# CresCine's Film Industry Data Repository (FIDA)

Indrek Ibrus[1], Manuel Pita[2], Gabriela Soares[2], João Pedro Carvalho[2], Thiago Paiva[2], Ksenia Mukhina[1], Andres Karjus[1], Ana Falcón[1], Zuil Pirola[2], Bruno Saraiva[2], Marius Øfsti[3], Aksel Nõmmela[1], Alisa Zeleva[4], Jukka Huhtala[4], Jenny Grouiller-Ruhland[4], Saara Hyvönen[4], Timo Norros[4], Sergei Posysaev[4], Vejune Zemaityte[5]

[1] Tallinn University, Baltic Film, Media and Arts School

[2] Lusofona University, The Centre for Research in Applied Communication, Culture, and New Technologies

[3] Aarhus University, Department of Media and Journalism Studies

[4] DAIN Studios

[5] independent research

**Corresponding author:** Indrek Ibrus, professor of media innovation, Tallinn University, Estonia; indrek.ibrus@tlu.ee

**Acknowledgements:** Production of FIDA and the writing of this article have been supported by the European Union (grant nr 101094988).

## Abstract

The Film Industry Data Repository (FIDA) is a lifecycle-wide, multi-source database developed by the CresCine consortium[1] to address the persistent data scarcity facing Europe's small and mid-sized film markets. Built on a scalable Databricks architecture and structured through a medallion pipeline (Bronze-Silver-Gold), FIDA integrates heterogeneous datasets covering production metadata, festival circulation, theatrical distribution (showtimes, admissions, box-office), streaming availability, television programming, and socio-economic context. Data from public and open infrastructures (TMDB, Wikidata, Lumiere, World Bank), institutional partners (Cinando, European Audiovisual Observatory), and selected commercial providers (International Showtimes, UsherU, media-press.tv) are cleaned, harmonised, and linked

[1] https://www.crescine.eu/

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them.

through an internal identifier (CresCine ID) using deterministic and fuzzy-matching techniques. The resulting star-schema repository enables cross-window, cross-territory analysis of European films with a granularity not previously available, especially for countries underrepresented in commercial analytics services. FIDA is disseminated through interactive analytical dashboards and simulation tools, supported by the release of specialised aggregated datasets that comply with licensing restrictions. Designed for long-term sustainability and interoperability, FIDA provides a durable evidence base for researchers, policymakers, and industry stakeholders seeking to understand and strengthen European film circulation, performance, and public value creation.

**Keywords**

Film production, film distribution, film data, film analytics, European film market

## Background

The European film industry operates through a multifaceted value chain that spans production, festival circulation, theatrical exhibition, and subsequent distribution windows. Yet the data generated across these stages is collected unevenly, stored in heterogeneous formats, and seldom interlinked in ways that enable systematic comparison across windows, territories, or time periods. Existing commercial data services that do integrate multi-window information typically prioritise the interests of large film markets, offering limited visibility into smaller European countries such as Estonia, Lithuania, Croatia, Portugal, Denmark, Ireland, or Belgium. As a result, researchers as well as public and private stakeholders in these countries face significant barriers to accessing comprehensive, high-resolution information about their domestic industries or related markets.

CresCine—a pan-European research consortium dedicated to strengthening the resilience of Europe's film ecosystem—addresses this structural gap through the development of the Film Industry Data Repository (FIDA). FIDA aggregates structured and unstructured datasets from across the European film domain, encompassing production metadata, festival screenings and awards, theatrical performance, streaming availability, and television showtimes.

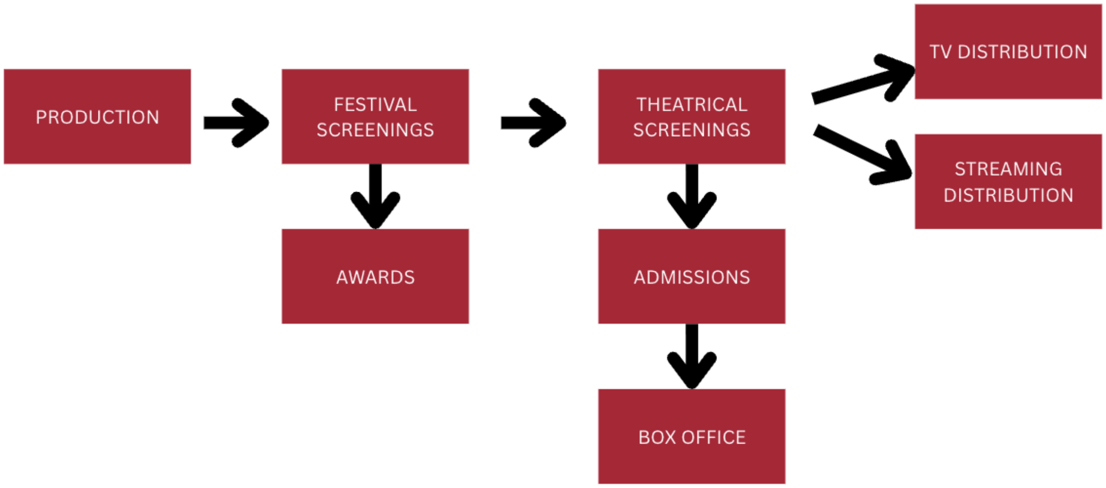

**Figure 1**: Simplified lifecycle of a film from production to distribution.

*This image depicts how a film produced in Europe is typically commercialised. While there are many different ways in which films can complete each stage that the one captured in this figure, this general overview can help to visualise the stages in which data is produced.*

Figure 1 offers a simplified depiction of the film lifecycle and the moments at which key data is generated. A film first enters production, generating metadata such as genre, production countries, spoken languages, and runtime. Upon completion, it circulates through festivals, where initial screenings and industry recognition produce further data points. The subsequent theatrical cycle—first domestic, and later international—creates country-level information on admissions and box-office revenues. After the theatrical window closes, films continue to circulate on broadcast television and video-on-demand platforms, producing additional layers of distribution and availability data. FIDA integrates data from all these stages, enabling lifecycle-wide analyses of how films from different European countries circulate and perform across markets and media environments.

The potential of FIDA is exemplified through two kinds of dashboards: first, analytics tools to visualise relationships across the film life cycle and, second, a theatrical distribution simulation tool. Together, these applications support the development of knowledge and strategies among small European film stakeholders, enhancing their competitiveness.

The design of FIDA is directly shaped by the persistent information gaps affecting Europe's smaller film markets. Commercial analytics firms—such and other global data providers—tend to focus their data-collection efforts on large territories where commercial returns justify detailed monitoring. As a result, stakeholders in smaller markets often lack access to high-resolution data on their own domestic film industries, making comparative analysis across windows and territories difficult. FIDA addresses this gap by integrating data sources that offer universal or near-universal coverage, such as TMDB, Wikidata, International Showtimes, and Usheru, alongside specialised European sources like Lumiere Pro. These datasets collectively provide granular information for all markets rather than privileging only the largest territories. By linking such sources into a unified lifecycle-wide architecture, FIDA enables analyses of production, festival circulation, theatrical distribution, streaming availability, and television programming even for countries that are otherwise invisible in commercial data systems. This design ensures that small-market stakeholders can access evidence that is comparable in quality and granularity to what is typically available only for large film economies.

The creation of FIDA is focused on technical interoperability, making disparate datasets to work together. Due to the lack of shared identifiers between the data sources, this work has meant various challenges and workarounds. In the next pages, the architecture of FIDA is explained, presenting its data sources, data set contents, methods, schema, tables, and usage regulations.

### Data sources used

The CresCine consortium undertook a systematic effort to collect data from multiple sources covering all stages of the film value chain. In some areas, this effort was fully successful, while in others it remains partial. Below, we present an overview of the datasets that had been ingested and integrated into the database by November 2025. Because several datasets originated from privately owned sources, it was necessary to negotiate the conditions under which such data could be made accessible. The resulting agreement stipulates that while FIDA may enable the discovery of analytic insights, it may not permit the publication of data relating to the production or distribution of individual films. Hence, all data is published in aggregate modes. All FIDA data sources are listed below, and the specific features of the data processing applied to each source are detailed in Annex 1.

A central design principle of FIDA was the long-term sustainability of the repository, which required selecting data sources that are legally accessible, financially viable, and technically suitable for integration into an open scientific infrastructure. Consequently, the CresCine consortium prioritised publicly accessible and non-proprietary datasets such as Lumiere and TMDB, as these enable persistent availability to European stakeholders without imposing restrictive licensing conditions. TMDB was preferred over IMDb specifically because the latter is proprietary, expensive to license at scale, and incompatible with the long-term open-access ambitions of the repository. In addition, FIDA incorporates data from Cinando and Lumiere Pro because these sources were made available to the project through institutional partnerships at no additional cost. Where essential data were not available through public channels, the consortium negotiated access to private commercial providers—most notably International Showtimes, Usheru, and media-press.tv—whose terms and pricing structures permitted sustainable inclusion in FIDA. Several potentially valuable datasets could not be obtained despite negotiation efforts, including piracy datasets (e.g., MUSO), audience-level data from major streaming platforms, and reception-oriented data from Rotten Tomatoes, Letterboxd, and festival management systems such as Eventival. These limitations reflect structural barriers to data access in the film and streaming industries, and they have shaped the contours of FIDA's present data architecture.

### The Movie Database (TMDB)

The TMDB dataset is a crowd-sourced repository of film metadata, with information such as title, cast, crew, genre, production companies, production countries, release dates, spoken languages, translations, production dates, release dates by country, and alternative titles.

FIDA features TMDB data from film productions released within the period 2013-2024. Alongside Lumiere Pro, TMDB provides the backbone of basic movie information for FIDA, which is then enriched with information from other ingested datasets.

### Wikidata

The Wikidata-based dataset provides a rich, structured knowledge graph of films and their creative contributors. It captures extensive metadata, including creative personnel, awards, reviews, filming and narrative locations, box office performance, and languages.

FIDA includes Wikidata information from 2017 to 2024. WikiData information facilitates the multidimensional exploration of cinema, linking geographical, periodical, and production-related information with standardised identifiers from a variety of sources.

### World Bank

The World Bank database provides socio-economic context for film market and audience studies. It contains demographic indicators such as population figures across countries and years, which can be integrated with cinema-specific data to normalise performance metrics. FIDA includes World Bank data from the years 1990, 2000, and 2014 to 2023. The inclusion of World Bank data enables per-capita analyses of cinema attendance, film market penetration, and broader cultural consumption trends at the national and regional levels.

### Cinando

Cinando is a proprietary online platform created in 2003 by the Marché du Film of the Cannes Film Festival. Initially a database of Cannes market attendees, it expanded into a year-round infrastructure which aids industry professionals navigate films, projects, contacts, and rights deals more efficiently. Beyond targeting individual users (such as producers, sales agents, distributors, buyers, festival programmers and sales companies), Cinando also serves institutional users: film festivals and film markets use it to schedule market screenings, host B2B screeners securely, coordinate rights sales, manage investment meetings, and access detailed information on films and professionals linked to their events. Its underlying relational database stores metadata on companies, people, films in development, screenings, market events, runtimes, production years, origin countries, languages, crew lists and content or “kind” tags (genre, target audience, identity labels, production type). For festivals specifically, Cinando records festival and event titles, location countries, event years and the full programming of films across thousands of events, enabling the reconstruction of film–festival co-occurrence networks and the analysis of programming diversity, hierarchy, and public value creation across more than 600 festival series worldwide (see Zemaityte et al., 2024). FIDA includes Cinando data covering festivals from 2012 to 2023. This data enables FIDA to study the screenings of European films in festivals across the world.

CresCine Festival Data

The CresCine Festival Data captures the programming, nominations and awards from the international festivals accredited by the International Federation of Film Producers Associations (FIAPF), as well as the remaining festivals that have been included in the Swedish Film Institute (SFI) Quality Index. The combination of these two lists is understood to capture the most influential film festivals across the world.

Each record captures key festival-level attributes such as the festival name, edition, section, and award category, alongside jury status and ecosystem context. Film-level metadata is also included, covering both English and original titles, production year, production country, and associated directors. FIDA features CresCine Festival data for the period 2014-2023. This allows for the examination of festival selections, award patterns, thus providing a bridge between artistic recognition and broader film market data.

UsherU

Usheru is an Irish data-intelligence and content discovery service provider for the film and entertainment industries. It works primarily with institutional partners such as national film bodies, film institutes, studios and distributors, helping them to build direct-to-consumer relationships, track where films are available in cinemas and on streaming platforms, and optimise marketing spend. The core platform aggregates and standardises data on films, showtimes, streaming availability, campaigns and audience interactions. CresCine purchased from UsherU data about the availability of European (member countries of the Council of Europe) on streaming services across the world from the years 2021 to 2024 on a monthly basis. It includes information such as title, genre, release year, type of platform, streaming service, and dates of catalogue addition and catalogue removal.

### Lumiere Pro

Lumiere Pro is the subscription-level offering from the European Audiovisual Observatory designed for film agencies, government bodies and institutional users who require deeper and more comprehensive theatrical market intelligence across Europe and key non-European territories. It builds on the publicly accessible LUMIERE admissions database but adds title-by-title box-office revenue and admissions data from a global measurement vendor (Comscore) for non-European markets as well as supplementary data sets for European markets. The data covers annual admissions and gross box-office figures for individual feature films in each participating market, enabling public film bodies to analyse export patterns, market shares, temporal trends and performance of national catalogues versus foreign films. Institutional users thus gain access to a relational database infrastructure that supports statistical aggregation, visualisation and benchmarking of theatrical film circulation and performance across more than 50 markets worldwide (European and non-European). FIDA includes Lumiere Pro data with regard to the admissions and box office data of European films from the years 1996 to 2023.

### International Showtimes

International Showtimes is a Berlin-based entertainment data company that is specialised in aggregating cinema and showtime information at a global scale. It operates a unified API that serves app developers, media and tech companies, marketing agencies and other institutional clients who need standardised data on which films are playing, where and at what time, across more than one hundred markets worldwide. The service ingests and normalises data from thousands of cinemas and ticketing partners, providing structured metadata on films, cinemas (locations, facilities) and individual showtimes. This infrastructure allows its clients to power film-discovery interfaces, “where to watch” services, campaign tracking dashboards and cross-border market analyses without building their own global data-collection systems, effectively turning showtimes and cinema programming into a machine-readable layer that other discovery, marketing and analytics tools can build upon. CresCine purchased for FIDA its showtimes data from the years 2021-2024. Since 2021-2023 data were raw JSON files that needed extensive cleaning efforts, and hence, not all data about all countries and years are currently included in FIDA.

European Audiovisual Observatory Yearbook

The European Audiovisual Observatory Yearbook is an online service which provides data on television, cinema, streaming and home video in European Countries. FIDA uses the EAO Yearbook as the source for ticket price information. The data featured in FIDA covers price information from multiple international markets spanning the years 2014 to 2023 (historical data). This information enables the advanced analysis of cinema consumption as it enriches admissions, showtimes and box-office performance.

Media-press.tv

Media-press.tv is a European entertainment metadata company that specialises in collecting, structuring and selling detailed data about digital TV schedules and programmes. It aggregates electronic programme guide (EPG) and related metadata for hundreds of broadcasters and platforms, then licenses this data to TV operators, IPTV and cable providers, VOD and streaming services, magazine publishers and other media or internet service providers who need reliable listings and programme information. Its databases model linear and on-demand schedules at channel and programme level, including titles, synopses, genres, series–episode relationships, cast and crew, technical attributes and, for enhanced services, enriched sports and entertainment metadata with structured information about events, teams, venues and visuals. By delivering this data through standardised feeds and tooling, media-press.tv underpins a wide range of commercial EPGs, recommendation engines and TV guides, effectively turning TV schedules and programme line-ups into a machine-readable asset that others use to build user interfaces, search and discovery features and analytics around contemporary television. CresCine purchased for FIDA 4 years (2021-2024) of TV showtimes data - when and how European films have featured across 8000 European TV-channels. This enables to study the interrelationships of TV showtimes in relation to other release windows.

**FIDA Dataset Contents**

FIDA is the result of a three-layered process in which data was cleaned, matched and aggregated into what is called a Gold Layer. The layer follows a traditional star schema in which a core table (dim_filmmaster) connects to the other tables of the repository.

Table 1: Data tables of the Film Industry Data Repository (FIDA)

| Table name | Type | Description |
|---|---|---|

| | | |
|---|---|---|
| dim_filmmaster | Film Master Table | Film Production Information |
| dim_countryclassification | Metrics | Country classification |
| dim_genre | Dimension | All genres attached to a single film |
| dim_main_genres | Dimension | Main genre attached to a single film |
| dim_productioncountries | Dimension | Production Country |
| dim_spokenlanguages | Dimension | Spoken Languages |
| fact_admissions | Metrics | Admissions data |
| fact_tvshowtimes | Metrics | TV screening data. |
| fact_boxoffice | Metrics | Box Office data |
| fact_distributions | Metrics | VOD data |
| fact_market_results | Metrics | Screenings data |
| fact_festivals | Dimension | Festival Information |
| fact_ticketprices | Metrics | Ticket Prices |

FIDA's principal strength lies in its capacity to combine heterogeneous data sources that represent different stages of the film industry value chain, yet this integrative ambition has also constituted its major technical challenge. Because the contributing databases were developed independently and rely on divergent identifier systems – including IMDb, TMDB, EIDR and ISAN – their records cannot be straightforwardly aligned. Considerable work has therefore been required to interlink and reconcile these sources, including the use of fuzzy-matching techniques to approximate correspondences where unique identifiers are absent or incomplete. Furthermore, certain APIs had restrictions on how data could be retrieved and navigated, which elongated the time dedicated to clean, prepare and aggregate data as described in the next section. This has meant that not all films or data points could be matched with full reliability, and some gaps in the unified master data table inevitably remain. Nonetheless, the architecture that has been developed enables robust analysis of film life cycles across multiple release windows and geographical markets, something that no single source previously allowed. Although FIDA cannot be made fully open due to licensing restrictions from commercial providers – which prohibit the publication of film-level records – the project mitigates this constraint by releasing a suite of analytical dashboards and aggregated datasets. These outputs make it possible for policymakers, researchers and industry stakeholders to examine the circulation, performance and reach of European films, including those from small and under-represented markets, through forms of integrated analysis that were not previously feasible. The actions taken to build FIDA are further discussed in the Methods Applied section and Annex 1.

\

## Methods Applied

Different single film identifiers were present in each data source, with not a single one being repeated in all sources. In FIDA, these sources are joined through a comprehensive matchmaking process and the use of an internal ID known as CresCine ID. In addition, the creation of FIDA required processing the data in three stages, which resulted in the creation of three layers: Bronze, Silver, and Gold.

### Data processing

One of the principles of FIDA is that data remains traceable, interoperable, reproducible, and updatable. Therefore, the databases used to build FIDA were processed through a three-layered approach based on the medallion architecture. Each layer (Bronze, Silver, Gold) was employed to fulfill a role within the data transformation pipeline spanning from raw data ingestion to the release of the refined datasets used to build the two FIDA industry dashboards.

The resulting layers support data updates as well as data pool extensions without disrupting its internal structure.

### Bronze Layer (Raw Data Ingestion)

During this stage, data was ingested and saved in its near-original form in order to maintain schema consistency with future data updates. API-sourced data, CSV exports, and JSON files were stored in Delta format, in order to guarantee that any data point can be traced back to its original source. While data transformation was minimal, every row was enriched with a rowhash (to track versioning) and a metadata timestamp (to log when data was loaded and updated).

### Silver Layer (Transformation and Integration)

Raw data was refined in the Silver Layer (SL) in order to improve interoperability between the FIDA datasets, enable the future integration of more data, and prepare the sources for integration in the Gold Layer. Some of the key actions performed in this stage included assigning an internal ID (CresCine ID), cleaning and wrangling data, resolving data inconsistencies and performing other Quality Control (QC) measures, which are overviewed in the next section and detailed in Annex 1.

### Gold Layer

The FIDA Gold Layer aims to support the creation of dashboards, data analyses, and simulation tools. Therefore, the layer follows a star schema, with a Film Master Table (dim_filmmaster) linking to multiple fact and dimension tables as detailed in the FIDA Dataset Contents section. In compliance with EU data protection regulations and to promote a fair usage of this public repository, identifiable data points are not included. The FIDA Gold Layer supports cross-domain analyses and data visualisation to study the lifecycle of films from world premiere to on-demand or TV distribution.

With its three-layered approach, FIDA is a data repository that can be scaled, updated, and extended. Therefore, FIDA is a database that approaches film analytics with a deep commitment to transparency and long-term sustainability. The details regarding the processing of each data source across the three FIDA layers are described in Annex 1.

### Quality Control

Due to the diversity of data sources, Quality Control (QC) was a priority during the integration of FIDA in order to ensure that the data pool was consistent, interoperable, reliable and updatable. QC processes were implemented iteratively during raw data ingestion and transformation. Some of the QC controls performed included:

- Structural validation: encompassing completeness and coherence checks plus the identification of malformed records, invalid data points, and/or null values,
- ID Standardization and enrichment: each dataset was normalized through the creation of an internal identifier (*CresCine ID*) that allowed matching between different IDs (e.g., TMDB ID, Lumiere Pro ID). Where identifiers were missing, fuzzy string matching was applied using field combinations such as the triplet film title x production year x production countries to identify as many titles as possible.

- Semantic standardization: fields such as genres, production countries and spoken languages were standardized. In addition, the contents of genres and spoken languages were also grouped into broader categories.
- Cross-source validation: Data points were cross-checked across sources for better accuracy.
- Data prep for traceability and version control: metadata fields identifying data source, loading data, and date of last update were added to identify the origin of each record. The FIDA Databricks environment is built for full versioning, so every transformation between the Bronze and Gold Layers can be audited and revised if necessary.

Through these Quality Control actions, FIDA achieves structural and semantic consistency. Thus the FIDA repository can serve both as a reliable corpus for research and as an interconnectable data framework that can be expanded.

**Schema and tables**

As mentioned before, the architecture of FIDA is a star schema, with the Film Master Table (dim_filmmaster) being a central database which connects to other repository components. Below is a short description of each of the tables contained within the FIDA Gold Layer.

Table 2: dim_filmmaster

The Master Table (*dim_filmmaster*) is the data backbone of FIDA. This table contains basic film data (e.g., production year, main production country, etc.) as well as the IDs necessary to connect its data to the other FIDA tables to create analyses and visualizations. To facilitate data searches, table joins, and filtering by fields, this table works independently.

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID |
| tmdb_id | int | TMDB ID |
| lumiere_id | double | Lumiere Pro ID |
| title_original | string | Original film title |
| title_english | string | Film title in English |
| production_year | double | Year of Production |
| production_country_main | string | Main Production Country |
| release_date | string | Date of First Release |
| wiki_id | string | WikiData ID |
| imdb_id | string | IMDB ID |
| budget_in_eur | bigint | Budget in euros |
| budget_classification | string | Classification of budget |

Table 3: dim_country

This table identifies the countries listed as production countries within FIDA, their total population, as well as their classification according to variables such as EU membership status.

| Column | Type | Description |
|---|---|---|
| iso_code | string | Country ISO code |
| name | string | Name of country in English |
| population | string | Number of inhabitants of a country |

| | | |
|---|---|---|
| council_of_europe_member | boolean | Status as European Union member |
| crescine | boolean | Status as CresCine focus country |
| council_of_europe_member | boolean | Status as a member of the European Union. |
| eu_small_countries | boolean | Status as a small country within the European Union |
| europe_small_countries | boolean | Status as a small country within Europe (EU and non-EU inclusive) |
| eu_big_5 | boolean | Status as a big country within the European Union |
| europe_big_5 | boolean | Status as a big country within Europe (EU and non-EU inclusive) |

Table 4: dim_genre

This table identifies all the genres that can be attached to a film production.The table is the result of aggregating the genres field of Lumiere Pro, TMDB, and WIkidata.

Across data sources, a single production could be linked to multiple genres. In some data sources, films were categorized based on narrative style (e.g., sci-fi, horror, romance, etc.) while in others genre was used to refer to the type of film production (e.g., animation, live-action, documentary, etc.), running time (e.g., short, feature, etc.), and other discrete categories (e.g., first feature, student film, etc.).

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID number |

| | | |
|---|---|---|
| genres | string | All genres attached to a single film |

Table 5: dim_main_genre

This table identifies main genres in which film productions can be classified. During the processing phase, the aggregated list of genres found across databases was sorted using Databricks Genie Large Language Model (LLM) in order to create a list of main genres that could contain all the genres present across data sources.

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID |
| main_genre | string | Main genre attached to a single film |

Table 6: dim_productioncountries

This table identifies the production country of each film. The production countries are classified to identify them as main production countries, and minority co-production countries.

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID |
| production_country | string | Production Country |
| order | int | Distinction between main production country (order = 1) and coproduction countries |

| | | |
|---|---|---|
| | | (order = [2,∞)) |
| number_of_production_countries | bigint | Number of production country |

Table 7: dim_spoken_languages

This table identifies the spoken language(s) of every film listed in FIDA. The aggregated list of languages was sorted using LLMs, thus creating language families (e.g., French) and their associated languages (e.g., Standard French, Canadian French, Belgian French, etc.).

Since many productions feature more than one spoken language, the field *pos* is used to rank them by their prominence within a film. The value *0* identifies the main language of a production, while successive numbers list other languages spoken within a film.

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID |
| language_family | string | Language Family (e.g., English, Spanish, French, etc.) without specifying a particular variant |
| sub_language | string | Language, specifying a particular variant (e.g., American English, Mexican Spanish, Belgian French, etc.) |
| pos | int | Order of spoken language, with the main one being 0 |
| source | string | Source from which the spoken language data was retrieved (e.g., |

| | | |
|---|---|---|
| | | WikiData, TMDB, Lumiere Pro) |

Table 8: fact_admissions

This table counts the number of theatrical admissions for each film production, organized by year and country in which the tickets were sold. To facilitate the analysis of domestic film consumption, films produced (either in a majority or minority capacity) within the studied country are also identified.

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID |
| year | int | Year in which the admissions took place |
| market | string | Country ISO code |
| national | boolean | Indicates whether the film is a majority domestic production. The value is true if the film was primarily produced domestically, and false if this is not the case |
| national_extended | boolean | Indicates whether the film is a minority domestic production. The value is true if the film was partly (but not primarily) produced domestically, and false if this is not the case |
| admissions | int | Number of Admissions. |

Table 9: fact_box_office

This table presents the box office in euros (€) achieved by each film, organized by year and country in which the revenue was generated. To ease the analysis of the performance of domestic productions, films produced (either in a majority or minority capacity) within the studied country are also identified.

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID |
| year | int | Year in which the box office was recorded. |
| market | string | Country ISO code |
| national | boolean | Indicates whether the film is a majority domestic production. The value is true if the film was primarily produced domestically, and false if this is not the case. |
| national_extended | boolean | Indicates whether the film is a minority domestic production. The value is true if the film was partly (but not primarily) produced domestically, and false if this is not the case. |
| box_office_eur | int | Box Office in Euros (€) |

Table 10: fact_distributions

This table features the distribution details for each production. Since films can have more than one first date of release (e.g., world premiere, festival debut, national release, etc.), additional information was included to facilitate the study of the film distribution cycle. A film can be released in more ways than ever before, with some movies being premiered in Video On Demand (VOD) platforms and others doing a theatrical release first. Thus, this table also contains information regarding the type of distribution, as well as the name of the distribution service used in the case of online releases.

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID |
| release_date | date | Release date by distributor |
| first_release_date | date | First release date known (regardless of distributor) |
| first_theatrical_release_date | date | First theatrical release date known |
| market | string | Country ISO code |
| national | boolean | Indicates whether the film is a majority domestic production. The value is true if the film was primarily produced domestically, and false if this is not the case |
| national_extended | boolean | Indicates whether the film is a minority domestic production. The value is true if the film was partly (but not primarily) produced domestically, and false if this is not the case |
| distribution_type | string | Type of distribution (e.g., Theatrical, TV, SVOD, etc.) |
| distributor | string | Name of distribution platform, if any (e.g. Amazon Prime, Netflix, Google Play, etc.) |

| | | |
|---|---|---|
| release_order_distribution_type | int | release order of distribution type for film (regardless in which market) |
| release_order_market_distribution_type | int | release order of distribution type for film within a market |
| release_order_market | int | release order (of any distribution type) of markets (or market/country groups) |

Table 11: fact_market_results

This table features the performance of each film within a specific market, including admissions and box office in euros (€). The table joins this information with the aim to support analyses related to ticket prices and revenue per capita. To ease the study of domestic films, the table also identifies films produced (either in a majority or minority capacity) within the studied country.

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | CresCine ID |
| year | int | Year in which admissions were recorded |
| market | string | Country ISO code |
| national | boolean | Indicates whether the film is a majority domestic production. The value is true if the film was primarily produced domestically, and false if this is not the case |
| national_extended | boolean | Indicates whether the film is a |

| | | |
|---|---|---|
| | | minority domestic production. The value is true if the film was partly (but not primarily) produced domestically, and false if this is not the case |
| admissions | int | Number of admissions |
| showings | bigint | Number of screenings |
| box_office_eur | int | Box Office in Euros (€) |

Table 12: fact_showings

This table features information regarding the number cinemas in which a film was screened on a particular date. Additional information was included to facilitate the study of theatrical screenings, since cinemas may have more than one screen and/or have several showtimes per screen.

| Column | Type | Description |
|---|---|---|
| id | string | Record ID |
| date | date | Date of screening |
| cinema_id | int | Cinema ID |
| cinema_country | string | Country ID |
| is_id | int | International Showtimes ID |
| title | string | Film Title |
| crescine_id | bigint | CresCine FIlm ID |

**Usage Limitations**

Scope of Use

The Film Industry Data Repository (FIDA) aggregates data on European film industries and markets, integrating datasets from public and private sources. Users may access, analyse, and visualise data through the FIDA Analytics and the FIDA Simulation dashboards as well as by downloading the associated freely accessible datasets subject to the conditions described below.

Permitted Uses

- Users may employ FIDA for research, education, policy development, and industry strategy purposes.
- Users may generate analytical outputs, visualisations, and reports, provided proper attribution is given. Proper attribution is referencing this paper.
- Any reuse of FIDA data must comply with the following ****data source license <INSERT DATA LICENSE>

Restrictions

- Users may not attempt to re-identify anonymised/aggregated data.
- Users may not redistribute FIDA datasets, nor use FIDA to provide paid data services.
- Any simulation or analytical outputs remain the responsibility of the user; FIDA provides no warranty of accuracy.

Attribution

When using FIDA data or tools, users must refer to the repository by referring to this paper as follows:

*"CresCine's Film Industry Data Repository (FIDA), CresCine, [Year of Access], DOI."*

### Liability

Users assume full responsibility for their analyses, outputs, and interpretations. FIDA authors and the CresCine consortium do not guarantee data completeness, accuracy, or fitness for specific queries.

### Termination

Any violation of this license, including misuse of restricted data or data re-identification, may result in immediate suspension of access and possible legal action.

### Reuse of the Data

FIDA data may be re-used for research, education, policy development, and industry strategy purposes. Users are allowed to write analyses, visualisations, and reports, provided that they include proper attribution. Re-use must comply with the license conditions described above.

### Governance and Sustainability

The long-term sustainability of FIDA is ensured through a governance model anchored in the CresCine consortium during the project period and jointly maintained thereafter by Tallinn University and Lusófona University—the two leading academic institutions responsible for the repository's technical development. During the 2023–2026 project cycle, FIDA is being populated with data covering all major release windows across multiple years, enabling extensive experimentation with cross-market and cross-window analyses. After the formal conclusion of CresCine in March 2026, Tallinn University and Lusófona University will continue to manage, update, and expand the repository. This includes integrating new annual data from public and open sources, negotiating continued access to selected commercial datasets, and maintaining FIDA's analytical dashboards as the primary mode of public dissemination. The consortium is also exploring options for establishing a stable institutional home for FIDA as a long-term European film-data infrastructure.

Updates to the repository will occur regularly, subject to the availability of new data from source providers and the capacity of the partner institutions to process and integrate them. Through this governance model, FIDA is positioned as a durable, evolving, and publicly oriented knowledge resource for Europe's film sector.

**References**

Zemaityte, V., Karjus, A., Rohn, U., Schich, M., & Ibrus, I. (2024). Quantifying the global film festival circuit: Networks, diversity, and public value creation. PLOS ONE, 19(3), e0297404. doi:10.1371/journal.pone.0297404

**Annex 1**

## Data Processing per Data Source

### The Movie Database (TMDB)

Data collection and Processing

The TMDB API v3 was employed to retrieve data, which consisted of JSON files with information about film productions including original title, genres, budget, TMDB ID, release date, revenue etc.

During preprocessing, this raw data was transformed and restructured into a table below to allow analysis. The columns updated_at and source were added to identify data source and information of when the data was last updated. Finally, a row hash was added to the rows to prevent double entries. One database was created:

- tmdb_raw_movie: containing all retrieved film metadata (original title, title, genres, budget, TMDB ID, release date, cast, crew, etc.)

This information would then be transformed in the Bronze Layer and become the backbone of the FIDA data pool.

Quality Control

Since TMDB, alongside Lumiere Pro, is one of the two key backbone components of FIDA, comprehensive Quality Control (QC) pipeline was followed to build structural and semantic integrity:

1. Ingestion and completeness control
    - Certain data-heavy information non-relevant to FIDA (e.g., video trailers, poster images, cast and crew images, etc.) were not selected for ingestion.
    - Blank or null-heavy records were pruned to save space, while making exceptions for film titles with explainable data gaps (such as titles not yet released).

2. Standardization of data

    - The meaning and usage of different ID codes was differentiated by re-naming some columns (e.g., cast_id, crew_id).
    - Verified the meaning of values in the field *gender* (0 = Not set / not specified. 1 = Female, 2 = Male, 3 = Non-binary)
    - Converted budgets recorded in USD ($) to EUR (€) using information from the World Bank.
    - Validated ISO country and language codes.

3. Join readiness

    - The availability of IDs against the backbone core database (Lumiere Pro) was tracked to support joins in the Silver and Gold Layers.

Bronze Layer (Raw Data Ingestion)

That data was ingested to the Bronze layer of the Databricks environment. The layer stored the data in Delta format, preserving its original characteristics to ensure traceability and updatability. One table was created:

Table 13: tmdb_raw_movie

| Field Name | Data Type |
|---|---|
| adult | boolean |
| backdrop_path | string |
| belongs_to_collection | string |
| budget | bigint |

| | |
|---|---|
| genres: {”items”: {”id”: ”int”, ”name”: ”string”}} | array |
| homepage | string |
| tmdb_id | int |
| imdb_id | string |
| origin_country: {”items”: ”string”} | array of strings |
| original_language | string |
| original_title | string |
| overview | string |
| popularity | float |
| poster_path | string |
| production_companies: {”items”: {”id”: ”int”, ”logo_path”: ”string”, ”name”: ”string”, ”origin_country”: ”string”}} | array |
| production_countries: {”items”: {”iso_3166_1”: ”string”, ”name”: ”string”}} | array |
| release_date | string |
| revenue | bigint |
| runtime | int |

| | |
|---|---|
| spoken_languages: {”items”: {”english_name”: ”string”, ”iso_639_1”: ”string”, ”name”: ”string”}} | array |
| status | string |
| tagline | string |
| title | string |
| video | boolean |
| vote_average | float |
| vote_count | int |
| alternative_titles: {”titles”: {”items”: {”iso_3166_1”: ”string”, ”title”: ”string”, ”type”: ”string”}}} | struct |
| external_ids: {”imdb_id”: ”string”, ”wikidata_id”: ”string”, ”facebook_id”: ”string”, ”instagram_id”: ”string”, ”twitter_id”: ”string”} | struct |
| keywords: {”keywords”: {”items”: {”id”: ”int”, ”name”: ”string”}}} | struct |
| credits: {”cast”: {”items”: {”adult”: ”boolean”, ”cast_id”: ”int”, ”character”: ”string”, ”credit_id”: ”string”, ”gender”: ”int”, ”known_for_department”: ”string”, ”id”: ”int”, ”name”: ”string”, ”order”: ”int”, ”original_name”: ”string”, ”popularity”: ”float”, ”profile_path”: ”string”}}, ”crew”: {”items”: {”adult”: ”boolean”, ”credit_id”: ”string”, ”department”: ”string”, ”gender”: ”int”, | struct |

| | |
|---|---|
| ”known_for_department”: ”string”, ”id”: ”int”, ”job”: ”string”, ”name”: ”string”, ”original_name”: ”string”, ”popularity”: ”float”, ”profile_path”: ”string”}}} | |
| release_dates: {”results”: {”items”: {”iso_3166_1”: ”string”, ”release_dates”: {”items”: {”certification”: ”string”, ”descriptors”: {”items”: ”string”}, ”iso_639_1”: ”string”, ”note”: ”string”, ”release_date”: ”string”, ”type”: ”int”}}}}} | struct |
| reviews: {”page”: ”int”, ”results”: {”items”: {”author”: ”string”, ”author_details”: {”name”: ”string”, ”username”: ”string”, ”avatar_path”: ”string”, ”rating”: ”double”}, ”content”: ”string”, ”created_at”: ”string”, ”id”: ”string”, ”updated_at”: ”string”, ”url”: ”string”}}, ”total_pages”: ”int”, ”total_results”: ”int”} | struct |
| images: {”backdrops”: {”items”: {”aspect_ratio”: ”double”, ”height”: ”int”, ”iso_639_1”: ”string”, ”file_path”: ”string”, ”vote_average”: ”double”, ”vote_count”: ”int”, ”width”: ”int”}}, ”logos”: {”items”: {”aspect_ratio”: ”double”, ”height”: ”int”, ”iso_639_1”: ”string”, ”file_path”: ”string”, ”vote_average”: ”double”, ”vote_count”: ”int”, ”width”: ”int”}}, ”posters”: {”items”: {”aspect_ratio”: ”double”, ”height”: ”int”, ”iso_639_1”: ”string”, ”file_path”: ”string”, ”vote_average”: ”double”, ”vote_count”: ”int”, ”width”: ”int”}}} | struct |
| translations: {”translations”: {”items”: {”iso_3166_1”: ”string”, ”iso_639_1”: ”string”, ”name”: ”string”, ”english_name”: ”string”, ”data”: {”homepage”: ”string”, ”overview”: ”string”, ”runtime”: ”int”, ”tagline”: ”string”, ”title”: ”string”}}}} | struct |

| | |
|---|---|
| videos: {”results”: {”items”: {”iso_639_1”: ”string”, ”iso_3166_1”: ”string”, ”name”: ”string”, ”key”: ”string”, ”site”: ”string”, ”size”: ”int”, ”type”: ”string”, ”official”: ”boolean”, ”published_at”: ”string”, ”id”: ”string”}}} | struct |

The information was kept as it was ingested to ensure that it can be updated and changes can be backtracked.

### Silver Layer (Transformation and Integration)

In the FIDA Silver Layer, the TMDB raw data was organized into a series of tables. The main table (tmdb_fact_movie) contains general film information, with the goal of it to become part of the FIDA backbone in the Gold Layer. Additional information that would for the other Gold Layer components was organized as separate tables resulting in the creation of ten tables with TMDB data.

Table 14: tmdb_fact_movie

This table contains basic film production information, and would be joined with LumierePro data to create the FIDA Master Table (*dim_filmmaster*) in the FIDA Gold Layer.

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique movie identifier |
| adult | boolean | Indicates if the movie is for adults |
| backdrop_path | string | Path to the backdrop image |

|  |  |  |
|---|---|---|
| belongs_to_collection | string | Collection the movie belongs to |
| budget | bigint | Movie budget amount |
| facebook_id | string | Facebook page ID or URL |
| homepage | string | Official homepage URL |
| imdb_id | string | IMDb identifier |
| instagram_id | string | Instagram page ID or handle |
| original_title | string | Original title of the movie |
| overview | string | Short description or synopsis |
| popularity | float | Popularity score from TMDB |
| poster_path | string | Path to the poster image |
| release_date | string | Movie release date |
| revenue | bigint | Total revenue earned |
| runtime | int | Movie duration in minutes |
| status | string | Current release status (e.g., Released, Post Production) |
| tagline | string | Movie tagline or slogan |

Table 15: tmdb_alternative_titles

This table table contains alternative titles, and is not included in the Gold Layer. However it remains accessible via the TMDB ID.

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique identifier |
| iso_3166_1 | string | Country code following ISO 3166-1 standard |
| title | string | Title or name associated with the record |
| type | string | Type or category of the entry |
| pos | int | Position or order value |
| unique_key | string | Unique key used for deduplication or joins |
| source | string | Data source or origin |
| loaded_at | timestamp | Timestamp when the data was initially loaded |
| updated_at | timestamp | Timestamp when the data was last updated |
| file | string | Source file name or reference |

| | | |
|---|---|---|
| rowhash | string | Hash value representing the row’s content |

Table 16: tmdb_cast

This table table contains cast information, and is not included in the Gold Layer. However it remains accessible via the TMDB ID.

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique movie identifier |
| adult | boolean | Indicates if the person is an adult actor |
| cast_id | int | Unique cast ID within the movie credits |
| character | string | Character name played by the actor |
| credit_id | string | TMDB credit identifier |
| gender | string | Gender of the cast member |
| known_for_department | string | Department or field the person is known for (e.g., Acting, Directing) |
| id | int | TMDB person ID |

| | | |
|---|---|---|
| name | string | Full name of the person |
| order | int | Order of appearance in credits |
| original_name | string | Original name of the person (if different from localized version) |
| popularity | float | Popularity score from TMDB |
| profile_path | string | Path to the profile image |
| pos | int | Position or order value |
| unique_key | string | Unique key used for deduplication or joins |
| source | string | Data source or origin |
| loaded_at | timestamp | Timestamp when the data was initially loaded |
| updated_at | timestamp | Timestamp when the data was last updated |
| file | string | Source file name or reference |
| rowhash | string | Hash value representing the row’s content |

Table 17: tmdb_crew

This table table contains crew information, and is not included in the Gold Layer. However it remains accessible via the TMDB ID.

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique movie identifier |
| adult | boolean | Indicates if the person is an adult contributor |
| credit_id | string | TMDB credit identifier |
| department | string | Department the person worked in (e.g., Directing, Writing, Editing) |
| gender | string | Gender of the crew member |
| known_for_department | string | Primary department or field the person is known for |
| id | int | TMDB person ID |
| job | string | Specific job title or role (e.g., Director, Producer) |
| name | string | Full name of the person |
| original_name | string | Original name of the person (if localized name differs) |

| | | |
|---|---|---|
| popularity | float | Popularity score from TMDB |
| profile_path | string | Path to the profile image |
| pos | int | Position or order value |
| source | string | Data source or origin |
| loaded_at | timestamp | Timestamp when the data was initially loaded |
| updated_at | timestamp | Timestamp when the data was last updated |
| file | string | Source file name or reference |
| rowhash | string | Hash value representing the row’s content |

Table 18: tmdb_genres

The information of this table was used to develop two Gold Layer tables:

- Main Genre (dim_main_genres), which list the genres that encompass all other listed genres of a film.
- Genres (dim_genres), which list all the available genres applicable to FIDA films.

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique identifier |

| | | |
|---|---|---|
| id | int | Internal or related record identifier |
| name | string | Name or label associated with the record |
| pos | int | Position or order value |
| unique_key | string | Unique key used for deduplication or joins |
| source | string | Data source or origin |
| loaded_at | timestamp | Timestamp when the data was initially loaded |
| updated_at | timestamp | Timestamp when the data was last updated |
| file | string | Source file name or reference |
| rowhash | string | Hash value representing the row’s content |

Table 19: tmdb_production_companies

This table table contains production company information, and is not included in the Gold Layer. However it remains accessible via the TMDB ID.

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique movie identifier |
| id | int | Company or organization identifier |
| logo_path | string | Path to the company's logo image |
| name | string | Name of the company or production entity |
| origin_country | string | Country of origin (ISO 3166-1 code) |
| pos | int | Position or order value |
| unique_key | string | Unique key used for deduplication or joins |
| source | string | Data source or origin |
| loaded_at | timestamp | Timestamp when the data was initially loaded |
| updated_at | timestamp | Timestamp when the data was last updated |
| file | string | Source file name or reference |

| | | |
|---|---|---|
| rowhash | string | Hash value representing the row’s content |

Table 20: tmdb_production_countries

This table table contains production country information. Some elements of this table were used to standardize the field of production country and were used to create the Gold Layer table for Production Country (*dim_productioncountries*).

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique movie identifier |
| iso_3166_1 | string | Country code following ISO 3166-1 standard |
| name | string | Name of the country or region |
| pos | int | Position or order value |
| unique_key | string | Unique key used for deduplication or joins |
| source | string | Data source or origin |
| loaded_at | timestamp | Timestamp when the data was initially loaded |

|  |  |  |
|---|---|---|
| updated_at | timestamp | Timestamp when the data was last updated |
| file | string | Source file name or reference |
| rowhash | string | Hash value representing the row’s content |

Table 21: tmdb_release_dates

This table table contains release date information. As a film could have several release dates (e.g., world premiere, international premiere, domestic premiere, etc.) the information for this table was used to complete the related fields in the FIDA Master (*dim_filmmaster*) and Distributions tables (*dim_distributions*).

. Some elements of this table were used to standardize the field of production country and were used to create the Gold Layer table for Production Country (*dim_productioncountries*).

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique movie identifier |
| iso_3166_1 | string | Country code following ISO 3166-1 standard |
| certification | string | Official content rating or certification (e.g., PG-13, R) |
| descriptors | array (items: string) | List of content descriptors or |

| | | |
|---|---|---|
| | | tags (e.g., ”violence”, ”language”) |
| iso_639_1 | string | Language code following ISO 639-1 standard |
| note | string | Additional notes about the release or certification |
| release_date | string | Date of the movie’s release |
| type | string | Type or category of the release (e.g., theatrical, digital) |
| meaning | string | Explanation or meaning of the certification or release type |
| order | bigint | Order or ranking of the record |
| unique_key | string | Unique key used for deduplication or joins |
| source | string | Data source or origin |
| loaded_at | timestamp | Timestamp when the data was initially loaded |
| updated_at | timestamp | Timestamp when the data was last updated |

| | | |
|---|---|---|
| file | string | Source file name or reference |
| rowhash | string | Hash value representing the row’s content |

Table 22: tmdb_spoken_languages

This table table contains information about spoken languages, which was used to build the Gold Layer table Spoken Languages (*dim_spoken_languages*).

| Name | Type | Description |
|---|---|---|
| Column Name | Data Type | Description |
| tmdb_id | int | TMDB unique movie identifier |
| english_name | string | English name of the language |
| iso_639_1 | string | Language code following ISO 639-1 standard |
| name | string | Native name of the language |
| pos | int | Position or order value |
| unique_key | string | Unique key used for deduplication or joins |
| source | string | Data source or origin |

| | | |
|---|---|---|
| loaded_at | timestamp | Timestamp when the data was initially loaded |
| updated_at | timestamp | Timestamp when the data was last updated |
| file | string | Source file name or reference |
| rowhash | string | Hash value representing the row's content |

Table 23: tmdb_translations

This table table contains information regarding the translated titles, taglines and overview of a film production. While this table is not included in the Gold Layer, it remains accessible in the Silver Layer via the TMDB ID.

| Name | Type | Description |
|---|---|---|
| tmdb_id | int | TMDB unique movie identifier |
| iso_639_1 | string | Language code following ISO 639-1 standard |
| iso_3166_1 | string | Country code following ISO 3166-1 standard |
| name | string | Localized or alternative movie name |

| | | |
|---|---|---|
| english_name | string | English version of the movie name |
| homepage | string | Official movie homepage URL |
| overview | string | Description or synopsis of the movie |
| runtime | int | Movie duration in minutes |
| tagline | string | Movie tagline or slogan |
| title | string | Official title of the movie |
| pos | int | Position or order value |
| unique_key | string | Unique key used for deduplication or joins |
| source | string | Data source or origin |
| loaded_at | timestamp | Timestamp when the data was initially loaded |
| updated_at | timestamp | Timestamp when the data was last updated |
| file | string | Source file name or reference |
| rowhash | string | Hash value representing the |

| | | |
|---|---|---|
| | | row’s content |

Gold Layer (Analytical Output)

The FIDA Gold Layer aggregates curated data into analytical models that support descriptive and predictive analyses. Thus, information that could identify single films was not ingested.  The following TMDB data has been ingested into the Gold Layer data pool:

- Production data (e.g., year of production, budget, production countries)
- Temporal and location data (e.g., release dates by country)
- Cross-database enrichment (e.g., linking with other tables for admissions or box office metrics).

TMDB data in the film master table of the Gold layer helps to widen the data pool by including films not found within Lumiere Pro or Wikidata. Thus, it enriches the insights regarding film production output and lifecycle in smaller markets.

**Wikidata**

Data collection and Processing

The Wikidata API was used to create a data dump with film production information such as original title, genres, cast, crew, budget, release date, box office, etc. During preprocessing,  the raw data was organised into tables.

The data retains the original column names and organisation of the source, in order to allow updates. Further transformations were performed to include certain data points as part of the FIDA backbone in the Gold Layer.

Quality Control

WikiData is a repository in which volunteer editors can add, remove, and update details. Therefore, the following activities were performed in order to verify the reliability of the data, remove identical duplicates, and merge multiple records about the same film:

1. Selection of trusted fields
    - Certain fields were selected to prioritise them as sources of truth for specific film characteristics (e.g. international title instead of original title).
2. Data harmonisation
    - Some field formats (e.g. currency, dates, etc.) were standardize in order to facilitate film matchmaking and interoperability between data sources.
3. External validation
    - The IDs included in each record was cross-checked against the wider dataset created by joining Lumiere Pro and TMDB.
4. Redundancy elimination and deduplication
    - Identical records, or rows in a table with small variations of a particular field (e.g., film festival name) were consolidated into single entries.

Bronze Layer (Raw Data Ingestion)

After QC, the data was ingested into the Bronze layer of the Databricks environment. This layer stores the data in Delta format, preserving its semi-structured nature while ensuring traceability and reproducibility. Twenty-nine tables were created:

- Wiki_raw_sound_designer: containing the entry for the sound designer.
- Wiki_raw_actor: containing the entry for cast information.

- Wiki_raw_award_received: containing the entry for awards received.
- Wiki_raw_base: containing all basic film data (original title, title, genres, budget, Wiki ID, runtime, crew, etc.)
- Wiki_raw_boxoffice: containing the entry for box office.
- Wiki_raw_composer: containing the entry for film music composer.
- Wiki_raw_costumdesigner: containing the entry for costume designer.
- Wiki_raw_director: containing the entry for film director.
- Wiki_raw_director_photography: containing the entry for director of photography.
- Wiki_raw_distributed: containing the entry for film distributor.
- Wiki_raw_executive_producer: containing the entry for executive producer.
- Wiki_raw_filmeditor: containing the entry for film editor.
- Wiki_raw_filming_location: containing the entry for filming location.
- Wiki_raw_follows: containing the entry for preceding or related work.
- Wiki_raw_genre: containing the entry for film genre.
- Wiki_raw_identifier: containing the entry for unique identifier.
- Wiki_raw_main_subject: containing the entry for main subject.
- Wiki_raw_movie_language: containing the entry for movie language.
- Wiki_raw_narrative_location: containing the entry for narrative location.
- Wiki_raw_nominated_for: containing the entry for award nomination.
- Wiki_raw_original_language: containing the entry for original language.
- Wiki_raw_part_of_series: containing the entry for film series membership.
- Wiki_raw_presented_in: containing the entry for presentation format or medium.
- Wiki_raw_producer: containing the entry for film producer.
- Wiki_raw_production_company: containing the entry for production company.

- Wiki_raw_production_designer: containing the entry for production designer.
- Wiki_raw_publication: containing the entry for publication or release.
- Wiki_raw_reviews: containing the entry for film reviews.
- Wiki_raw_screenwriter: containing the entry for screenwriter.

Silver Layer (Transformation and Integration)

Data was transformed within the Silver layer in order to create tables that supported  analytical queries. Certain fields names were harmonised against other fields belonging to external sources in order to ease matchmaking . When ID matchmaking was not possible, fuzzy string matching was performed.

WikiData tables were joined in order to create an additional table containing basic film production information through ID numbers which connect this master table to other dimension tables.

Table 24: wiki_fact_movies

| Column Name | Data Type | Description |
|---|---|---|
| id | string | Unique identifier for the film. |
| id_actor | string | Identifier of actor(s) in the film. |
| id_director | string | Identifier of the film's director. |
| id_distributed | string | Identifier of the distribution company or distributor. |
| duration | string | Runtime or total length of the film |
| id_genre | string | Identifier for the film's genre |

| | | |
|---|---|---|
| id_main_subject | string | Identifier for the main subject or theme of the film. |
| id_original_language | string | Identifier for the original language of the film. |
| id_director_photography | string | Identifier for the director of photography |
| id_executive_producer | string | Identifier for the executive producer(s). |
| id_filmeditor | string | Identifier for the film editor. |
| id_filming_location | string | Identifier for where the film was shot. |
| id_narrative_location | string | Identifier for the story's setting or narrative location. |
| id_producer | string | Identifier for the film's producer(s). |
| id_screenwriter | string | Identifier for the screenwriter(s). |
| id_composer | string | Identifier for the composer of the film's score. |
| id_costumdesigner | string | Identifier for the costume designer. |
| id_production_designer | string | Identifier for the production designer. |
| costs | int | Production budget. |
| id_part_of_series | string | Identifier for the series the film is part of, if any. |
| id_production_company | string | Identifier for the production company. |
| id_follows | string | Identifier of the preceding film, if any. |

| | | |
|---|---|---|
| start_time | date | Date when filming started. |
| end_time | date | Date when filming ended. |
| premiere | date | Date of the film’s first public showing or release. |
| source | string | Source of the data. |
| loaded_at | timestamp | Timestamp when the data was first loaded. |
| updated_at | timestamp | Timestamp when the data was last updated. |
| rowhash | string | Hash value used for deduplication or data integrity. |

Gold Layer (Analytical Output)

The Gold Layer joins curated WikiData data into the Master Table and dimension tables of FIDA. This includes:

- Market-level KPIs (e.g., year of release, countries of distribution per title);
- Production information (e.g. budget, cast, crew, etc.)
- Cross-database enrichment (e.g., Film IDs for a variety of platforms).

By using WikiData as a secondary source to complement the data backbone composed of TMDB and Lumiere Pro information, FIDA extends its data pool to include details that provide context to the observations made across film markets.

### World Bank

#### Data collection and Processing

The World Bank data consisted of a single Excel file:

- Worldbank-population: containing the full name, ISO code, and population of every country listed in the World Bank database in the years 1990, 2000, and 2014 to 2023.

During the pre-processing stage the table and field names were kept as in the raw data, in order to keep consistency in future updates.

#### Quality Control

The data from the World Bank was already organized, yet the following steps were taken to ensure reliability:

1. Data harmonisation

- Countries with multiple records (due to country name changes)  were identified. This information would be used to build a unified country list in the Silver Layer.

2. Consistency check

- The data across years within a single country was compared to detect anomalies or missing values.

3. Validation of format

- The data was reviewed in order to confirm it was properly organized, with standardized country names, ISO country codes, and consistent number formats.

Through this process, it was possible to validate that the data was complete, organized, and ready for further processing.

#### Bronze Layer (Raw Data Ingestion)

The World Bank data was ingested into the Bronze layer of the Databricks environment in Delta format, ensuring traceability and reproducibility.

Table 25: crescine.bronze.worldbank-population

| Column | Type |
|---|---|
| Series Name | string |
| Series Code | string |
| Country Name | string |
| Country Code | string |
| 1990 [YR1990] | bigint |
| 2000 [YR2000] | bigint |
| 2014 [YR2014] | bigint |
| 2015 [YR2015] | bigint |
| 2016 [YR2016] | bigint |
| 2017 [YR2017] | bigint |
| 2018 [YR2018] | bigint |
| 2019 [YR2019] | bigint |
| 2020 [YR2020] | bigint |
| 2021 [YR2021] | bigint |

| | |
|---|---|
| 2022 [YR2022] | bigint |
| 2023 [YR2023] | bigint |

At this stage, data was retained in near-raw form to preserve original fidelity and enable backtracking if needed.

### Silver Layer (Transformation and Integration)

In this stage, column names were reformatted to enhance readability while country names were standardized by adding the column *country_as_used_in_crescine*. This created a redundant way to connect to the table by either using this field or the country code. In addition, timestamps were added in order to log data updates and detail version history.

Table 26: crescine.silver.worldbank_population

| Column | Type |
|---|---|
| Country_name | string |
| country_as_used_in_crescine | string |
| Country_code | string |
| Year_1990 | bigint |
| Year_2000 | double |
| Year_2014 | double |

| | |
|---|---|
| Year_2015 | double |
| Year_2016 | double |
| Year_2017 | double |
| Year_2018 | double |
| Year_2019 | bigint |
| Year_2020 | double |
| Year_2021 | bigint |
| Year_2022 | double |
| Year_2023 | bigint |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

### Gold Layer (Analytical Output)

The Silver Layer data was ported into the Gold Layer, and connects to the main via the standardized column *country_as_used_in_crescine*. The World Bank information in Gold Layer allows for the following types of queries:

- Time and Location (e.g., year and country)
- Population
- Cross-database enrichment (e.g., linking with other tables for per capita and country-level calculations).

The inclusion of this World Bank data enables the contextualization of analyses that examine the performance of films within a particular country in a specific year, as well as enabling the per-capita calculations needed to compare the economic performance of different film markets.

## Cinando

### Data collection and Processing

The collection and preprocessing workflow transformed a MS Access data dump provided by Cinando into structured, analyzable data.  During the pre-processing stage, labeling conventions, table names, and file names from the raw data were kept to create consistency for future data updates.

The raw data was split across tables, which could  be linked and merged using the cFilm field for films, and ref* variables for various film properties (e.g. country, company) with more data in their respective tables.

The data convention in the provided Cinando dump is unusual. Fields have inconsistent names between larger and smaller metadata tables. The variable isCinando variable indicates whether the data point was entered and verified by Cinando employees (1) or platform users (0, the majority). Smaller Cinando tables have a NoOrder variable, which roughly indicates priority (e.g. primary production country) but there is no guarantee users have used this properly. Film IDs were duplicated if they had multiple categories. Finally, there is a variable (Active or Actif) which identifies (with a 0) if an entry was deleted. Thus it is recommended that only Actif=1 entries are used.

Below are short summaries for the tables exported from the MS Access database dump provided by Cinando

Seventeen tables were created within a single volume:

- Countries: contains countries plus ID codes for linking.
- Event (temp): is the general festival events table. It contains information such as event name, event ID code, series ID, location country, and dates. [2]
- Events (main): contains the festival series present in the database.
- Country (film): contains films and their main production country.
- Festivals (temp): this is a merged table of films at festivals.
- Film Crew: contains the names crew members and roles per film, with the roles of producer and director being the most likely ones present.
- Films (Films Roles): contains information about the film production companies.
- Films (generic): this is another film crew table, seems to contain the same data as the above, but without names and few other variables.

[2] Note: As the data from Cinando is user-submitted, there is a considerable number of duplicate events. For example, Cannes Film Festival appears under several slightly different names and different event IDs. This table can be linked using the festival series(idFestival) and libelleFestival (title) columns. While most events are not attached to a festival series, this does not rule out said possibility.

- Kind: contains films and their genres, themes, and other variable tags referring to the director (e.g. first film) or technical aspects (e.g. VR, 3D).
- Market history: contains details about a film's participation at film markets.
- Match film: contains meta-information of when data was added and by whom.
- Media: contains information about media assets (e.g. posters) associated with films.
- Films (main): contains cFilm ID that can be used to link to other tables, original film title, alternative title, completion status, and year of production.
- Films (generic): contains the same data as the above, but lacks names and few other variables.
- TIDX Film: contains the link to identify which film appeared at which festival, awards received, name of the award, and place where an award was given.
- TIDX Film Language: contains film languages
- Film language: contains film languages and other fields about languages such as language codes.

These tables were later ingested in the Bronze layer into the Databricks data architecture.

Quality Control

Given the heterogeneity of the raw API responses, a comprehensive Quality Control (QC) pipeline was implemented to ensure structural and semantic integrity.

1. Structural validation
    - Duplicated records were filtered and merged.
2. Data hierarchisation
    - Meaningful labels (e.g. full country names) were contained in smaller tables, with the main Cinando tables referring to them via c/ref codes.
3. External validation

- The field *eidr_id* was cross-checked against the eidr from the wider dataset created by joining Lumiere Pro and TMDB.

4. Redundancy elimination and deduplication

    - Identical records, or rows in a table with repeated variations of a particular field (e.g., film festival name) were consolidated into a single entry.

Through these measures, the dataset achieved both internal coherence and external referential integrity.

Bronze layer (Raw Data Ingestion)

After QC, the data was ingested into the Bronze layer of the Databricks environment. This layer stores the data in Delta format, preserving its semi-structured nature while ensuring traceability and reproducibility. Seventeen tables were created:

Table 27: crescine.bronze.cinando_raw_tcountries

| Column | Type |
|---|---|
| cCountry | bigint |
| NameEN | string |
| cContinent | bigint |
| Continent | string |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 28: crescine.bronze.cinando_raw_tevent_temp_20250115

| Column | Type |
|---|---|
| UID | bigint |
| VID | double |
| CREEPAR | bigint |
| MODIFIEPAR | double |
| DATECREATION | string |
| DATEMODIFICATION | double |
| DateModificationAdm | double |
| refUtilisateurAdm | double |
| ACTIF | bigint |
| cEvent | bigint |
| refTypeEvent | bigint |
| idFestival | double |
| LibelleEvent | string |
| DateDebut | string |
| DateFin | string |
| Year | double |
| refMarketEdition | double |
| isCinando | bigint |
| isFocusCatchup | double |
| IstTypeEvent | double |

Table 29: cinando_raw_teventmain_temp_20250115

| Column | Type |
|---|---|
| idFestival | bigint |
| libelleFestival | string |
| refCompany | double |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 30: crescine.bronze.cinando_raw_tfilms_country_temp

| Column | Type |
|---|---|
| cFilm | bigint |
| refCountry | bigint |
| Main_Country | string |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |

| | |
|---|---|
| file | string |

Table 31: crescine.bronze.cinando_raw_tfilms_festivals_temp

| Column | Type |
|---|---|
| cFilm | bigint |
| TitleVA | string |
| refCompany | bigint |
| Company | string |
| idFestival | double |
| libelleFestival | string |
| refEvent | bigint |
| Year | double |
| isCinando | bigint |
| IstTypeEvent | double |
| refEventSection | double |
| LibelleSection | string |
| rowhash | string |

| | |
|---|---|
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 32: crescine.bronze.cinando_raw_tfilms_filmcrew_temp

| Column | Type |
|---|---|
| cFilm | bigint |
| refFilmCrew | bigint |
| IstType | bigint |
| txtType | string |
| NoOrder | double |
| FirstName | string |
| LastName | string |
| Company | string |
| Actif | bigint |
| rowhash | string |

| | |
|---|---|
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 33: crescine.bronze.cinando_raw_tfilms_filmsroles_temp

| Column | Type |
|---|---|
| cFilm | bigint |
| TitleVA | string |
| fVis | bigint |
| IstProductionStatus | double |
| txtProductionStatus | string |
| refCompany | bigint |
| Company | string |
| ceVis | bigint |
| SALES | bigint |
| DISTR | bigint |

| | |
|---|---|
| PROD | bigint |
| FIN | bigint |
| FEST | bigint |
| PROMO | bigint |
| REP | bigint |
| CONS | bigint |
| BROAD | bigint |
| rowhash | string |
| source | string |
| loaded_at | timestamp |

Table 34: cinando_raw_tfilms_generic_temp

| Column | Type |
|---|---|
| cFilm | bigint |
| refFilmCrew | bigint |
| IstType | bigint |
| txtType | string |

| | |
|---|---|
| NoOrder | double |
| Actif | bigint |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 35: crescine.bronze.cinando_raw_tfilms_kind_temp

| Column | Type |
|---|---|
| Active | comp |
| cFilm | bigint |
| Main_Genre | string |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |

| | |
|---|---|
| file | string |

Table 36: crescine.bronze.cinando_raw_tfilms_markethistory_temp

| Column | Type |
|---|---|
| cFilm | bigint |
| TitleVA | string |
| refCompany | bigint |
| Company | string |
| refMarketEdition | bigint |
| Year | bigint |
| cMarket | bigint |
| Name | string |
| IstType | double |
| txtType | string |
| IstMarketStatus | double |
| txtMarketStatus | string |
| historicStatus | double |

| | |
|---|---|
| IstProductionStatus | double |
| txtProductionStatus | string |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 37: crescine.bronze.cinando_raw_tfilms_match_film_temp

| Column | Type |
|---|---|
| IDFILM | bigint |
| REFFILMPARTENAIRE | bigint |
| IDPARTENAIRE | bigint |
| ISPROJET | double |
| ANNEEPRODUCTION | double |
| ISAN | string |
| ACTIF | bigint |

| | |
|---|---|
| DATECREATION | string |
| DATEMODIFICATION | string |
| CREEPAR | string |
| MODIFIEPAR | string |
| REFMATCHING | double |
| LSTTYPETRAITEMENT | bigint |
| TXTTYPETRAITEMENT | string |
| COMMENTAIREELEMENT | string |
| ISHORSTRAITEMENT | bigint |
| ISENCOURSTRAITEMENT | bigint |
| TRAITEPAR | string |
| DATEDEBUTTRAITEMENT | double |
| rowhash | string |

Table 38: crescine.bronze.cinando_raw_tfilms_media_temp

| Column | Type |
|---|---|
| cFilm | bigint |

| | |
|---|---|
| dbo_idxFilmMedia_Actif | bigint |
| refMedia | bigint |
| dbo_Media_Actif | bigint |
| IstType | bigint |
| txtType | string |
| Status | double |
| ProcessId | double |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 39: crescine.bronze.cinando_raw_tfilms_temp

| Column | Type |
|---|---|
| cFilm | bigint |
| TitleVA | string |

| | |
|---|---|
| fVis | bigint |
| IstTypeFilm | bigint |
| txtTypeFilm | string |
| IstCategory | double |
| txtCategory | string |
| IstProductionStatus | bigint |
| txtProductionStatus | string |
| refCountry | bigint |
| refCompany | bigint |
| Company | string |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 40: crescine.bronze.cinando_raw_tfilms_temp_20250115

| Column | Type |
|---|---|
| value | string |

Table 41: crescine.bronze.cinando_raw_tidxfilmevent_temp_20250115

| Column | Type |
|---|---|
| uid | bigint |
| DateCreation | string |
| CreePar | bigint |
| refEvent | bigint |
| refFilm | bigint |
| refEventSection | double |
| awardDescription | string |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 42: crescine.bronze.cinando_raw_tidxfilmlanguage_temp_20250115

| Column | Type |
|---|---|
| uid | bigint |
| vid | double |
| CreePar | bigint |
| ModifiePar | double |
| DateCreation | string |
| DateModification | string |
| DateModificationAdm | string |
| refUtilisateurAdm | double |
| Actif | bigint |
| EtatFiche | double |
| refFilm | bigint |
| refLanguage | bigint |
| txtLanguage | string |
| NoOrder | bigint |
| rowhash | string |

| | |
|---|---|
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 43: crescine.bronze.cinando_raw_tidxgeneric_temp_20250115

| Column | Type |
|---|---|
| uid | bigint |
| vid | double |
| CreePar | bigint |
| ModifiePar | double |
| DateCreation | string |
| DateModification | string |
| DateModificationAdm | string |
| refUtilisateurAdm | double |
| Actif | bigint |
| EtatFiche | double |

| | |
|---|---|
| refFilm | bigint |
| refFilmCrew | bigint |
| IstType | bigint |
| txtType | string |
| NoOrder | double |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Table 44: crescine.bronze.cinando_raw_tidxgeneric_temp_20250115

| Column | Type |
|---|---|
| uid | bigint |
| vid | double |
| CreePar | bigint |
| ModifiePar | double |

| | |
|---|---|
| DateCreation | string |
| DateModification | string |
| DateModificationAdm | string |
| refUtilisateurAdm | double |
| Actif | bigint |
| EtatFiche | double |
| cLanguage | bigint |
| refCountry | double |
| NameFR | string |
| NameEN | string |
| Code | string |
| isActifListingFilms | bigint |
| isActifListingFilms2 | bigint |
| Name_1 | string |
| Name_2 | string |
| Name_3 | double |

At this stage, data were minimally processed and retained in near-raw form to preserve original fidelity and enable backtracking if needed.

Silver Layer (Transformation and Integration)

The Silver Layer consolidated the raw data into tables suitable for analytical queries and integration with external systems. To create the tables, data fields between Cinando and the master tables were normalized on both sides before matching.  Three key transformation outputs were obtained:

- cinando_clean_hasduplicates: Contains variables relevant for matching with other databases. It only includes entries classified as “Actif”, completed productions, or that have a production year. To ease linking the table contains ISAN, EIDR and the field refFilm (standardized as cinando_id). This table intentionally has duplicates as in the original database, which should be removed in matching, keeping the more complete entries. Different entries may link to different festivals though. The table contains a variable number_of_festivals that indicates how many entries are linked to a given refFilm ID in the festivals table. Please note that the same film can appear with multiple IDs in this table.
- cinando_films_filmmaster_joined: This table is a combination of the film master data and Cinando film data. The table contains different columns that enable it to be linked to other tables: cinando_id column allows for joining with other Cinando data objects (called refFilm as in the original source), crescine_id allows linking to all other Crescine datasets, and lumiere_id enables links to Lumiere and CresCine awards data, since it also contains it.
- cinando_films_festivals_filmmaster_joined: this table contains data on Cinando festivals that can be linked back to any other Crescine data via the crescine_id or to CresCine’s award data via the lumiere_id, for further analysis.

Table 45: cinando_clean_hasduplicates

| Column | Type | Description |
|---|---|---|
| title_original | string | |
| title_english | string | |
| production_year | bigint | |
| duration | string | |
| cinando_id | bigint | |
| refFilm | bigint | |
| Isan | string | |
| Eidr | string | |
| txtTypeFilm | string | |
| txtCategory | string | |
| logline | string | |
| synopsis | string | |
| spoken_language_main | string | |
| spoken_languages | string | |
| genres | string | |

| | | |
|---|---|---|
| production_country_main | string | |

Table 46: cinando_films_filmmaster_joined

| Column | Type | Description |
|---|---|---|
| cinando_id | int | |
| crescine_id | int | |
| title_original_cinando | string | |
| lumiere_id | string | |
| title_original | string | |
| title_english | string | |
| production_year | int | |
| production_country_main | string | |

Table 3: cinando_films_festivals_filmmaster_joined

| Column | Type | Description |
|---|---|---|
| crescine_id | bigint | |
| lumiere_id | string | |

| | | |
|---|---|---|
| cinando_id | bigint | |
| refEvent | bigint | |
| refEventSection | string | |
| awardDescription | string | |
| LibelleEvent | string | |
| idFestival | string | |
| txtCountry | string | |
| Year | string | |
| title_original_cinando | string | |
| title_original | string | |
| title_english | string | |
| production_year | bigint | |
| production_country_main | string | |

Gold Layer (Analytical Output)

The Gold layer aggregates curated data into analytical models that support descriptive and predictive analytics within FIDA. The Cinando information from the Silver Layer is integrated within the film master table.

The Gold layer integrates the following CInando data:

- Temporal and location data (e.g., event years, festival countries)
- Industry-level KPIs (e.g., festival selections, awards per title);
- Cross-database enrichment (e.g., linking with other tables for admissions or screening metrics).

The inclusion of Cinando data in the film master table of the Gold layer enables the creation of  analytical dashboardsand modeling pipelines that examine festival participation, awards performance, and festival screening cycles across markets.

CresCine Festival Data

Data collection and Processing

In FIDA, the processing of CresCine Festival Data refers to the transformation of three Excel files into analyzable data. The Excel file was based on data collected for the period 2014-2023 via web scrapper Octoparse from the festival web pages, festival web transcriptions if scrapping was not possible, and Wikipedia if the information was not available in the former two. The original Excel had two tabs:

- Participations and Nominations, containing festival, edition, section, title, director, production year.
- Awards, containing festival, edition, award, title, director, production year.

Quality Control

The QC stage for CresCine Festival aimed to maximize matchmaking between this source and the film master table, by performing:

1. Combining tables
    - The two tabs from the original file were combined in a master film list, with columns covering each of the fields previously separated in different tabs.

2. Matchmaking of incomplete records.
    - Matchmaking was performed via the Lumiere ID or via fuzzy matching.
    - To ensure that identifiers were consistent, Lumiere IDs in the original file were cross-checked with LumierePro identifiers.
3. Redundancy elimination and deduplication:
    - Identical records with the same fields(international_title, directors, country, platform, available_from, and available_to) were reduced to a single instance.

Bronze Layer (Raw Data Ingestion)

Data was ingested into the Bronze layer of the Databricks environment in a separate volume. This layer stores the data in Delta format, preserving its semi-structured nature.

Table 47: raw_awards

| Field | Type | Description |
|---|---|---|
| Festival | string | Name of the festival where the film was shown. |
| SFI Scoring | string | Scoring or rating given by SFI (if applicable). |
| Edition | bigint | Edition number of the festival. |
| Section | string | Festival section or program the film belongs to. |
| Award | string | Awards the film has received (if any). |
| English Title | string | Official English title of the film. |
| Original Title | string | Title of the film in its original language. |

| | | |
|---|---|---|
| Director | string | Name of the director. |
| Jury Status | string | Film’s status regarding jury selection or competition. |
| Lumiere ID | string | Lumiere database identifier for the film. |
| Ecosystem | string | Ecosystem classification (industry, market, etc.). |
| Prod. Year | bigint | Year the film was produced. |
| Prod. Country | string | Country or countries where the film was produced. |
| Director(s) | string | Full list of directors. |
| Other Titles | string | Alternative or working titles. |
| Description | string | A summary or synopsis of the film. |

Silver Layer (Transformation and Integration)

In the Silver Layer, the CresCine Festival table was enriched with a row hash for identification and trace back changes.

Gold Layer (Analytical Output)

The Gold layer aggregates curated data into analytical models that support descriptive and predictive analytics within FIDA. The CresCine information from the Silver Layer is integrated within the film master table.

The Gold layer integrates the following data:

- Temporal and location data (e.g., event years, festival name, edition)
- Industry-level KPIs (e.g., festival selections, awards per title);

The inclusion of CresCine festival data in the film master table of the Gold layer enables the creation of analytical dashboards and modeling pipelines that examine festival participation and awards performance across markets.

UsherU

Data collection and Processing

The collection and preprocessing workflow transformed the raw data into analyzable data. The data was received as a CSV export form UsherU, and contained information regarding the films, platforms, and catalogue availability.

Quality Control

The QC stage for UsherU aimed to maximize matchmaking between this source and the film master table, by performing:

4. Structural validation

    - Records with malformed or missing fields were removed.
5. Matchmaking of incomplete records.
    - Fuzzy string matching was performed for films missing an ID. By matching records via film characteristics (e.g. international title x year x country of origin triplet) it was possible to match incomplete records against the wider film master table.
6. External validation
    - To ensure that identifiers were consistent, retrieved tmdb_id and imdb_id were cross-checked through external APIs (notably TMDB).
7. Redundancy elimination and deduplication:
    - Identical records with the same fields(international_title, directors, country, platform, available_from, and available_to) were reduced to a single instance.

Through these measures, the dataset achieved both internal coherence and external referential integrity.

Bronze Layer (Raw Data Ingestion)

Data was ingested into the Bronze layer of the Databricks environment in a separate volume. This layer stores the data in CSV format, preserving its semi-structured nature while ensuring traceability and reproducibility.

Table 48: usheru

| Column | Type | Description |
|---|---|---|
| international_title | string | International film title in English |

| | | |
|---|---|---|
| original_title | string | Film title in original language |
| imdb | string | IMDB identifier |
| eidr | string | EIDR identifier |
| directors | string | Name of director |
| release_date | timestamp | Platform release date |
| platform | string | Name of Streaming Platform |
| country | string | Country of distribution |
| url | string | Link to catalogue |
| available_from | timestamp | Date of release in a platform |
| available_to | timestamp | Date of release in a platform |
| delivery_type | string | Type of streaming service |

Silver Layer (Transformation and Integration)

The Silver Layer consolidates and standardizes the raw data into relational tables suitable for analytical queries and integration with external systems.

Table 49: usheru_fact_movie

Contains harmonized cinema-level metadata for geographical and organizational analysis.

| Column | Type | Description |
|---|---|---|
| international_title | string | International film title in English |
| original_title | string | Film title in original language |
| imdb | string | IMDB identifier |
| eidr | string | EIDR identifier |
| directors | string | Name of director |
| release_date | timestamp | Platform release date |
| platform | string | Name of Streaming Platform |
| country | string | Country of distribution |
| url | string | Link to catalogue |
| available_from | timestamp | Date of release in a platform |
| available_to | timestamp | Date of release in a platform |
| delivery_type | string | Type of streaming service |
| unique_key | string | |
| rowhash | string | |
| source | string | |

| | | |
|---|---|---|
| loaded_at | timestamp | |
| updated_at | timestamp | |
| file | string | |

### Gold Layer (Analytical Output)

The Gold layer aggregates curated data into analytical models that support descriptive and predictive analytics within FIDA. The Gold layer integrates:

- Market-level KPIs (e.g., platforms per title, countries of distribution per title);
- Cross-database enrichment (e.g., linking with TMDB for genre and popularity metrics).

This layer powers analytical dashboards, and modeling pipelines that examine streaming distribution, platform density, and temporal exhibition trends across markets.

### LumierePro

### Data collection and Processing

Market information data was exported from the Lumiere Pro web service and then ingested into Databricks. The labeling convention, table names, and file names from the raw data was maintained, in order to create consistency for future data updates.

Data was organised into six tables covering different aspects of market performance:

- lumiere_pro_country_iso_codes: contains country ISO codes.
- lumierepro_market_data: contains location data
- lumierepro_ranking: contains box office data

- lumierepro_raw_admissions: contains admissions data
- lumierepro_raw_box_office:contains box office data
- lumierepro_raw_movies: contains basic film information details (e.g., title, year, genres, etc.).

This was the foundation of the Bronze Layer in the Databricks architecture.

Quality Control

To ensure interoperability, the following Quality Control actions were performed:

1. Elimination of incomplete records
    - Malformed records and rows with null values were eliminated.
2. Data harmonisation
    - Some field formats (e.g. currency, dates, order of production country etc.) were standardized to match other data sources.
    - Schemas were audited to identify and correct naming convention inconsistencies
3. Data organisation.
    - Columns featuring information for different fields were split into separate columns.
4. External validation
    - The IDs included in each record were cross-checked against TMDB in order to ensure coherence.

With these measures, the ingested data was ready for future transformations.

Bronze Layer (Raw Data Ingestion)

Data ingested into the Bronze layer of the Databricks environment was stored in Delta format. The ingested tables preserve the semi-structured nature of the raw data while ensuring traceability and reproducibility.

Table 50: lumiere_pro_country_iso_codes

| Field Name | Data Type | Description |
|---|---|---|
| ISO code | string | ISO country code |
| Country | string | Country name |
| All countries | string | Identifies country as part of the group “All countries” |
| Asia | string | Identifies a country as part of the group of countries located in Asia. |
| European Union after Brexit | string | Identifies a country as part of the group of countries located in Europe after Brexit. |
| European Union before Brexit | string | Identifies a country as part of the group of countries located in Europe before Brexit. |
| Europe (CoE) | string | Identifies a country as part of the group of countries located in Europe (CoE). |
| Latin America & Caribbean | string | Identifies a country as part of the group of countries located in Latin America & Caribbean. |
| Middle East & North Africa (MENA) | string | Identifies a country as part of the group of countries located in Middle East & North |

| | | |
|---|---|---|
| | | Africa (MENA). |
| North America | string | Identifies a country as part of the group of countries located in North America. |
| Oceania | string | Identifies a country as part of the group of countries located in Oceania. |
| Other Europe (non-CoE) | string | Identifies a country as part of the group of countries located in Other Europe (non-CoE). |
| Sub-Saharan Africa | string | Identifies a country as part of the group of countries located in Sub-Saharan Africa. |
| United Kingdom + Ireland | string | Identifies a country as part of the group of countries located in the United Kingdom + Ireland. |

Table 51: lumierepro_market_data

| Field Name | Data Type | Description |
|---|---|---|
| year | bigint | Year |
| market | string | Market country. |
| region | string | Market country region. |

| | | |
|---|---|---|
| providers | string | Data provider (e.g., BFI, MEDIA, Screen Ireland, etc.) |
| nb_films | bigint | Amount of films |
| admissions | bigint | Admissions. |
| coverage | double | |
| timestamp | timestamp | Data in which the data was loaded. |

Table 52: lumierepro_ranking

| Field Name | Data Type | Description |
|---|---|---|
| rank | int | |
| lumiere_id | int | Lumiere ID |
| imdb_id | string | IMDB ID |
| tmdb_id | string | TMDB ID |
| market | string | Market country |
| year | int | |
| original_title | string | Film title (original) |
| prod_year | int | Production year |

|  |  |  |
|---|---|---|
| prod_country | string | Production country |
| directors_gender | string |  |
| box_office_eur | int | Box office in euros. |

Table 53: lumierepro_raw_admissions

| Field Name | Data Type | Description |
|---|---|---|
| admissions | bigint | Admissions |
| year | bigint | Year |
| market | string | Market country |
| national | boolean |  |
| region | string |  |
| id | bigint |  |
| timestamp | timestamp |  |

Table 54: lumierepro_raw_box_office

| Field Name | Data Type | Description |
|---|---|---|
| rank | int |  |

| | | |
|---|---|---|
| lumiere_id | int | Lumiere ID |
| imdb_id | string | IMDB ID |
| tmdb_id | string | TMDB ID |
| market | string | Market country |
| original_title | string | |
| prod_year | int | Production year |
| prod_country | string | Production country |
| directors_gender | string | Film director’s genre |
| box_office_eur | int | Box Office in euros (€) |

Table 55: lumierepro_raw_movies

| Field Name | Data Type | Description |
|---|---|---|
| id | int | |
| original_title | string | |
| production_countries | string | |
| directors | string | |
| other_titles | string | |

| | | |
|---|---|---|
| links | string | |
| distributions | string | |
| total_admissions_obs | int | |
| genres | string | |
| prod_year | int | |

At this stage, data were minimally processed and retained in near-raw form to preserve original fidelity and enable backtracking if needed.

### Silver Layer (Transformation and Integration)

The Silver Layer consolidates and standardizes Lumiere Pro information in tables suitable for analytical queries and integration with external systems.  Each record was enriched with a row hash for identification and trace back changes. Eight different tables were created:

Table 56: lumierepro_admissions

| Field Name | Data Type | Description |
|---|---|---|
| id | int | |
| year | int | |
| market | string | Market country |
| national | boolean | |

|  |  |  |
|---|---|---|
| national_extended | boolean |  |
| region | string | Market country region. |
| admissions | int | Total of Admissions |
| created_at | timestamp | Date of creation. |
| rowhash | string |  |

Table 57: lumierepro_box_office

| Field Name | Data Type | Description |
|---|---|---|
| lumiere_id | int |  |
| imdb_id | string |  |
| tmdb_id | string |  |
| market | string |  |
| year | int |  |
| prod_year | int |  |
| prod_country | string |  |
| directors_gender | string |  |
| box_office_eur | int |  |

Table 58: lumierepro_countries_iso

In the Silver Layer, this table change the data type used to mark a country as part of group from string (X) to boolean (True, False) values.

| Field Name | Data Type | Description |
|---|---|---|
| iso_code | string | ISO country code |
| country | string | Country name |
| all_countries | boolean | Identifies country as part of the group “All countries” |
| asia | boolean | Identifies a country as part of the group of countries located in Asia. |
| eu_after_brexit | boolean | Identifies a country as part of the group of countries located in Europe after Brexit. |
| eu_before_brexit | boolean | Identifies a country as part of the group of countries located in Europe before Brexit. |
| europe_coe | boolean | Identifies a country as part of the group of countries located in Europe (CoE). |
| latin_america_caribbean | boolean | Identifies a country as part of the group of countries located in Latin America & Caribbean. |

| | | |
|---|---|---|
| middle_east_north_africa_mena | boolean | Identifies a country as part of the group of countries located in Middle East & North Africa (MENA). |
| north_america | boolean | Identifies a country as part of the group of countries located in North America. |
| oceania | boolean | Identifies a country as part of the group of countries located in Oceania. |
| europe_other_non_coe | boolean | Identifies a country as part of the group of countries located in Other Europe (non-CoE). |
| sub_saharan_africa | boolean | Identifies a country as part of the group of countries located in Sub-Saharan Africa. |
| uk_ireland | boolean | Identifies a country as part of the group of countries located in the United Kingdom + Ireland. |

Table 59: lumierepro_distributions

| Field Name | Data Type | Description |
|---|---|---|
| id | int | |
| release_date | date | |

| | | |
|---|---|---|
| first_release_date_movie | date | |
| last_release_date_movie | date | |
| company | string | |
| country | string | |
| rowhash | string | |

Table 60: lumierepro_market_data

| Field Name | Data Type | Description |
|---|---|---|
| market | string | |
| year | int | |
| region | string | |
| providers | string | |
| nb_films | int | |
| admissions | bigint | |
| coverage | double | |
| created_at | timestamp | |
| rowhash | string | |

Table 61: lumierepro_movies

| Field Name | Data Type | Description |
|---|---|---|
| id | int | |
| imdb_id | string | |
| original_title | string | |
| directors | string | |
| other_titles | string | |
| url_and_external_id | string | |
| distributions | string | |
| total_admissions_obs | int | |
| film_type | string | |
| film_make | string | |
| obs_genre | string | |
| prod_year | int | |
| directors_gender | string | |
| rowhash | string | |

| | | |
|---|---|---|
| created_at | timestamp | |
| wiki_id | string | |

Table 62: lumierepro_other_titles

| Field Name | Data Type | Description |
|---|---|---|
| id | int | |
| name | string | |
| is_original | boolean | |
| country_code | string | |
| created_at | timestamp | |
| rowhash | string | |

Table 63: lumierepro_production_countries

| Field Name | Data Type | Description |
|---|---|---|
| id | int | |
| production_country | string | |
| order_of_appearance | int | |

| | | |
|---|---|---|
| number_of_production_countries | int | |
| created_at | timestamp | |
| rowhash | string | |

In addition to these tables which included only Lumiere Pro data, a joint table was created in order to create a first version of the Master Table of the FIDA Gold Layer.

Table 64: movies_similarity_joined

| Field Name | Data Type | Description |
|---|---|---|
| title_original | string | |
| title_english | string | |
| production_year | double | |
| production_country_main | string | |
| release_date | string | |
| lumiere_id | double | |
| tmdb_id | int | |
| wiki_id | string | |
| imdb_id | string | |

Gold Layer (Analytical Output)

The Gold layer aggregates curated data into analytical models that support descriptive and predictive analytics within FIDA. Lumiere Pro is one of the two backbone components of the FIDA Gold Layer by providing:

- Market-level KPIs (e.g., admissions, box-office);
- Location information (e.g., admissions country, country of origin):
- Cross-database interconnectivity (e.g., TMDB ID, Lumiere ID, IMDB ID).

Thus, Lumiere Pro is one of the two critical components that delimit the FIDA data pool as well as a critical element which enables interconnecting other databases ingested.

International Showtimes

Data collection and Processing

The collection and preprocessing workflow involved multiple stages to transform raw API responses into structured, analyzable data. Initially, data were retrieved from the Showtimes International API, which returns semi-structured JSON documents containing information about movies, cinemas, and showtimes.

The Movie Matcher service, also provided by Showtimes International, was used to retrieve unique movie_id values by querying the endpoint with the movie title. The results typically include additional metadata such as alternative titles, directors, release dates, and, when available, the tmdb_id.

Raw responses were stored in local JSON files and subsequently imported into a MongoDB database via a custom local API, enabling controlled ingestion into the Databricks data lake. During this phase, redundancy elimination was applied to remove multiple identical entries for the same (cinema × movie × screening) triplet.

Two base MongoDB collections were created:

- cinema – containing all known cinema metadata (name, address, country, coordinates, etc.);
- showings – containing detailed screening information with movie references and timestamps.

These formed the foundation for the Bronze layer in the Databricks data architecture.

Quality Control

Given the heterogeneity of the raw API responses, a comprehensive Quality Control (QC) pipeline was implemented to ensure structural and semantic integrity.

1. Structural Validation:

   Records with malformed or missing fields—particularly inconsistent or invalid start_at timestamps—were filtered and stored in a dedicated invalid__ directory.

2. Identifier Enrichment:

   Since raw data often lacked identifiers (movie_id, cinema_id, and showtimes_id), the QC process integrated API calls to retrieve missing information:

   - The Movie Matcher endpoint was queried repeatedly until a valid movie_id was obtained.
   - For cinema identification, a fuzzy matching pipeline was applied, comparing attributes such as latitude, longitude, slug, address, and zipcode against the official Showtimes cinema registry per country.
     - Geolocation-based matching was prioritized when valid coordinates were available.

- Fuzzy string matching was used when only textual data existed, with similarity thresholds of 100% for slug and zipcode, and 95% for address.

3. External Validation:

   Retrieved tmdb_id and imdb_id were cross-checked through external APIs (notably TMDB) to validate identifier consistency.

4. Redundancy Elimination and Deduplication:

   Identical screenings—defined by the same (country, movie_id, cinema_id, start_at)—were reduced to a single instance (showings).

Through these measures, the dataset achieved both internal coherence and external referential integrity.

Bronze Layer (Raw Data Ingestion)

Data ingested into the Bronze layer of the Databricks environment was stored in Delta format. The ingested tables preserve the semi-structured nature of the raw data while ensuring traceability and reproducibility.

Two main tables were created:

Table 65: showtimes_raw_cinemas

| Column | Type | Description |
|---|---|---|
| cinema_id | string | Unique cinema identifier |
| name | string | Cinema name |
| address | string | Cinema address |

| | | |
|---|---|---|
| country | string | Country code |
| lat | double | Latitude |
| lon | double | Longitude |
| slug | string | URL-friendly cinema identifier |
| chain | struct | Nested JSON with chain_id |
| loaded_at | timestamp | Data ingestion timestamp |

Table 66: showtimes_raw_showings

| Column | Type | Description |
|---|---|---|
| show_id | string | Unique showing identifier |
| date | string | Screening date |
| cinema | struct | Contains cinema_id and country |
| movie | struct | Contains is_id, title, tmdb_id, imdb_id |
| page | string | Source page identifier |
| loaded_at | timestamp | Data ingestion timestamp |

At this stage, data were minimally processed and retained in near-raw form to preserve original fidelity and enable backtracking if needed.

Silver Layer (Transformation and Integration)

The Silver Layer consolidates and standardizes the raw data into relational tables suitable for analytical queries and integration with external systems, which resulted in the creation of two tables.

Table 68: showtimes_dim_cinemas

Contains harmonized cinema-level metadata for geographical and organizational analysis.

| Column | Type | Description |
|---|---|---|
| id | int | Cinema identifier |
| name | string | Cinema name |
| address | string | Cinema address |
| country | string | Country code |
| lat | double | Latitude |
| lon | double | Longitude |
| slug | string | URL-friendly cinema identifier |
| chain_id | int | Reference to cinema chain |

Table 69: showtimes_fact_showings

Contains normalized screening-level data linking cinemas and movies

| Column | Type | Description |
|---|---|---|
| id | string | Unique showing identifier |
| date | date | Screening date |
| cinema_id | int | Foreign key to showtimes_dim_cinemas |
| cinema_country | string | Country of cinema |
| is_id | int | Internal movie identifier |
| title | string | Movie title |
| tmdb_id | bigint | TMDB identifier |
| imdb_id | string | IMDB identifier |

Gold Layer (Analytical Output)

The Gold layer aggregates curated data into analytical models that support descriptive and predictive analytics within FIDA. At this stage, each screening is uniquely identified by the tuple (country, cinema_id, movie_id, start_at).

The Gold layer integrates:

- Temporal aggregations (e.g., daily, weekly, and yearly screening counts);
- Market-level KPIs (e.g., screenings per cinema, screenings per title);

- Cross-database enrichment (e.g., linking with TMDB for genre and popularity metrics).

This layer powers analytical dashboards, recommendation engines, and modeling pipelines that examine audience behavior, screening density, and temporal exhibition trends across markets.

European Audiovisual Observatory Yearbook

Data collection and Processing

Data was retrieved for the EAO Yearbook Online Service, which was stored in an Excel table containing information about average ticket price, country ISO code, and year. Raw data was subsequently imported into the Bronze Layer of the Databricks data lake for further transformation.

Quality Control

While data retrieved was well organized, some Quality Control (QC) activities were performed in order to ensure interoperability.

1. Standardization of data.
    - The field *country* was cross-checked against the aggregated country list in order to ensure compatibility and support interoperability with other tables.
    - The data was reviewed in order to confirm it had a consistent number format
2. Consistency Check
    - The data across years within a single country was compared to detect anomalies or missing values.

Bronze Layer (Raw Data Ingestion)

Data was ingested into the Bronze layer of the Databricks environment in a separate volume (market). The data was stored in a table in  Excel format, preserving the original field names and field conventions to ensure traceability and reproducibility.

Silver Layer (Transformation and Integration)

The information from the Bronze layer was transformed in order to create the following output:

- Crescine.silver.ticket_prices, which details the currency used, country (in ISO code), average ticket price by year, update timestamps and a rowhash.

Table 70. ticket_prices

| Column | Type |
|---|---|
| Country | string |
| currency | string |
| Year_2014 | double |
| Year_2015 | double |
| Year_2016 | double |
| Year_2017 | double |
| Year_2018 | double |
| Year_2019 | double |

| | |
|---|---|
| Year_2020 | double |
| Year_2021 | double |
| Year_2022 | double |
| Year_2023 | double |
| Year_2023_to_2022 | double |
| rowhash | string |
| source | string |
| loaded_at | timestamp |
| updated_at | timestamp |
| file | string |

Gold Layer (Analytical Output)

The Gold layer integrates the EAO Yearbook data as a dimension table thus allowing:

- Analyzing the relationships between admissions, box-office and ticket prices.
- Enriching the predictions made based on historical market data.

Therefore, this information helps to deepen the study of theatrical markets within Europe and provides context related to admissions and box-office data.

Linked References