extraction pipeline relied upon, and the full genre listings displayed in the rendered page content.

Table 2 summarises the quantitative results of the gold-standard evaluation.

Table 2. Validation results on the gold-standard sample ($N = 20$).

Metric	**Exact Match**	**Levenshtein ($\tau = 0.8$)**
Precision	0.993	0.993
Recall	0.900	0.900
F1 Score	0.944	0.944
Avg. Levenshtein Similarity	0.993	

Coverage Analysis on Full Graph To assess structural consistency (V1) and population completeness (V2) at scale, we executed the full set of 18 SPARQL queries across the complete set of 376 fully annotated movies. This large-scale evaluation served as an automated stress test to verify that the knowledge graph supports all intended query patterns across its entire schema-driven scope. The analysis yielded an overall query success rate of 99.3%, with 13 of the 18 queries achieving complete (100%) coverage. The remaining gaps were confined to optional or sparsely populated attributes, including Metacritic scores (95.2% coverage) and production budgets (95.5% coverage). These omissions stem from incomplete or inconsistent availability of such information in the underlying source data rather than from failures in the extraction pipeline. Notably, no syntax or execution errors occurred during SPARQL evaluation, thereby confirming the structural soundness of the RDF generation process.

4.2 Link Validation

While the symbolic metadata is extracted directly from structured sources, the links to external audio content are generated through the neuro-symbolic pipeline

Table 3. SPARQL Query Coverage Analysis across 376 Movie Knowledge Graphs

ID	Query Description	Success	Empty	Coverage (%)
Q1	Who directed the movie?	376	0	100.0
Q2	Who wrote the script for the movie?	376	0	100.0
Q3	Who are the actors of the movie?	376	0	100.0
Q4	What is the rating of the movie?	376	0	100.0
Q5	How many people have rated the movie?	376	0	100.0
Q6	What is the plot of the movie?	376	0	100.0
Q7	When was the movie released?	376	0	100.0
Q8	What is the runtime of the movie?	376	0	100.0
Q9	What is the Metacritic Score?	358	18	95.2
Q10	What are the keywords?	376	0	100.0
Q11	What is the budget of the movie?	359	17	95.5
Q12	What is the trailer of the movie?	373	3	99.2
Q13	What is the genre of the movie?	376	0	100.0
Q14	What is the poster of the movie?	376	0	100.0
Q15	Which are the production companies?	376	0	100.0
Q16	What are alternate names of the movie?	372	4	98.9
Q17	What is the content rating?	370	6	98.4
Q18	Which are the images and their captions?	376	0	100.0
Overall		**6,720**	**48**	**99.3**

described in Section 3.3. Validating these links is essential to ensure that the `schema:audio` properties reference the correct musical works associated with each film. To support this process, we developed a custom human-in-the-loop validation tool that streamlines the assessment of linked audio resources. The tool presents annotators with the movie metadata, the soundtrack title, and the automatically generated YouTube link, enabling rapid binary verification (Correct/Incorrect) of the linked audio content.

We conducted this validation on a subset of 165 soundtracks associated with the sample movies. The retrieval component of the pipeline successfully identified candidate videos for 148 soundtracks, corresponding to approximately 90% coverage. Subsequent manual assessment confirmed that 129 of these links were correct, resulting in an overall accuracy of 87.16% for the neuro-symbolic linking workflow. Most incorrect links were attributable to cover versions or fan-made uploads that were musically similar to the official soundtrack but differed in provenance. These findings demonstrate the effectiveness of incorporating LLM-based reasoning to disambiguate multimedia content in scenarios where unique identifiers are unavailable.

5 Knowledge Graph Overview

In this section, we provide a detailed statistical analysis of the IMDB4M knowledge graph, characterising its topological structure, entity distribution, and mul-

timodal coverage. The final resource comprises 1,815,922 RDF triples describing 660,039 unique nodes interconnected through 58 distinct predicates. The ontology, grounded in `schema.org`, encompasses 17 entity types, with `schema:PerformanceRole` (232,492 instances), `schema:Movie` (50,756 instances), `schema:ImageObject` (36,844 instances), and `schema:Person` (16,994 instances) constituting the primary classes. A complete summary of the graph statistics is provided in Table 4.

Table 4. Knowledge Graph Statistics

Category	**Metric**	**Value**
RDF Metrics	Triples	1,815,922
	Unique Nodes	660,039
	Unique Predicates	58
	URI Nodes	142,282
	Literal Nodes	264,388
	Blank Nodes	253,369
Graph Structure	Connected Components	1
	Largest Component	141,913
	Graph Density	5.05×10^{-5}
Degree Statistics	Average In-Degree	2.75
	Average Out-Degree	2.75
	Max In-Degree	232,492
	Max Out-Degree	787
Structural Properties	Leaf Nodes	254,713
	Source Nodes	0
	Sink Nodes	297,304
	Hub Nodes (top 1%)	6,656
Entity Types	Unique Types	17
	Typed Entities	362,477
	Multi-Typed Entities	0

5.1 Structural Analysis

A notable design pattern in IMDB4M is the extensive use of blank nodes, which account for 38.4% of all nodes (253,369 instances). This high proportion results from the reification of n-ary relationships, such as modelling an actor's performance in conjunction with their specific character name using the `schema:PerformanceRole` pattern. The predicate distribution exhibits the power-law characteristic typical of real-world scale-free networks; the top four relations (`schema:actor`, `rdf:type`, `schema:performerIn`, and `schema:characterName`) account for 72% of all triples, while the bottom 30 predicates collectively represent merely 2% of the graph.

Structurally, the entity graph forms a single connected component encompassing all 141,913 URI entities. The graph density is low (5.05×10^{-5}) with an average clustering coefficient of 2.33×10^{-4}, indicating a sparse but fully connected topology consistent with biographical and bibliographic knowledge graphs. The degree distribution reveals significant asymmetries: while the average node degree is 5.5, leaf nodes (degree=1) constitute 38.6% of all nodes—primarily representing literal values—and sink nodes (entities without outgoing edges) represent 45% of the graph. Hub analysis identifies 6,656 highly connected nodes (the top 1% by degree), which are predominantly type declarations and prolific actors, with the most connected artist having up to 787 film credits.

5.2 Modality Coverage

A primary contribution of IMDB4M is its comprehensive multimodal coverage across the 376 fully annotated movies in the seed set. For the 376 movies in the dataset seed, Table 6 details the availability of the four core modality categories: text, image, video and audio. Textual content is universally present (100% coverage), encompassing properties such as titles, abstracts, plot descriptions, reviews, and keywords, with an average of 48.63 text elements per movie, as seen in Table 5. Visual content demonstrates similarly robust coverage, with 100% of movies containing image data (averaging 7.91 images per entity, primarily film stills and promotional materials) and 99.20% featuring video content such as trailers. Audio modality, represented by soundtrack clips linked to YouTube recordings, achieves 94.15% coverage with an average of 11.20 audio clips per movie. For the 5,484 actors represented in the knowledge graph, we observe complete coverage across all three applicable modalities—text, images, and videos—with actors exhibiting richer textual annotations (124.81 elements on average) due to comprehensive filmography data including character names and biographical information. Analyzing the joint availability of modalities reveals that 93.62% of movies possess all four modalities simultaneously, while 6.12% contain three modalities (typically lacking audio due to unavailable soundtrack recordings). Only a single movie (0.27%) is limited to two modalities.

Importantly, IMDB4M overcomes the missing modality problem common in many MMKGs. As shown in Table 7, 93.62% of movies in the seed set of the knowledge graph possess all four modality types simultaneously. Only 6.39% of

Table 5. Average Elements by Entity Type across the seed set.

Modality	**Movies (n=376)**	**Actors (n=5,484)**
Text elements	48.63	124.81
Image elements	7.91	7.24
Video elements	0.99	1.73
Audio clips	11.20	N/A

Table 6. Modality Coverage (% of entities with modality) across the seed set.

Modality	**Movies**	**Actors**
Text	100.00%	100.00%
Images	100.00%	100.00%
Videos	99.20%	100.00%
Audio	94.15%	N/A

Table 7. Distribution of Modality Availability

Modalities	**Movies (n=376)**		**Actors (n=5,484)**	
Available	**Count**	**%**	**Count**	**%**
4	352	93.62%	–	–
3	23	6.12%	5,484	100.00%
2	1	0.27%	0	0.00%
1	0	0.00%	0	0.00%
0	0	0.00%	0	0.00%

movies miss one or two modalities, primarily due to the unavailability of official trailers or soundtracks for specific older or less prominent films. For actors, multimodal completeness is absolute, with 100% of entities containing all three modalities. This high degree of multimodal coverage positions the knowledge graph as a valuable resource for research in multimodal knowledge graph completion, cross-modal entity linking, and vision-language reasoning tasks. The structured representation using Schema.org vocabulary ensures semantic interoperability, while the linked multimedia resources—including high-resolution images, video trailers, and audio recordings with YouTube provenance—provide rich grounding for multimodal learning applications.

6 Applications

The unique characteristics of IMDB4M – specifically its quad-modal coverage and schema standardisation – open new avenues for research in the Semantic Web and Multimedia communities. In this section, we outline the primary downstream applications enabled by this resource.

6.1 Movie Recommendation

IMDB4M provides a rich testbed for next-generation recommendation systems that move beyond collaborative filtering or simple metadata matching. By integrating audio and visual modalities, researchers can develop content-based recommendation algorithms that suggest movies based on sensory features, such as the visual style of a poster or the mood of a soundtrack. For instance, a multimodal recommender could analyse the visual style of movie posters (e.g., colour palettes indicating "noir" or "horror") and the acoustic features of soundtracks (e.g., "upbeat jazz" vs. "melancholic orchestral") to identify latent similarities between films. The explicit linking of `schema:MusicRecording` entities to YouTube videos enables researchers to extract audio embeddings directly, facilitating content-based recommendation that goes beyond simple genre matching. Furthermore, the inclusion of video trailers allows for the extraction of temporal visual features, enabling sophisticated queries such as "find movies with high-paced action sequences and electronic scores".

6.2 Multimodal Question Answering

Furthermore, the dataset is particularly well-suited for Knowledge Graph Question Answering (KGQA) and Retrieval-Augmented Generation (RAG). IMDB4M supports the development of "Multimodal RAG" systems by providing a structured, verifiable backbone of facts that are grounded in sensory data. Specifically, the graph allows an AI system to answer complex queries that require joint reasoning over symbolic and perceptual modalities, such as "Who is the actor shown in this scene, and what other movies have they directed?". The validation process produced a dataset of natural language questions, SPARQL queries, and ground

truth answers (Section 4), which can serve as a gold standard benchmark for evaluating KGQA systems over multimodal data. Specifically, the inclusion of complex queries involving reified relations (e.g., "Who played character X in movie Y?") presents a valuable challenge for current QA models [CITE].

6.3 Multimodal Knowledge Graph Completion

IMDB4M provides a robust testbed for foundational tasks in the Semantic Web community, particularly Multimodal Knowledge Graph Completion (MMKGC).

- **Link Prediction:** Researchers can utilise the rich node attributes (images, audio clips, descriptions) to predict missing edges in the graph, such as inferring the `schema:genre` of a movie based solely on its poster and plot summary.
- **Entity Alignment:** The extensive set of `owl:sameAs` links to Wikidata allows for the study of cross-lingual and cross-platform entity alignment. Models can be trained to align entities between IMDB4M and Wikidata by leveraging the visual similarity of actor portraits or movie posters, rather than relying exclusively on string similarity of names.
- **Knowledge Graph Embeddings:** The resource supports the training of multimodal embedding models (e.g., TransE with visual features) that map entities into a continuous vector space, preserving both structural and perceptual semantics.

7 Conclusion and Availability

In this paper, we introduced IMDB4M, a large-scale, quad-modal knowledge graph for the movie domain that overcomes the bimodal bottleneck of existing resources. Unlike prior efforts that rely on ad-hoc vocabularies or limited modality coverage, IMDB4M is rigorously engineered on `schema.org` to ensure semantic interoperability and structural quality. We implemented a novel construction pipeline that harmonises symbolic metadata with text, image, video, and audio, employing a neuro-symbolic linker to reliably associate soundtracks with external media sources. Our validation confirms the resource's high quality, achieving a 99.3% query success rate and a 94.4% F1 score on metadata accuracy. By adhering to a "linking over hosting" policy, we provide a legally sustainable model for sharing multimedia knowledge graphs. IMDB4M is publicly available to the research community to foster innovation in multimodal reasoning, recommendation systems, and the broader Semantic Web.

Resource Availability: The IMDB4M dataset, including the RDF dump (Turtle format), the construction scripts, and the validation tools, is available at the following repository: https://github.com/onradio/imdb4m. The resource is released under a Creative Commons Attribution-NonCommercial (CC-BY-NC) license.

References

1. Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. Can knowledge graphs reduce hallucinations in llms? : A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3947–3960, 2024.
2. Gianluca Apriceno, Valentina Tamma, Tania Bailoni, Jacopo de Berardinis, and Mauro Dragoni. A pattern to align them all: Integrating different modalities to define multi-modal entities. *arXiv preprint arXiv:2410.13803*, 2024.
3. Dan Brickley, Matthew Burgess, and Natasha Noy. Google dataset search: Building a metadata search engine for all the web's data. *Communications of the ACM*, 62(7):26–29, 2019.
4. Zhuo Chen, Yuxia Zhang, Zixuan Wang, Handong Zhao, Ji-Rong Wen, and Xu Chen. Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv preprint arXiv:2402.05391*, 2024.
5. Michael Fell, Elena Cabrio, Maroua Tikat, Franck Michel, Michel Buffa, and Fabien Gandon. The wasabi song corpus and knowledge graph for music lyrics analysis. *Language Resources and Evaluation*, 57(1):89–119, 2023.
6. Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan. Imgpedia: a linked dataset with content-based analysis of wikimedia images. In *International Semantic Web Conference (ISWC)*, 2017.
7. Ramanathan V Guha, Dan Brickley, and Steve Macbeth. Schema.org: Evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016.
8. Pascal Hitzler and Md Kamruzzaman Sarker. Neuro-symbolic artificial intelligence: The state of the art. 2022.
9. Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision – ECCV 2020*, pages 709–727. Springer, 2020.
10. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.
11. Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. Mmkg: Multi-modal knowledge graphs. *arXiv preprint arXiv:1903.05485*, 2019.
12. Haojie Pan, Zepeng Zhai, Yuzhou Zhang, Ruiji Fu, Ming Liu, Yangqiu Song, Zhongyuan Wang, and Bing Qin. Kuaipedia: a large-scale multi-modal short-video encyclopedia. *arXiv preprint arXiv:2211.00732*, 2022.
13. Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
14. Hyeongcheol Park, Jiyoung Seo, MinHyuk Jang, Hogun Park, Ha Dam Baek, Gyusam Chang, Hyeonsoo Im, and Sangpil Kim. Vat-kg: Knowledge-intensive multimodal knowledge graph dataset for retrieval-augmented generation. *arXiv preprint arXiv:2506.21556*, 2025.
15. Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024.

16. Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123:94–120, 2017.
17. Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8581–8590, 2018.
18. Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. Richpedia: A large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, 22:100159, 2020.
19. Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. Tiva-kg: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2391–2399, 2023.
20. Junlin Zhang, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*, 2024.
21. Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):715–735, 2022.