Digital Film Listings (DIGIFIL)

Final Report

(version 25 May 2020)

I. Kisjes, T. van Oort, K. Lotze, I. Staliunaite, J. Veerbeek, K. Beelen, J. Noordegraaf

CLARIAH
Common Lab Research Infrastructure
for the Arts and Humanities

create
creative Amsterdam
an E-Humanities
perspective

Table of contents

1. Introduction[1]

DIGIFIL (DIGITAL FILM LISTINGS (CLARIAH, 2018)) aims at automatically extracting, digitizing and publishing film screening data from the weekly *Filmladders* (or films listings) as published in historical newspapers. Besides the listings, DIGIFIL aims to extract contextual information about the wider movie landscape as reported in historical newspapers (such as movie reviews and descriptions). The screenings constitute the focal point of film culture: they are the place where distributors, exhibitors and audiences meet. Collecting information about these encounters, and embedding them in their wider discursive context, yields an invaluable resource for linguists, socio-economic historians and media scholars to study the ways in which cinema-going contributed to the formation of modern societies.

Socio-economic historians and media scholars have for decades studied the various ways in which cinema-going contributed to the formation of modern societies.[2] Manual attempts to gather data on film listings have proven to be time-consuming and resulted in case studies that focus on a particular place and a limited period.[3]

The first large attempt to collect and digitally present film screening data for the Netherlands was initiated by Karel Dibbets within the framework of the large-scale research project “Cinema, Modern Life and Cultural Identity in the Netherlands, 1896-1940” (NWO, 2002-2006), which resulted in the development of Cinema Context (cinemacontext.nl). Cinema Context is an online encyclopedia and research instrument for studying the history of film culture in the Netherlands.[4] Currently, it contains the data for 107.235 screenings of 45.623 films in 400 Dutch cities, towns and villages (cinemacontext.nl; 20 March 2019). In addition, Cinema Context contains the names of all cinemas that have been active in the Netherlands. While the period from 1896 to 1940 is largely covered, film screening data for more recent decades are missing.

By focusing on the postwar movie landscape, DIGIFIL extends the work of film historian Karel Dibbets on “Cinema Context”. In doing so, DIGIFIL demonstrates how semi-structured text—the agendas published in newspapers—can be automatically converted to entries in a database (which, so far, has been manually compiled). The project provides scholars who work on similar sources with various computational tools and best practices to transform their *own* sources to machine readable and manipulable data.

[1] **Acknowledgements**: This research was made possible by a CLARIAH-CORE project financed by NWO, see http://www.clariah.nl, project number CC-WP3-18-004. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. Project id was e-infra180150.

[2] See, for example, the overview provided by Biltereyst, Maltby & Meers 2012 and 2018.

[3] For the Netherlands in the 1930s, see, for example, Sedgwick, Pafort-Overduin & Boter 2012, for the Flemish cities of Gent and Antwerp between 1925 and 1975, see Meers, Biltereyst & Van de Vijver 2010. On prewar Germany, see Garncarz, 2015. For a comprehensive overview of recent research projects focusing on local histories of film exhibition and consumption see the website of the History of Moviegoing, Exhibition, and Reception (HoMER) network: http://homernetwork.org/dhp-projects/homer-projects-2/?viz=1; last access 26 March 2019.

[4] For details see Noordegraaf, Lotze & Boter 2018, Dibbets 2010,

For its input data, DIGIFIL relies upon the digitization effort of the Royal Library (*Koninklijke Bibliotheek,* henceforth KB). Their current collection, available via Delpher, already contains an impressive set of digitized, segmented and enriched newspapers. The point of DIGIFIL is to improve digitization and enrichment of *specific sections in the newspaper corpus*. We use the available digitized materials as a starting point but refine and extend.

This report mainly describes the computational techniques we applied (and invented) to automatically extract film listings from newspapers and transform these to structured and semantically enriched data, which can be propagated to existing scholarly databases such as Cinema Context. The overview below gives a detailed account of the methods used and tools developed within the DIGIFIL project. This report also serves as a manual for other researchers who wish to make use of the scripts we created and enhance the data we generated.

2. Data & Methods

2.1. Collecting Newspaper Data

As DIGIFIL focused on the postwar period, we started with sourcing the major national newspapers for the period 1948-1995 from Delpher. This resulted in a collection of around 20.650.000 articles from *De Tijd*, *De Telegraaf*, *De Volkskrant*, *De Waarheid*, *Algemeen Handelsblad*, *Trouw*, *Het Parool*, and *NRC*. All articles were indexed with a PostgreSQL database, which allows for rapid querying and processing of such an immense amount of information.

2.2. Harvesting and Structuring Film Listings

2.2.1 Identifying the listings

After collecting the articles from Delpher, we set out to automatically identify the weekly film listings in the digitized newspapers. Firstly, we attempted to use the article titles to isolate movie listings. We created folders for each article title, in which we saved all the articles with that corresponding title. However, this effort wasn’t completely satisfactory as

1) there was too much variation in the literal article title string (sometimes including, for example, the date of an article);

2) the article titles of movie listings varied considerably over time and between the different newspapers;

3) article titles were not always consistent with article content. Sometimes, for example, the segmentation used by the KB lumped multiple articles together, on other occasions, it split

them where they should be joined;

2.2.2 Locating the listings in space

Because such a simple strategy failed to produce the desired results (i.e. it could not separate movie listings from non-movie listings), we tried to assign each candidate listing a “cinema score” that indicates the likelihood of an article being an actual listing for a specific city. We counted the occurrences of cinema names that we knew had been active in that city at a particular point in time. We used an export from the Cinema Context database, to group cinemas by cities.

To identify the city where a movie was screened, we assigned articles to a particular city using a classifier based on a bag-of-words representation—assuming that a document’s ngrams contain strong enough lexical signals to classify a *Ladder* by place. However, things turned out to be more complicated as the movie listings frequently covered multiple cities: some cover an area as wide as The Hague, Amsterdam, Rotterdam etc. So this idea was discarded, but relics of the results were still used in the next step. An artefact of this strategy is the 'city' query in the file '**biosscores.py**'.[5]

The script '**biosscores.py**' calculates the scores for the articles in the database.[6] These scores were then forwarded to the script that created the training files for tagging the content of the listings (see below) in order to obtain a cleaner set of listings and avoid that annotators have to plough through swaths of irrelevant data. This has proven to be rather effective in weeding out unwanted articles.

2.2.3 Identifying titles

Besides the name of a cinema, the presence of movie titles provides critical clues for determining whether or not an article belongs to the category of *Filmladders*. We enhanced the cinema score by estimating, for each article, the number of titles it contains (and, as noted above, used these clues as additional information for selecting training files that had to be manually annotated (see below. 2.2.3)). The main workhorse for this task **'titlescores.py'**, which only retrieves movie titles from a particular year from the database and scores each article according to the number of occurrences of movie titles in the article. Unfortunately, the procedure proved less effective than anticipated, partly because the OCR was often far from perfect. This scoring mechanism did find some articles that were definitely listings, but only the

[5] All codes can be found at https://gitlab.com/uvacreate/digifil

[6] This should, in light of new insights, be slightly rewritten to calculate scores for multiple cities instead of the one city assigned to it in the database.

very high scores turned out to be a reliable measure of “Filmladder-ness”—for the lower ranges, the results were less stable.

2.2.4 Tagging the listings’ content

Besides automatically identifying whether an article constitutes a Filmladder, we had to create an efficient procedure for segmenting listings into various elements of interest, such as cinema name, screening time, or movie title. We conceived this task as being similar to part-of-speech tagging, but instead of tagging elements by their word type, we classify tokens by their function within the film listing. To save time—and circumvent the costs of training a new model—we retrained an out-of-the-box tagger taken from Spacy (a very popular NLP library in Python).

The script used to train the POS-model is called '**tagtrain.py**'. It takes a tokenized movie listing as input, stored as a CSV file that maps tokens to a semantic tag. We created a set of tags that captured the structure of the film listings in various newspapers. For example 'B' stands for cinema, 'T' for movie title, 'H' for temporal data (the full set of tags are defined in **‘tagtrain.py**’).

Creating a sufficiently large set of training examples, required input from multiple annotators. In order to facilitate annotation, we created a script ('tagtrain_xls.py') that would grab a set of listings from the database and convert them into Excel spreadsheets, with one movie listing article per spreadsheet. All rows are colour-coded (one colour for each tag, which helped spotting errors quickly). The thus generated Excel files were manually corrected and used iteratively to retrain and finetune the model.

2.3 Refining and improving the pipeline

2.3.1 The Movieparser class

After producing a basic dataset of annotated film listings, we trained a classifier that grasped more adequately the structure of the data we were working with. We furthermore optimized the pipeline based on information we obtained during the process. This produced the file **'parserclass.py**', which contains a single class named Movieparser that provides methods for finding and parsing movie listings. Users can instantiate the class, passing the city and a year as arguments to look for listings that match these requirements.

The script makes use of Cinema Context data, such as movie title and cinema name (by city and year). It also relies on other information such as a list of street names (which commonly follow the name of a cinema) and frequent variations on cinema names (caused by the failing of the OCR engine). We use a master list of common misspellings, as we observed that fuzzy matching yielded too many false positives. Additionally, we applied a list of movie titles that included Dutch movie titles from the 1960s, 1970s and 1980s.

We recommend running '**find_latin.py**' after instantiating a Movieparser class. This script removes any non-Latin movie titles for the cached title set. The fewer titles the better, when it comes to fuzzy matching.[7]

The script '**prepare_per_year.py**' can be used to pre-populate the listings for a particular city in a particular year with the movie listing articles for that year. After creating a folder with the name of the city in the root folder[8], the script tries to find at least one movie listing for each week, which is followed by a second, somewhat less restrictive query for the weeks for which no listings were returned. The script '**find_listings.py**' takes the output of this step to capture the content of the listings and dump the information in a CSV file.

Optionally, '**check_cinemanames.py**' compiles any suspicious cinema names returned by 'find_listings.py' and stores these in a CSV file, that can be checked manually later on (and copied into 'extra-data/cinema_name_variations.csv'). This stage concludes with validating and filtering the retrieved listings, which discards remaining non-movie-listing-articles. This process is handled by '**filter.py**'.

2.3.2 Segmenting Titles

It turned out that, especially in the 1980s and 1990s, many movie titles were lumped into a 'title' for a single cinema – sometimes up to twelve titles would be listed as one. Separating those titles by segmentation based on common markers such as 'en' or ' om 10 uur' failed because titles were regularly separated by just the number of the theatre auditorium (e.g.: City 1, 2, 3, etc.). Distinguishing titles requires a lot of real-world (semantic) knowledge and is therefore difficult to accomplish with computational means. Knowing, for example, that the string “Rambo 3 The Beginning” contains two titles, is a non-trivial task (even for humans).

In order to tackle this issue, we constructed a conditional random fields classifier to segment the titles. First, we had to create some training material consisting of titles collated into one string. Toward this end, we created '**crf_prepare_titles.py**' which retrieves titles from the database and writes them to a file. Subsequently '**crf_prepare.py**' randomly merges titles into training strings, after which ‘**crf_train.py**’ trains the actual model on the generated examples. Based on 100.000 examples, it reported a 95% correct segmentation of the titles. The model is stored in the 'extra-data' directory. Once the model is trained, '**label_crf.py**' prepares the OCR'ed data for classification, after which '**crf_skl.py**' loops over all instances and separates them into the actual titles.

[7] When creating the Movieparser class, we assumed that cinemas had 'screening weeks' that started on Thursdays (meaning that in some places it is set up to look, for instance, for titles playing in a cinema in a particular week). Later on, we found out that these weeks were not completely consistent and would not always start on Thursdays. This necessitates that those parts of the code have to be adjusted.

[8] We did have to normalize 'Den Haag' and 'sGravenhage' to 'sgravenhage' for the Movieparser class to pick up the listings (this was just hardcoded into the class).

2.3.3 Identifying and Linking Titles

Completing all these steps allowed us to start identifying and linking the titles. Here, '**match.py**' serves as the main workhorse. This script fuzzily matches a large set of titles shown in a particular year. Since this step is computationally expensive, we do not simply match all the listings, but first compile them into a list of unique strings which are subsequently matched. The compiled titles are saved to a pickle file in the folder 'matchmap'.

The script '**matchmap2csv.py**' converts the pickle file to CSV format for manual checking. It also normalizes the identifiers from the Internet Movie Database (IMDb).[9] We opted for IMDb as a masterlist since it currently figures as the most comprehensive online database for information related to film production—and allows us to assign commonly used unique identifiers to movie titles.

Running this script yields a CSV file in the 'matchmap/check' folder for the year selected. To monitor the quality of the linking process, identifiers attributed to titles are manually validated; e.g. setting the value in the 'isok' column '1' signals that the title is correctly identified. This facilitates validation of the individual listings per city at a later stage.

The script '**splice_oks_in_result.py**' digests the output stored in 'matchmap/check' to update the CSV files per city. Along the way it also validates titles according to context: if a very similar title did play within seven days of an identified one in the same cinema, it is automatically corrected to the same title. As titles tend to be screened multiple times this saves a large amount of manual correction.

Finally **'profile_cin.py'** validates the listings for cinemas in a particular year. It estimates the median day when films are changed[10] and shows where multiple titles per week were screened, thus making it easier to find mistakes and extra screenings in the final CSV listings.

3. Inspecting and Evaluating the DIGIFIL Dataset

3.1. Completeness

To estimate the completeness of the harvested *Filmladders*, we identified all “weeks” (i.e a contiguous period of seven days) in which no listing appeared in the dataset for a specific city. We then calculated how many days fell within a period of such listing drought. For Amsterdam, the missing percentage was 4.6%, for Rotterdam 3.1%, for Den Haag 5.4% and for Utrecht

[9] www.imdb.com

[10] Either once or twice a week for now, we will need to expand that for the later periods.

7.4%. 1948 and 1949 are the years for which the most listings are missing: for Amsterdam 35.4%, for Rotterdam 6.4%, for Den Haag 17.8% and for Utrecht 63.7%.

3.2. Redundancy

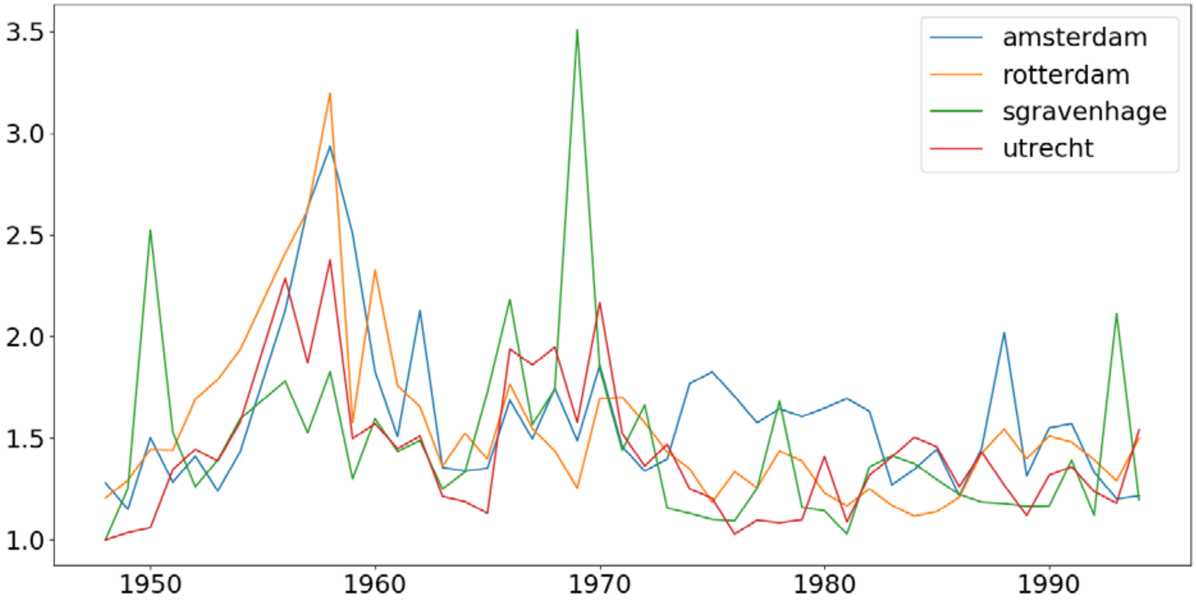


Figure 1. The average number of unique movie listings per day for each city per year.

For most years and cities, we collected about 1.5 listings on average for any given week. Taking into account that there are weeks for which we have no data at all, the average number of listings per screening-week is closer to 2. Figure 1 plots the average number of unique movie listings per day (split by city).

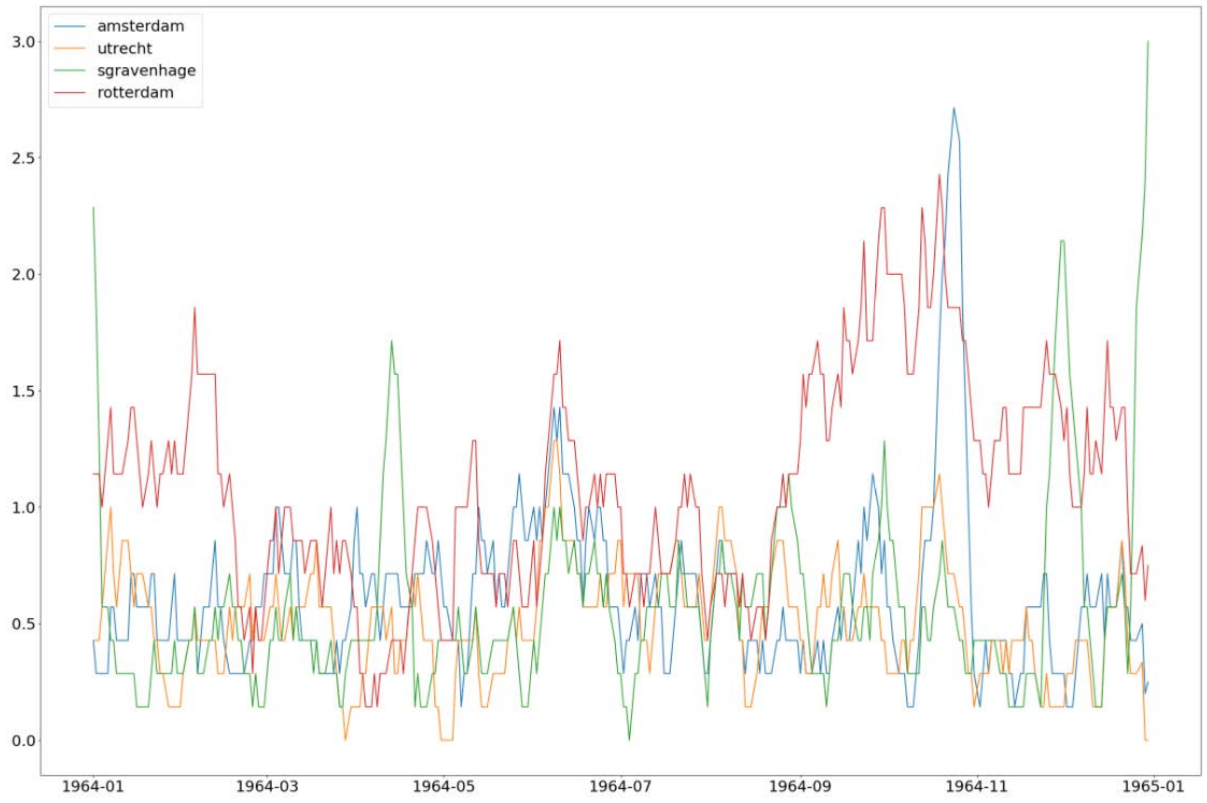


Figure 2. The average number of unique movie listings in 1964 per day for each city.

Figure 2 plots the temporal evolution for a single year (in this case 1964) to inspect variation on a more fine-grained (weekly) level. As can be observed, few weeks lack data, while others contain many listings.

3.3. Benchmarking: Comparing against a golden standard

3.3.1 Rotterdam

Besides these metrics, we further gauged the quality of the dataset by comparing a subset to a golden standard—in this case, a manually collected list of data which we know is correct. We decided to use a subset of programming data for Rotterdam cinemas for the period 1951-1953, which had been created by Thunnis van Oort. Checking the raw output of DIGIFIL against the golden standard shows that:

1) Out of 395 listings for weekly cinema unique film listings, 295 are correctly identified automatically, for the correct cinema in the correct week.
2) 126 films that were marked as screened in the golden standard dataset were not picked up correctly by the script.

3) 296 false positives (titles that the output marks as shown that were not shown according to information based on the golden standard dataset). Most of these are relatively easily correctable; for example, in week 41 Capitol is screening “Het teken van Don Marcos”, which in one of the 10 listed DIGIFIL entries is misidentified as “Jøden Süss” because of a trailing bit in: “Het teken van Don Marcos . en sSsüt” which was identified as a second listing for that theatre.

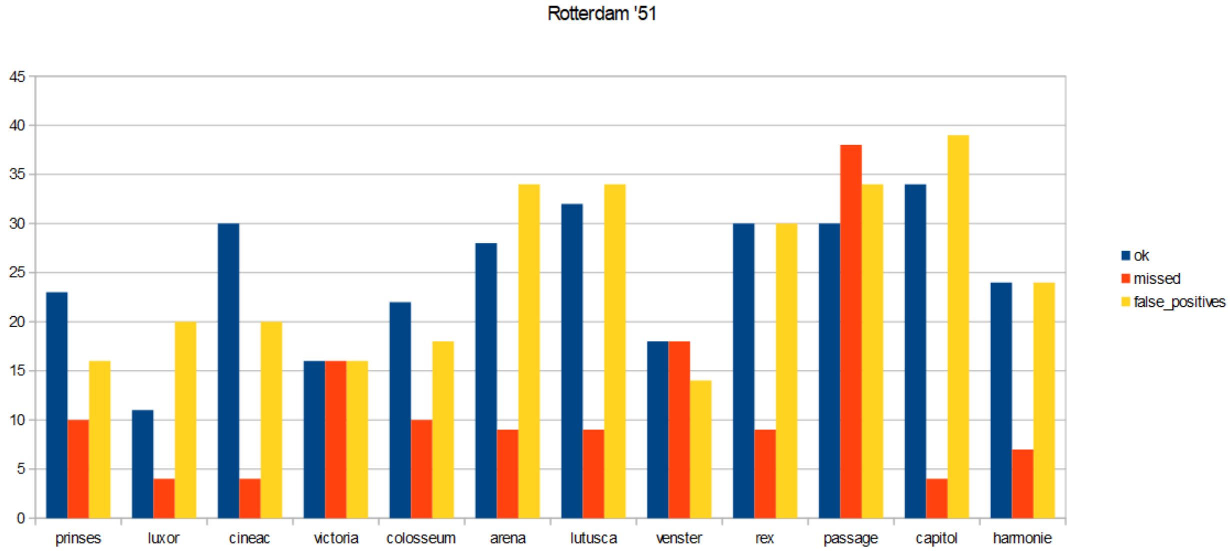


Figure 3: Film identification correctness and missing data overview for Rotterdam 1951 per cinema.

3.3.1.1 Rotterdam: Error-analysis

Some errors can be attributed to the structure of the Ladder: oftentimes multiple cinemas are listed as the venue for the same title (e.g. “Colosseum, Luxor en Victoria” screen the same film). This type of notation was rare enough that we didn’t account for it in the code. Also, a few film titles are simply not yet recorded in IMDb (e.g. ‘Vrede voor de wereld’, ‘La tour de Babel’ (France, 1951)) Another issue was that similar, but not identical film titles, are being identified as different films (e.g. ‘De wolf van Sila’ being identified not just as ‘De wolf van Sila’ but also as ‘De wolf rail’ when the OCR reads ‘De wolf rail Sela’).

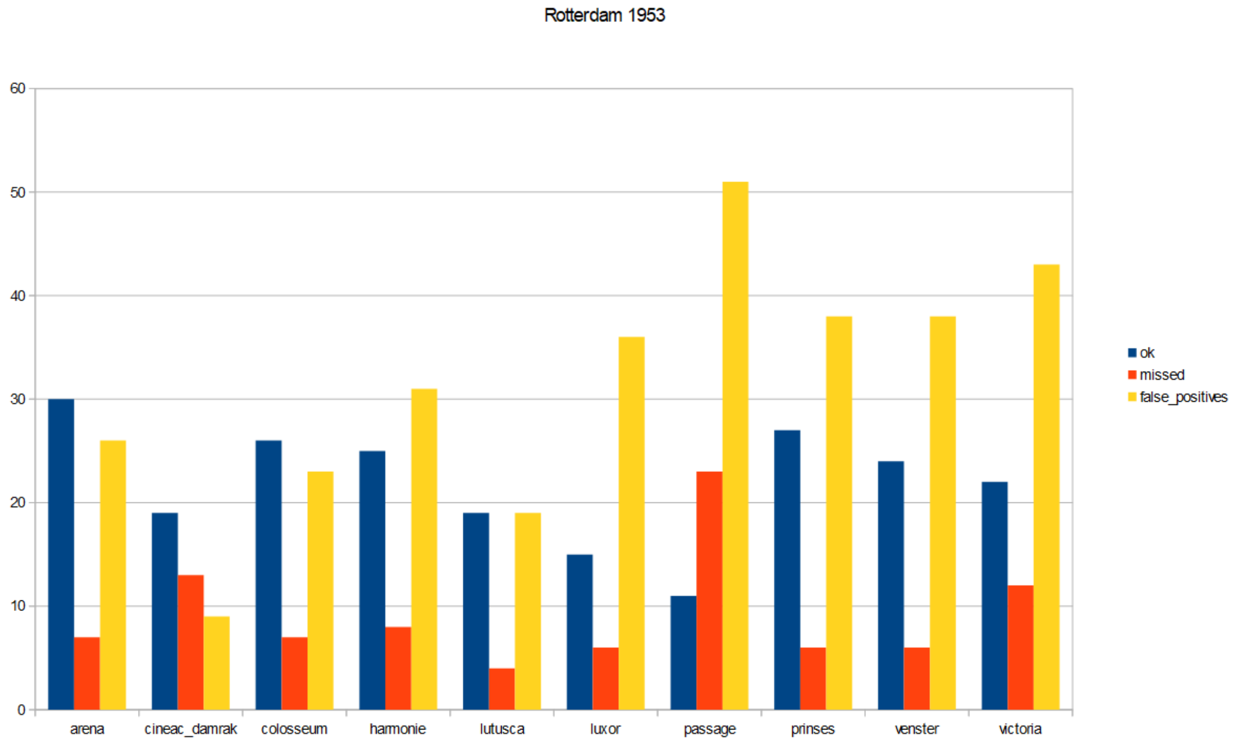


Figure 4: Film identification correctness and missing data overview for Rotterdam 1953 per cinema.

Figure 4 plots the accuracy and missing data overview for Rotterdam 1953—which looks similar to 1951 overall. Cinema Rex has an enormous number of false positives, not only because the OCR engine has frequently misrecognized the name of the cinema name but also because it appears to be a very common name for cinemas (therefore difficult to disambiguate): many of the entries for ‘Rex, Rotterdam’ actually listed titles played at other venues named “Rex” in the Netherlands.

The high amount of false positives for Cinema Passage is attributable to confusion with an eponymous venue in The Hague.[11] Overall, for ‘53, out of 390 screenings listed, 275 were correctly identified, with 115 missed or misidentified titles and 433 false positives (mostly attributable to Rex and Passage).

3.3.2 Utrecht

[11] The one contained in the golden standard dataset is actually located in the municipality of Schiedam, not Rotterdam, and is often listed under the heading ‘Schiedam’ in the newspapers and therefore not included consistently in the Rotterdam listings.

Besides Rotterdam, we also inspected data for Utrecht, again by comparing the generated data against a manually corrected list: for these tests we limited our data to the cinemas listed in Cinema Context as active in the relevant year and city.

3.3.2.1 Utrecht: Error-analysis

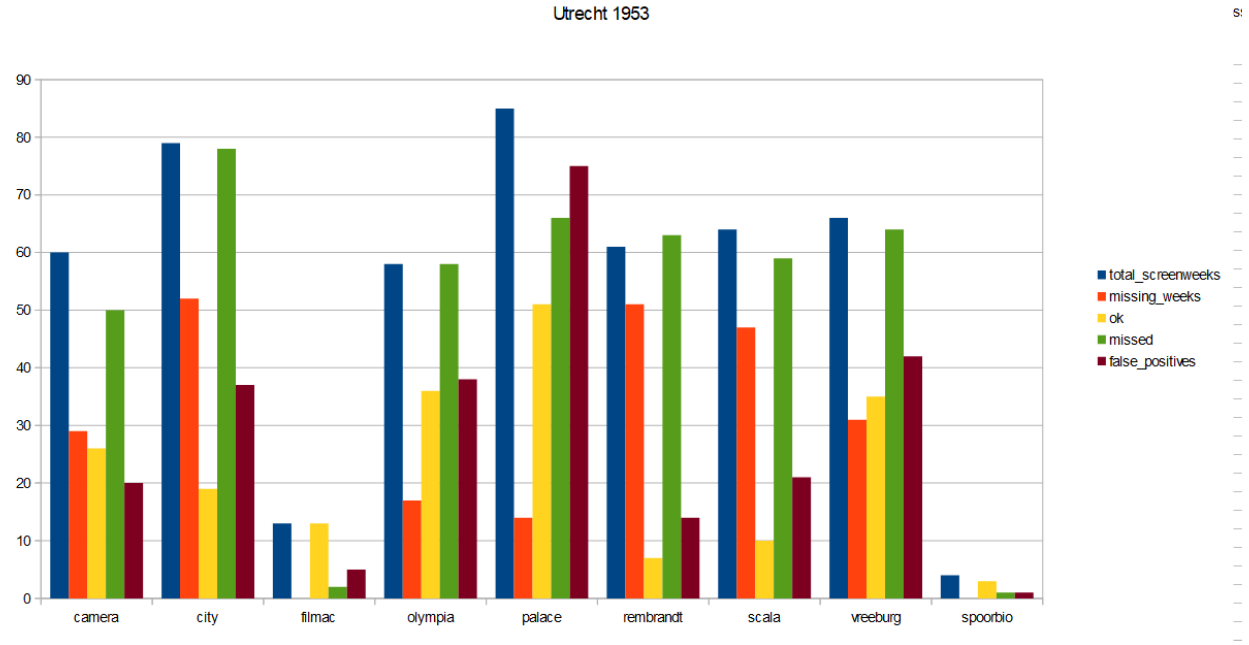


Figure 5: Film identification correctness and missing data overview for Utrecht 1953 per cinema

Scrutinizing the output for Utrecht in 1953 we encountered multiple issues. Cinemas Filmac and Spoorbio mostly showed newsreels, i.e. film titles that could not be identified and, consequently, are missing in this figure. The figure shows that listings for cinema names that are more unique to Utrecht (e.g. Olympia, Vreeburg, Palace, Camera) produce better results than more common names (e.g. City, Rembrandt, Scala). “City” is especially problematic because the word also occurs in film titles (“City of Bad Men”, “Naked City” etc.) and, as a name, it is rather short and contains multiple characters that are problematic for OCR (‘i’, ‘t’, and ‘y’). [12]

Furthermore, Utrecht is quite a challenging city in terms of the cinema names, as there are relatively few cinemas and only about three of those are common enough to distinguish the city from The Hague or Amsterdam (see above. 2.2.2). When a listing comprises multiple cities, those are mostly ranked in alphabetic order. Utrecht, therefore, tends to appear at the very

[12] ABBYY’s OCR engine - the software the KB used to digitize the text - often mangles the name “Asta” into for example “als” (we suspect because the language for converting scans to text was set to Dutch). Because “City” is a very short word, it is also difficult to use a large OCR correction map (with misread versions of ‘City’ mapped to the correct version) as this would result in too many terms that actually are not ‘City’ being corrected to the cinema ‘City’. This is a problem we also encountered with, for example, the cinema name ‘Asta’: if we included something like ‘ast’ as a variant, too often the Dutch word ‘als’ would be corrected to ‘Asta’. Hence the need for a manually curated list of variants.

bottom, which has as a consequence that when the page segmentation occasionally cuts off the bottom of an article, the text is clipped and titles are more likely to be misidentified.

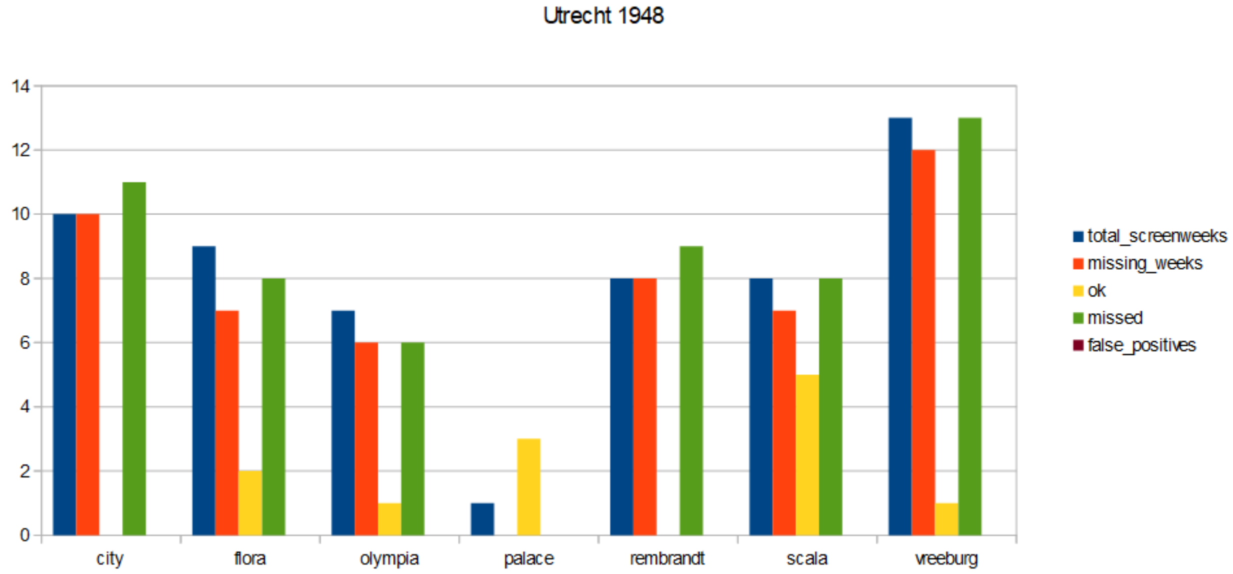


Figure 6: Film identification correctness and missing data overview for Utrecht 1948 per cinema

The number of listings found for Utrecht in 1948 is much smaller, but the pattern in the results (and errors) is very similar to those observed in 1953: listings for unique cinema names are better recognized than common names. However, the figure shows that many more titles are misidentified for this period, which—we think—is attributable to the fact that the correct titles were not yet part of the set of alternative titles: we used Utrecht listings as a source for gathering a set of titles, mostly Dutch versions of foreign movie titles. Many of these do not yet exist in either the IMDb or Cinema Context. We chose Utrecht because it has the smallest number of cinemas in our sample of four cities.

3.3.3 Experimenting with an extended set of film titles

We anticipated that by adding these identified titles, subsequent parsings of other cities should improve. To test this, we re-parsed Rotterdam in 1949, including the added titles identified for the Utrecht set.

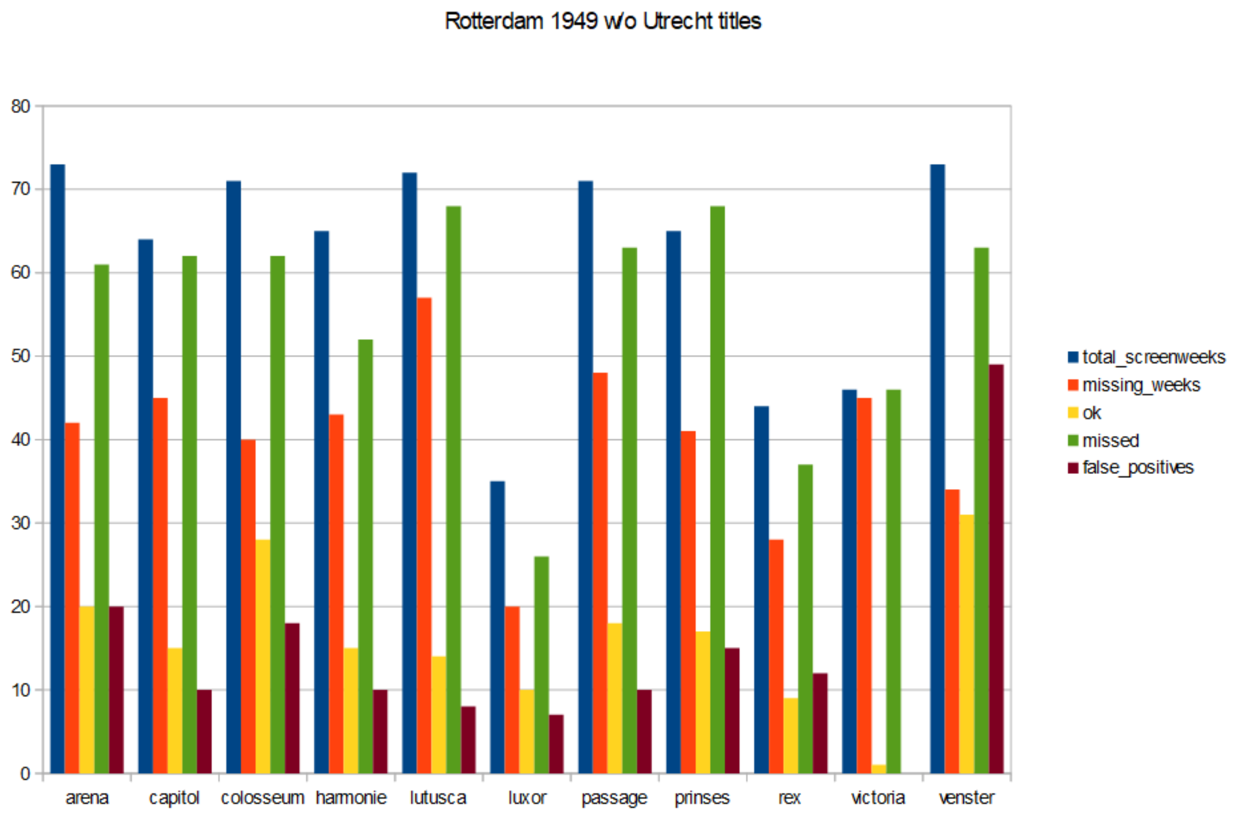


Figure 7: Film identification correctness and missing data overview for Rotterdam in 1949 per cinema

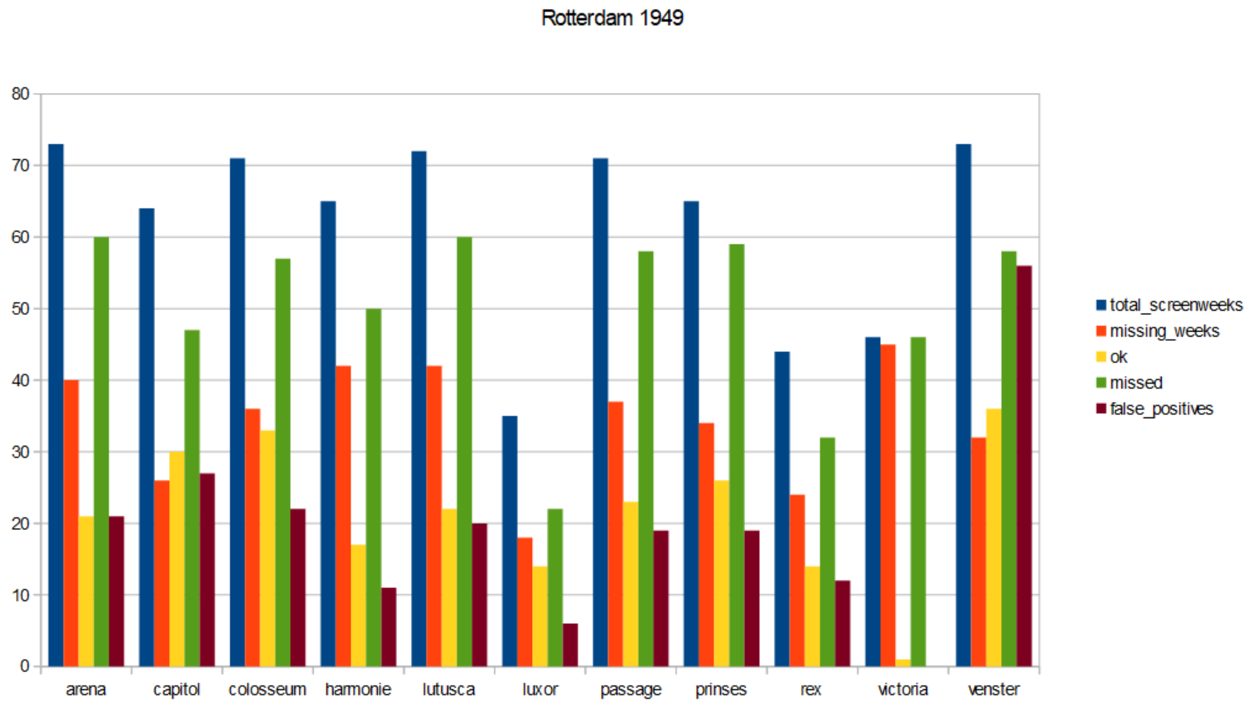


Figure 8: Film identification correctness and missing data overview for Rotterdam in 1949 per cinema using the extended set of film titles

This step resulted in more true positives (and fewer false positives) which are reported in figure 8. When comparing these results to the previous ones (obtained before adding the titles identified in the Utrecht files (see figure 7)) we observe this approach manages to correct 24% of the wrongly identified titles (a reduction of 11% in misidentified titles and an increase of 25% in falsely identified titles).[13]

4. Using the DIGIFIL data set

4.1 Usability of “dirty” data

To estimate the effect of the presence of noise in the data on the usage, we compared the relative frequencies of screening production countries between the raw and clean data sets for Rotterdam 1949 (countries with percentages under 1% omitted):

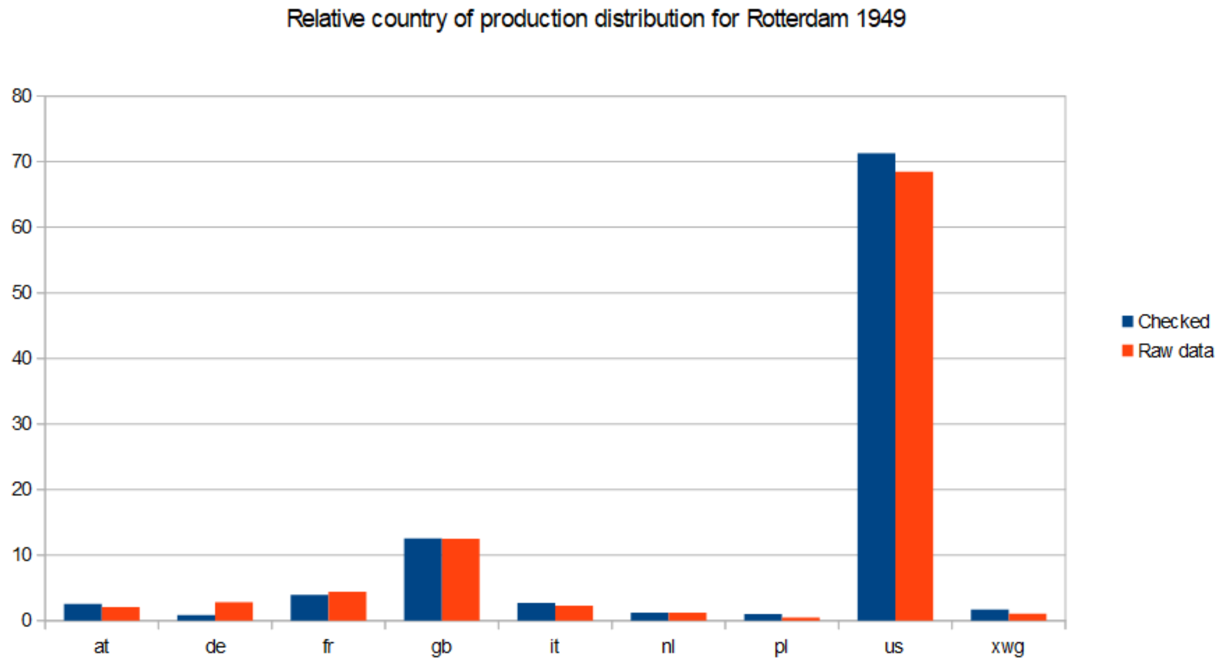


Figure 9: Comparison of countries of productions between DIGIFIL data and a golden standard.

[13] The latter is likely a result of titles previously matched to Cinema Context IDs and being left out of the graph now being matched to IMDb IDs and included: whereas without using the Utrecht titles for identification 337 films are shown here, including the Utrecht titles identifies 450 listings. Note: this last test was performed without properly normalising the cinema names.

As can be observed in Figure 9, The DIGIFIL data produces results which are very close to the golden standard—with only one notable exception: the raw data suggests that there were more German films screened than actually was the case. For the other countries, the frequencies are nearly identical. Germany is an anomaly here because unidentified Dutch titles are often close—in terms of letters and partial words—to German titles.

4.2 Preliminary results

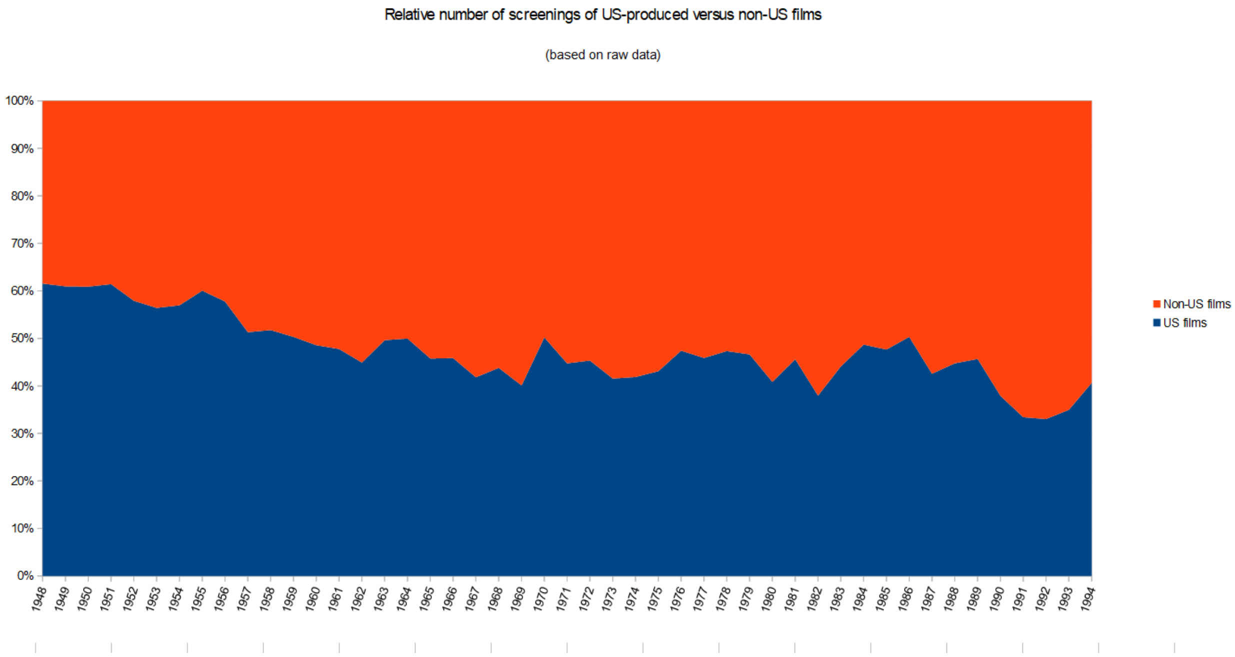


Figure 10: based on 566,623 total screenings. Unidentified titles and films for which we do not have country data are omitted.

Figure 10 shows the proportion of screenings of US-produced versus non-US produced films, using the data from all four cities, over the period 1948 through 1994. Given the sanity checks above (Figure 9) we tend to have a higher level confidence in the data we produced. The figure indicates a slow gradual decline of the market share of US film, in terms of screenings, which may conflict with the findings by Bart Hofstede (2000, pp. 106-107), who observed a different dynamic: a relative decline of US market share during the 1960s and 1970s followed by a sharp increase during the 1980s to a share of over 80%. Hofstede is not clear about the sources used for his statistics, but he claims to have used box office statistics, whereas the DIGIFIL data counts the number of screenings. However, we have to remain careful about this finding: our sanity checks were limited to the 50s, so the divergence could be driven by other, unobserved factors.

To demonstrate other opportunities for research we included additional figures below.

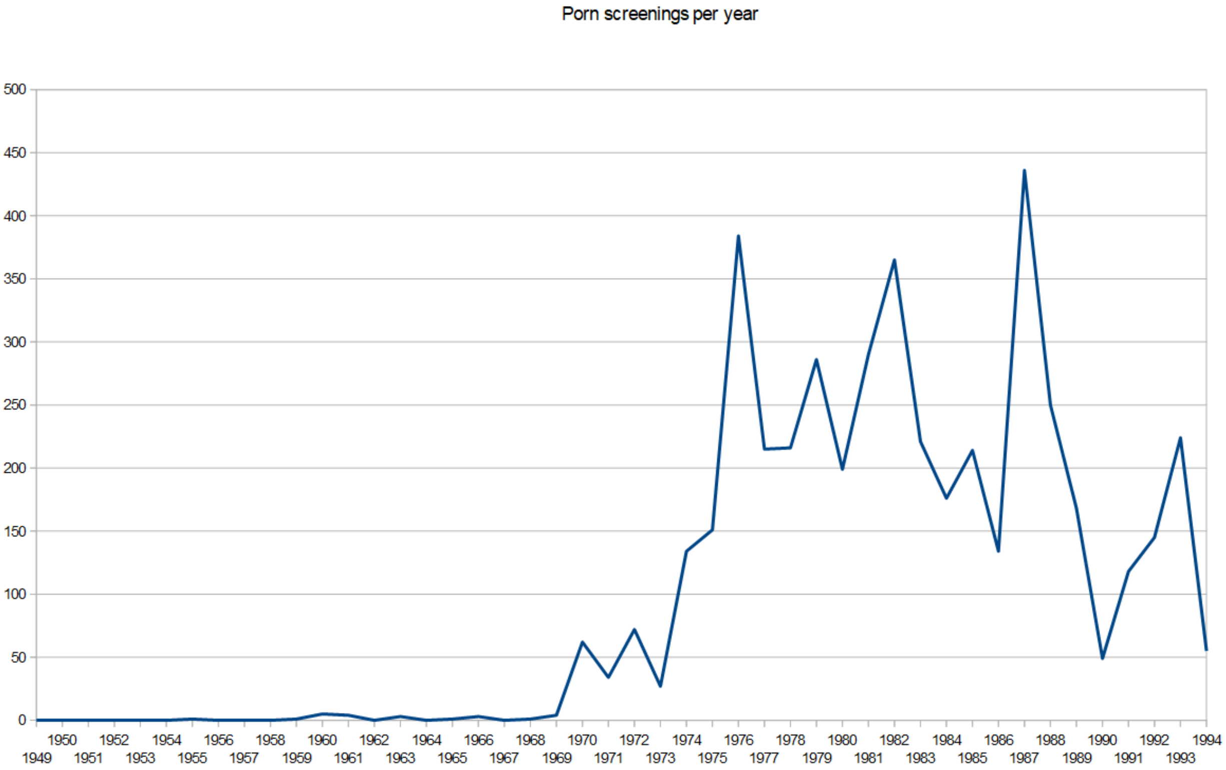


Figure 11: sum of adult titles screened in the four cities

Figure 11 shows the evolution of adult film screenings. Not surprisingly, the increase starts only at the end of the sixties, and keeps climbing to the mid-seventies. The figure then suggests a long-term decline starting from the eighties onwards—which, again, confirms our expectations, given that the competition of other media, such as VHS, gradually emptied film houses that screened adult materials. The peak in the mid-eighties was surprising, but upon closer inspection, attributable to an linking-error: it is solely explained by “The Morning After”, a perfectly innocent crime movie from 1986, which our linker confused with an older adult movie that shared the same title. This results shows that even though DIGIFIL does allow researchers to study the movie landscape, given the current quality of the data, findings have to be handled with care (and double-checked if possible).

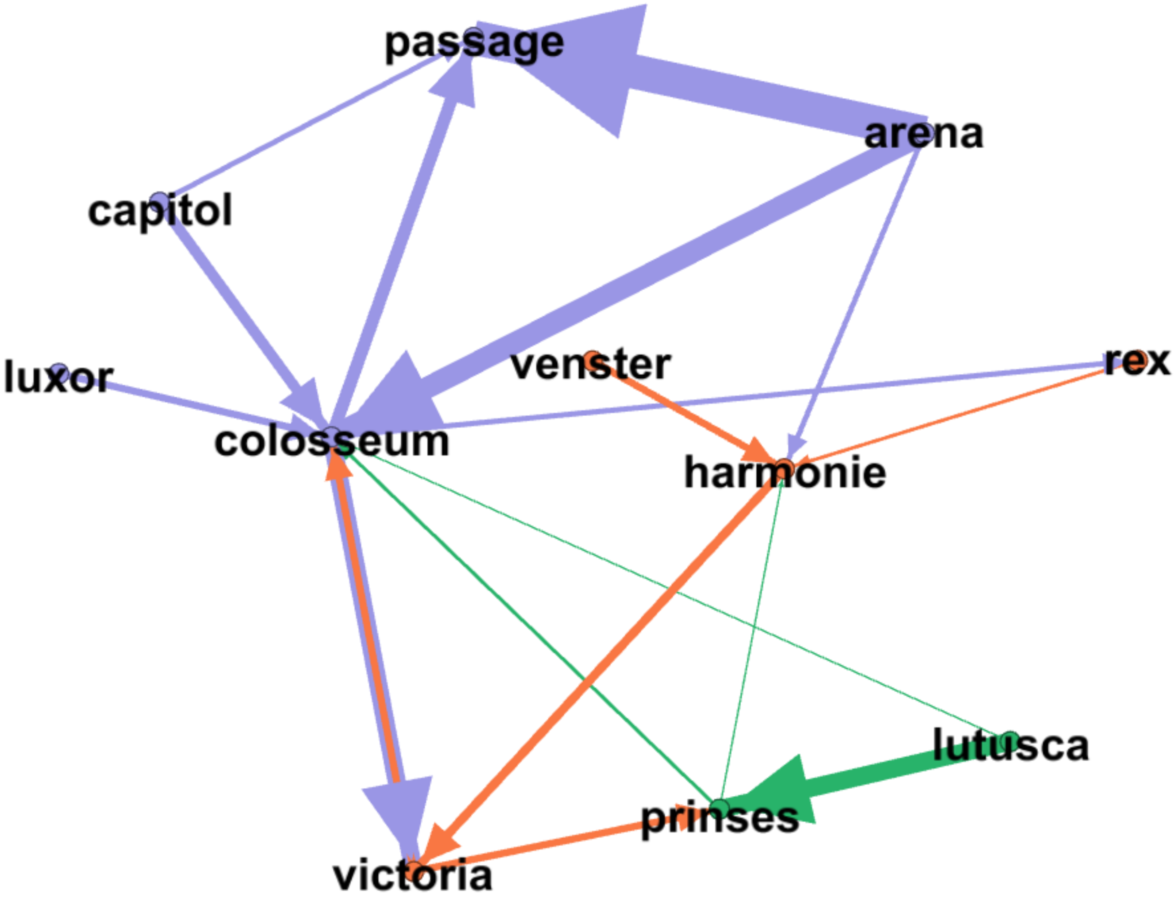


Figure 12: Rotterdam ‘49 film circulation

Figure 12 shows the circulation of films through the Rotterdam cinema landscape over the year 1951. An arrow from cinema A to cinema B indicates that a certain title was screened at a certain date in A and was screened at a later date in B. Some cinemas functioned as premiere cinemas, screening films at the earliest date (their first run), such as Arena or Lutusca. Some cinemas ranked lower in the hierarchy, screening films for the second or later runs, such as Victoria or Colosseum. Cinemas apparently at the bottom of the food chain would mostly have arrows pointing towards it, not away, screening the ‘oldest’ movies, such as Prinses or Harmonie. The graph also shows to what extent cinemas were embedded in the network (thick lines indicating a high number of shared film titles programmed). An art house cinema such as Venster would not hardly participate in the network because it screened a type of niche films with a limited appeal for the ‘regular’ cinemas.

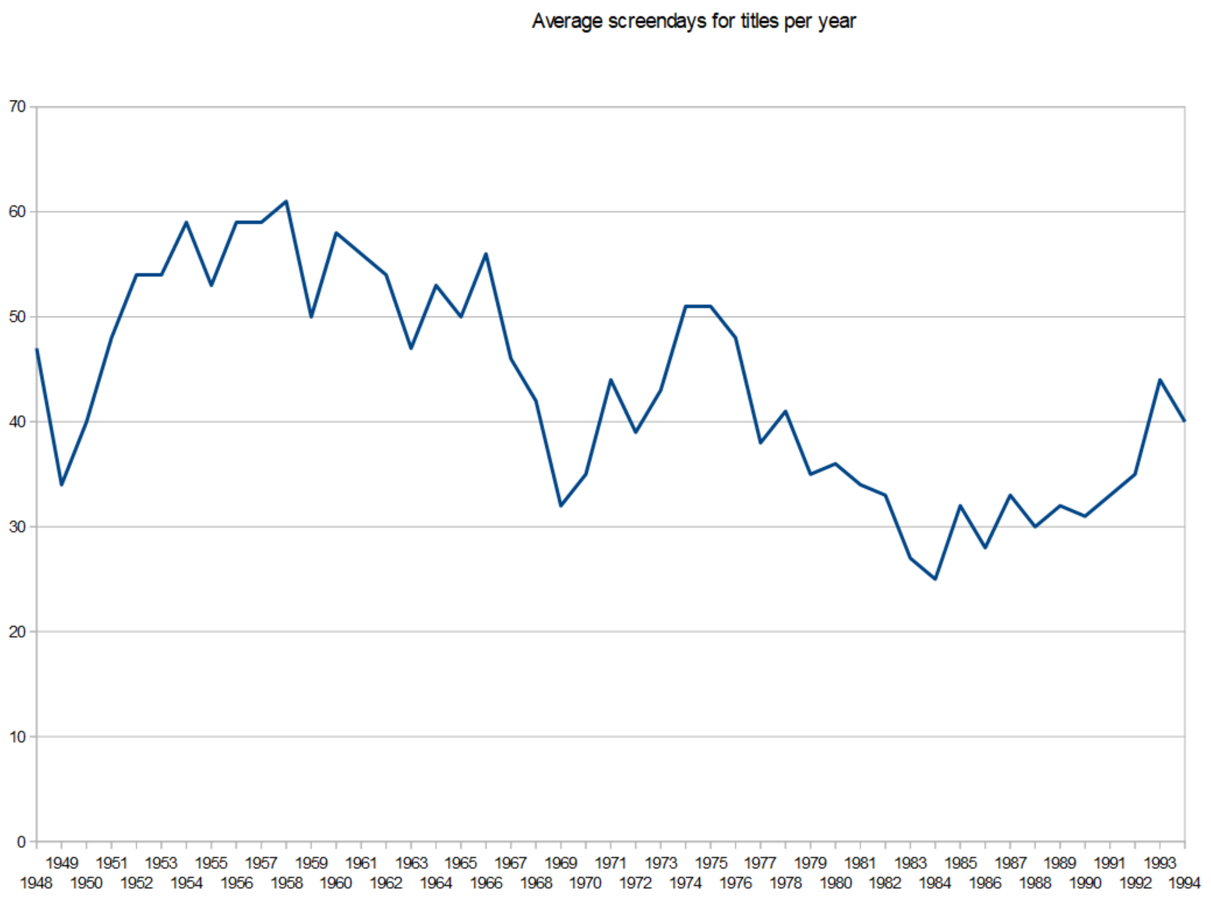


Figure 13: Average screenings for all four cities

Figure 13 is another attempt to extract long-term trends from the dataset. After calculating the average duration a movie was screened (the yearly average screen days per title) we observed a general decline: it took less-and-less time for movies to be taken out-of-circulation, suggesting a consistent shift in consumption patterns over the postwar period.

4.3 DIGIFIL and Cinema Context: Converging Data

4.3.1 Introduction

DIGIFIL has resulted in a dataset that could be a very useful addition to the Cinema Context database, covering at least the four major cities Amsterdam, Rotterdam, The Hague and Utrecht in the period 1948-1995. It would allow investigations of the Dutch post-war cinema exhibition and distribution landscape with a depth and detail that is unprecedented in film historiography. Moreover, the procedure as it has been developed, could be applied to other periods, geographical areas and topics as well, allowing even more additions in the future.

4.3.2 Merging datasets

Several datasets that have been produced in the DIGIFIL project, could be added to Cinema Context at this point. One dataset to add is the list of additional IMDb IDs for film titles that now only have Cinema Context IDs, for want of matches in previous sessions of data entry. A second addition would be the list of (mostly Dutch) title variations for known films that are not yet included in Cinema Context. Also, the lists of film titles distributed in the Netherlands in the 1960s-1980s would be a valuable contribution to the collection of film titles in Cinema Context, that presently does not run beyond 1960. A further possible dataset would be the OCR variation for cinema names as found in Delpher. Finally, another useful result from the DIGIFIL project is a list of suggested corrections of the active periods for cinemas: while the list of cinemas originally included in Cinema Context originates in membership listings distributed and managed by the Dutch association for film exhibitors, distributors and producers Nederlandse Bioscoopbond (NBB), for DIGIFIL we were able to identify additional active cinemas, not included in these listings.

4.3.3 Cleaning data

In the current state, however, the programming data are not sufficiently clean: the data set contains faulty matches, most significantly caused by either misidentifying a film (a common occurrence, especially with similar or identical titles), or misallocating a film program to the wrong cinema (rare, but it does occur). For some periods (or cities) we have more complete data than for others, which mostly hinges on the number of digitized newspapers that are available for a certain period and/or the quality of the OCR. For instance for the late 1940s we have a much sparser collection of film listings than for later periods.

There are several scenarios for how to proceed.

1. Use the data in the current, unrefined form: Even if there are still many mistakes in the data, results derived from them might allow us to generate some larger, long-term trends.
2. Use the parts of the data for which we are confident about the quality. It is possible to extract a subset of the data that meets the quality criteria for inclusion in Cinema Context. For example, if the movie title is difficult to confuse with other titles, the cinema is unique to a certain city, the OCRed title matches the matched title closely and four different movie listings all list that same title for that same cinema on the same date, we can be quite certain that the record is correct.
3. Refine the data up to a standard that meets the requirements of Cinema Context. Obviously, this would be preferable. The next questions then would be: what are those requirements and how can we meet them?

4.4. Encoding Uncertainty: Adjusting the Cinema Context Data Model

The Cinema Context data model was not explicitly designed to capture uncertainty. In case of conflicting sources, it is up to the editors to decide which source is the most reliable. However, this may not always be possible; for example, if one major newspaper lists a cinema playing “Een Flinke Jongen” while another lists, for the same cinema and the same date, a screening for “Een Linke Jongen”, it may not be possible to judge. For instance, such a model that scores probability could show two conflicting opening dates of a certain cinema, each with a probability score of 50%.

In Cinema Context, only one opening date for a cinema can be registered; perhaps an alternative opening date is mentioned in an information field, but this would be hard to use in a calculation, it remains anecdotal. Especially for the earliest cinema history, a lot of uncertainty exists in the sources and Cinema Context users must be aware of the fact that information on early (traveling) cinema exhibition, to name an example, is incomplete at best. We would suggest introducing a start and an end date in order to allow for ranges of dates (for instance indicating that a cinema was opened or closed at least before or after a certain date).

The way uncertainty (and this also applies to the identification of film titles) is dealt with in the data model is relevant for the DIGIFIL results. Because any decision on whether and how to accept DIGIFIL data for incorporation into the Cinema Context dataset would entail considering how much uncertainty one is willing to accept. Possible options include:

A. Incorporate data that is not corrected/checked by a human eye, but flag it as such, e.g. by assigning a certain user (perhaps with the option to remove/change the flag when this specific record has been checked)
B. Incorporate data that has been checked by a human eye (and here also a flag function could be used)
C. Create a clone of Cinema Context specifically for DIGIFIL data. Users then can integrate both data sets at will. This option could be applied to either unrefined data (as under A) or cleaned data (as under B or C).

Mistakes will likely remain unless relevant advertisements for the listings are consistently looked up and can actually be found. For example, we found a listing for a film called ‘Happy Landing’, which upon checking the advertisement turned out not to be any of the three ‘Happy Landing’ films IMDb listed as possible candidates for that year, but instead is an alternative English title (not listed in IMDb) for the movie ‘Jumping Jacks’. Similarly, there are two films called “De Trommen van Fu Manchu”, released a few years apart and in some years both these films are screened under the same title in movie theatres in the Netherlands.

Option A would import a lot of mistakes in the Cinema Context database. The question is whether such a problematic dataset would still invite users to do research? Option C obviously would be preferable but is time and cost intensive. The whole purpose of the DIGIFIL project was to find ways to automate the data extraction process and prevent the time-consuming task of verifying each individual source. A version of option B would perhaps be most feasible. Next questions are then: what needs to be done and how can we realize this?

4.5 Unresolved Issues: What needs to be done?

4.5.1 Supervision and cleaning of the data

The preliminary results that DIGIFIL has yielded so far are annual lists of cinema programs for each of the four largest cities for the period 1948-1995, almost 200 lists in total. Basically, each list contains at least a date, a cinema name and a film title, plus various additional columns, for instance, references to the newspaper sources (including a URL), and some information on the matched film title, such as year of production and an (external) ID of the film. For each list, several things need to be checked:

- Is the cinema match correct? Are the correct cinemas selected for the city under consideration? This should be about 5 minutes of work per file, even if it is 8000 rows long.
- Is the film title match correct? The match that links the title taken from the OCRed source to a reference catalogue of film titles (consisting mostly IMDb and Cinema Context). When in doubt, the original source can be referenced via the link to the digitized newspaper in Delpher. When working cumulatively from the oldest to the most recent years, the corrected title matches will be available in the updated list for the next year. So all titles from the corrected year that are still circulating in the following year, will already have been matched correctly. When a certain title has been matched for the previous or following week, the system automatically allocates it a code that signifies that the match probably is correct.
- After checking all records, some problematic titles will remain. Mostly, these are titles that refer to different film versions, for example, adaptations of famous plays, or remakes of popular stories (e.g. various Hamlet films; Tarzan films), or films with many sequels or episodes (e.g. Policy Academy Part 1 to 7; James Bond). These cases would either need more specific checking of the sources, or would need generic labels (e.g.: “A film from the Police Academy series”)
- Once all records have been checked, we can map where the gaps in the data are. We could try to fill those gaps by going back to the raw data.

4.5.2 Issues related to timestamps

Another unresolved issue concerns the temporal unit of the screening data. For each of the screenings, we need to pin exactly the day and the hour that the film was screened. As the data is still noisy, many of these timestamps are misrecognized: for example, “om 7, 9 11 uur” often is OCR-ed as something like “8. ; 1911 vr”. However, as cinemas adhere to rather stable schedules (alterations being explicitly marked by, for example, notes as ‘Attention! Diverging screening times!’ in certain listings), we expect to be able to solve this problem by recoding the odd timestamps.[14]

As an example, we show the timestamp strings for cinema Arena (Rotterdam) for 1958 (leaving out those occurring less than three times):

Number of occurrences	String
56	2730
32	27930
14	27930Zazo24307930
6	2en730
5	27en930
5	zo2
4	vrdag8
3	fr""nt
3	27930Wo24307930

Obviously, most of the screenings actually took place at 14:00 and either 19:30 or at 19:00 and 21:30. Reading the fourth row suggests that those would be the screening times during the week while during the weekend screenings would take place at 14:00, 16:30, 19:00 and 21:30. There are a few exceptions, markedly a few occurrences where there would be additional screenings on Wednesdays.

Looking at a more recent year, we can observe a few changes. Below is the overview for cinema Alhambra (Rotterdam) in 1988 (again leaving out any strings occurring just once):

Number of occurrences	String
72	
32	za2415
26	za0015

[14] An exception being singular extra screenings, but it should be possible to recognize these because of an extra title being listed in a certain week for a particular cinema.

14	15
12	15
10	200645930
10	za15
8	1418452130
8	1420
8	133020
7	130en
5	200830
4	14192130
4	345545745945
4	142030
4	300515730945
3	zanacht0015
3	200800
3	130800
3	za0015HaiM
3	18452130
3	19302230245500700000

The vast majority of listings does not even have an associated time, and the most frequent listed times seem to point to special night screenings.

4.5.3 Integration of reviews

Besides collecting listings, DIGIFIL attempted to include contextual data, primarily movie reviews. To this end, we built a specialized classifier that managed to pick out reviews from the article database with high precision (more than 90% on our test set). We further explored this collection by means of topic modeling and sentiment classification. We attempted to link reviews to IMDb identifiers, but, unfortunately, we ran into so many issues that we had to abandon this aim. In other words: while we have extracted a, we believe, substantive series of movie reviews the proper integration of these data still has to happen.

5. The way forward

Based on some pilot work on the lists, we found that it might take at least one hour per list to correct. We could make some tools to expedite this, some of which are already in place, for example: (i) a “common title lookup tool” where a user can enter a few IMDb links to cast members who collaborated in a specific film project; (ii) a tool to retrieve Delpher search results (for example for a particular film title) in a particular year without having to load all the slow Delpher thumbnails or be limited to ten results per page.

However, this remains a very rough estimate; it is likely that the process goes much faster after a while, through habituation and because of the cumulative effect of the matched titles from the preceding years. This means that we would need about 200 hours of correction work, plus some extra hours for the remaining problematic titles. This could either be done internally by CREATE employees (Thunnis van Oort and Ivan Kisjes, for instance) or, probably the more feasible option, be outsourced to a student assistant (or a crowdsourcing platform), supervised by Ivan and Thunnis. If we would hire a student-assistant for 200 hours, this would cost around 6000 euros. To continue refining the DIGIFIL dataset, we intend to apply for further funding such as ‘klein data project’, KNAW/DANS.

Lastly, the tools and techniques developed within DIGIFIL, have proven useful for extracting other types of implicitly structured information, such as theatre listings, stock market information, but also the location of ships sailing under a Dutch flag. DIGIFIL has paved the way for larger attempts to collect “micro-events” from big data, opening up new avenues of historical research.

6. List of outputs

Publications and presentations

Beelen, K., ‘CLARIAH-project Digital Film Listings. Reconstructie van oude filmladders’, *E-data & research* 14.1 (2019), https://www.edata.nl/1401/pdf/Reconstructie_van_oude_filmladders.pdf.

Beelen, K., I. Kisjes, K. Lotze, T. van Oort, Automatic Extraction of Film Programming Data from Digitized Newspapers Reflections on the DIGIFIL project. Paper presented at HOMER Conference ‘Anchoring New Cinema History’, Nassau, 2019.

Beelen, K., K. Lotze, I. Kisjes en T. van Oort, ‘Digifil. Digital film listings’, in: E. Renckens ed., CLARIAH - A Digital Research Infrastructure for Humanities Researchers in The Netherlands (CLARIAH, 2019), 70-71, https://clariah.nl/boekje/clariah-digitaal.pdf.

Beelen, K., T. van Oort, I. Kisjes, K. Lotze, J. Noordegraaf, Digital Film Listings (DIGIFIL). Paper presented at CREATE Salon. From text to table: extracting information from semi-structured data, Amsterdam, 2019.

Kisjes, I., T. van Oort, K. Lotze, I. Staliunaite, J. Veerbeek, K. Beelen, J. Noordegraaf, *Digital Film Listings (DIGIFIL). Final report* (Clariah/CREATE, 2020)

Kisjes, I., T. van Oort, K. Lotze, I. Staliunaite, J. Veerbeeck, K. Beelen, K. Noordegraaf, Mining the Movie Landscape: Extracting Film Listings from Digital Newspapers. Paper presented at Digital Humanities Benelux, Liege, 2019. Shortlisted for best paper award.

Data, code & tools

The code used in the project, referenced throughout this report, can be found on gitlab: https://gitlab.com/uvacreate/digifil

The data that was produced in the project is still being processed. Once portions of the data are cleaned, they will be added to www.cinemacontext.nl (see also regularly updated dumps at DANS (10.17026/dans-z9y-c5g6). (Portions of the) raw data can be made available upon request.

7. References

Biltereyst, D., R. Maltby & P. Meers, Cinema, Audiences and Modernity. An Introduction, Eds. D. Biltereyst, R. Maltby & P. Meers, *Cinema, Audiences and Modernity. New Perspectives on European Cinema History*. Routledge, London & New York, 2012: 1-16.

Biltereyst, D., R. Maltby & P. Meers, *The Routledge Companion to New Cinema History. Routledge*, London & New York, 2018

Dibbets, K. (2010) 'Cinema Context and the genes of film history', *New Review of Film and Television Studies*, 8: 3, 331 — 342, DOI: 10.1080/17400309.2010.499784

Garncarz, J.. *Wechselnde Vorlieben: Über die Filmpräferenzen der Europäer, 1896-1939*. Frankfurt am Main, Basel: Stroemfeld, 2015

Hofstede, B.P., *In het wereldfilmstelsel. Identiteit en organisatie van de Nederlandse film sedert 1945*. Eburon, Delft, 2000.

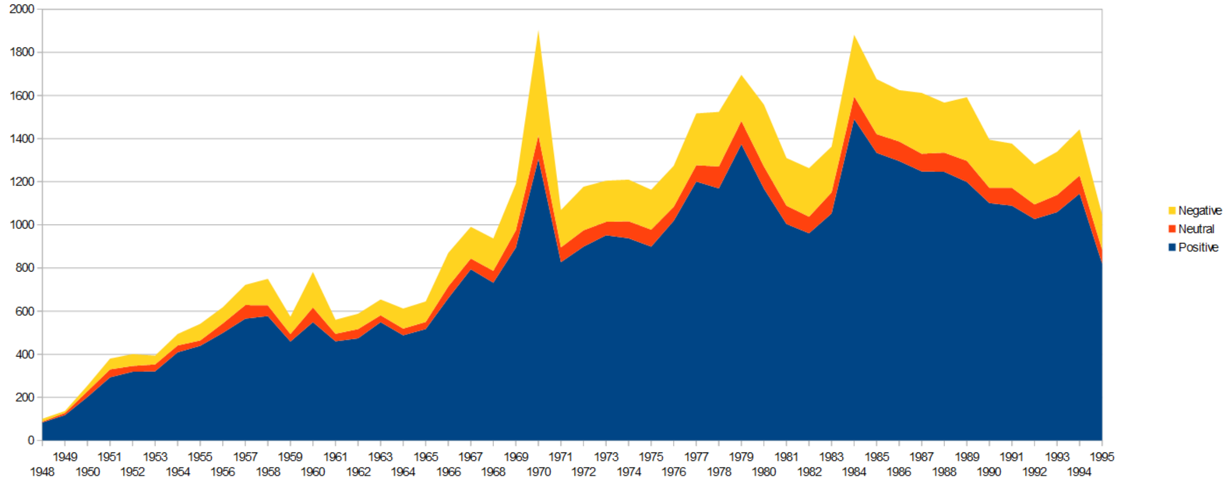
Meers, P., D. Biltereyst & L. Van de Vijver, Metropolitan vs rural cinemagoing in Flanders 1925-75, *Screen* 51 (3) 2010: 272-280.

Noordegraaf, J., K. Lotze & J. Boter, Writing Cinema Histories with Digital Databases: The Case of Cinema Context, *Tijdschrift voor mediageschiedenis* 21(1) 2018: 106-126.

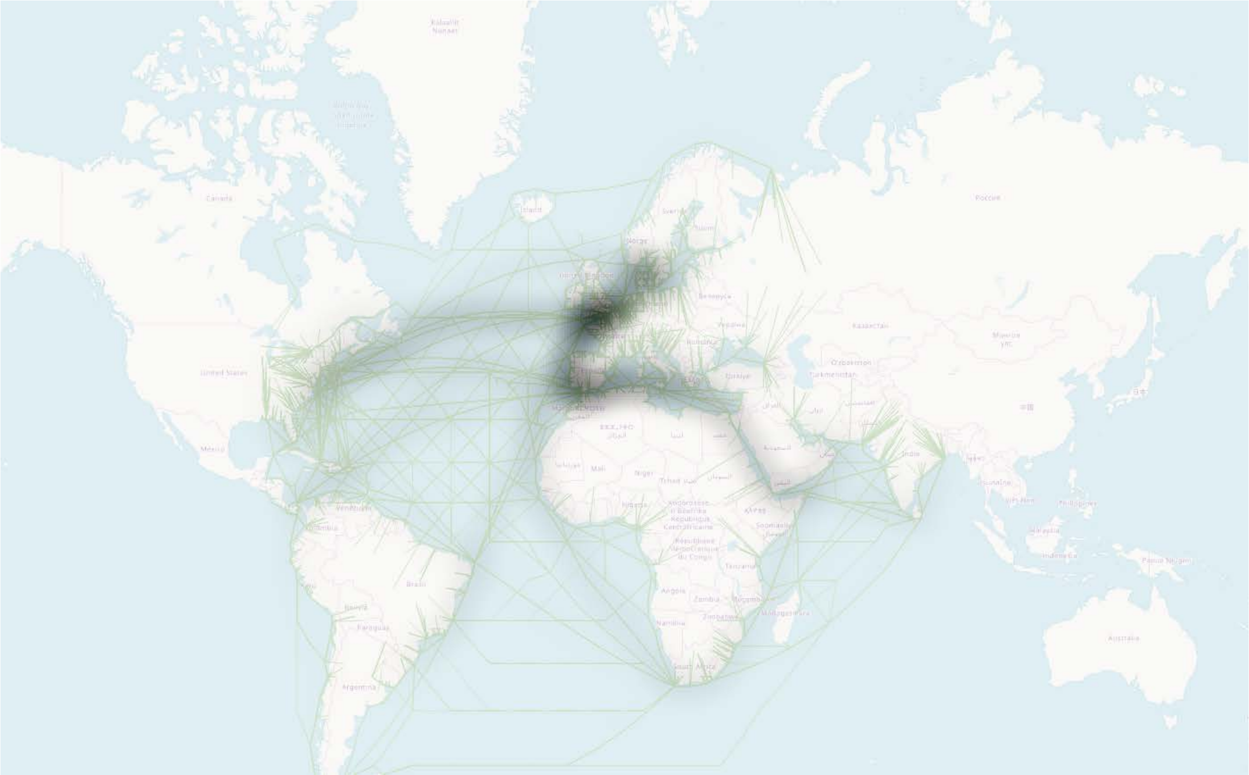
Sedgwick, J., C. Pafort-Overduin & J. Boter, Explanations for the Restrained Development of the Dutch Cinema Market in the 1930s, *Enterprise and Society* 13(3) 2012: 634-671.

Addendum: other visualizations

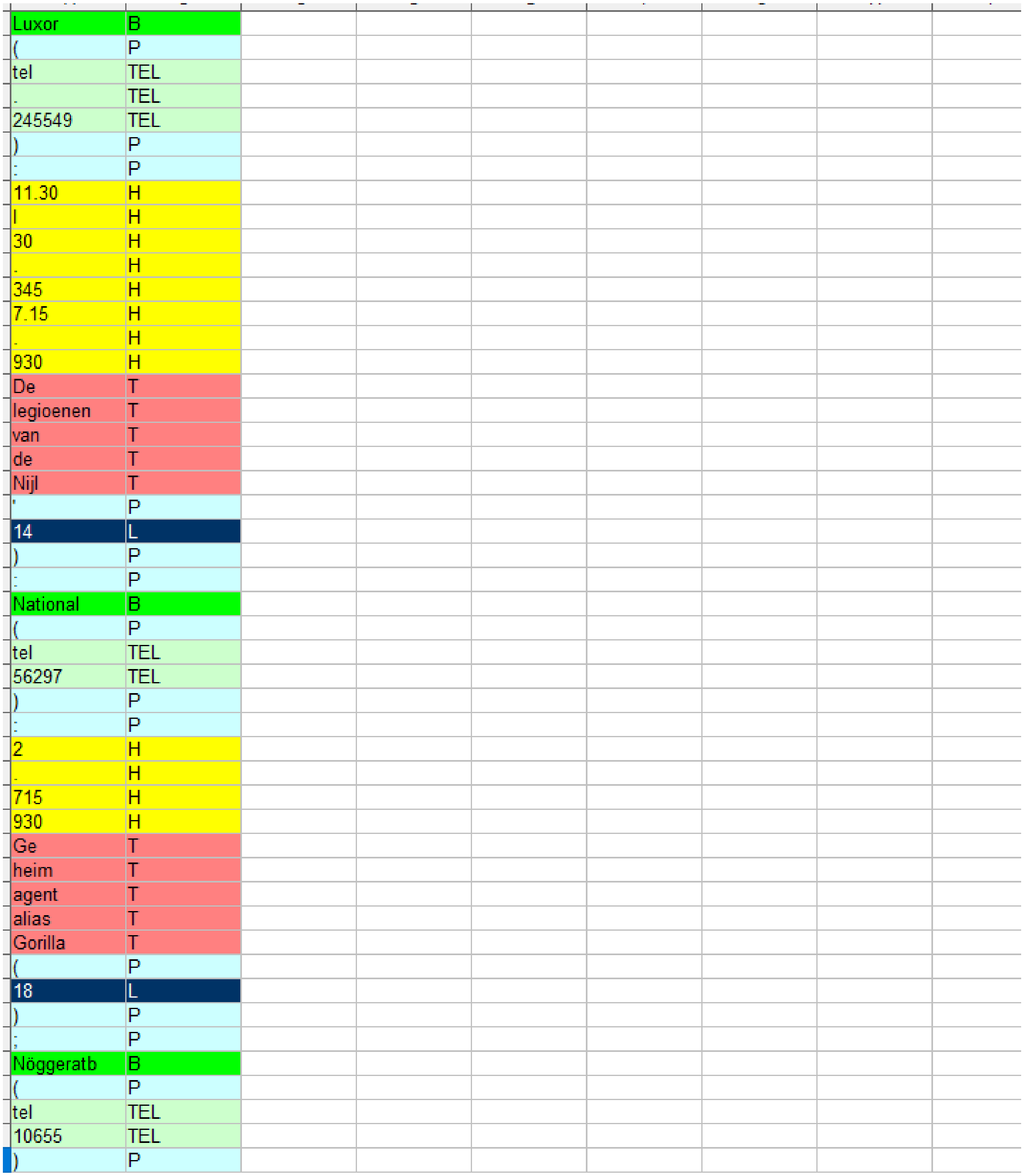
Result of emotion classification of movie review over time:



Ship movements in 1962: information extracted newspaper using tools developed within DIGIFIL



Example of manual correction of tagged film listings



Luxor	B
(	P
tel	TEL
.	TEL
245549	TEL
)	P
:	P
11.30	H
l	H
30	H
.	H
345	H
7.15	H
.	H
930	H
De	T
legioenen	T
van	T
de	T
Nijl	T
'	P
14	L
)	P
:	P
National	B
(	P
tel	TEL
56297	TEL
)	P
:	P
2	H
.	H
715	H
930	H
Ge	T
heim	T
agent	T
alias	T
Gorilla	T
(	P
18	L
)	P
;	P
Nöggeratb	B
(	P
tel	TEL
10655	TEL
)	P