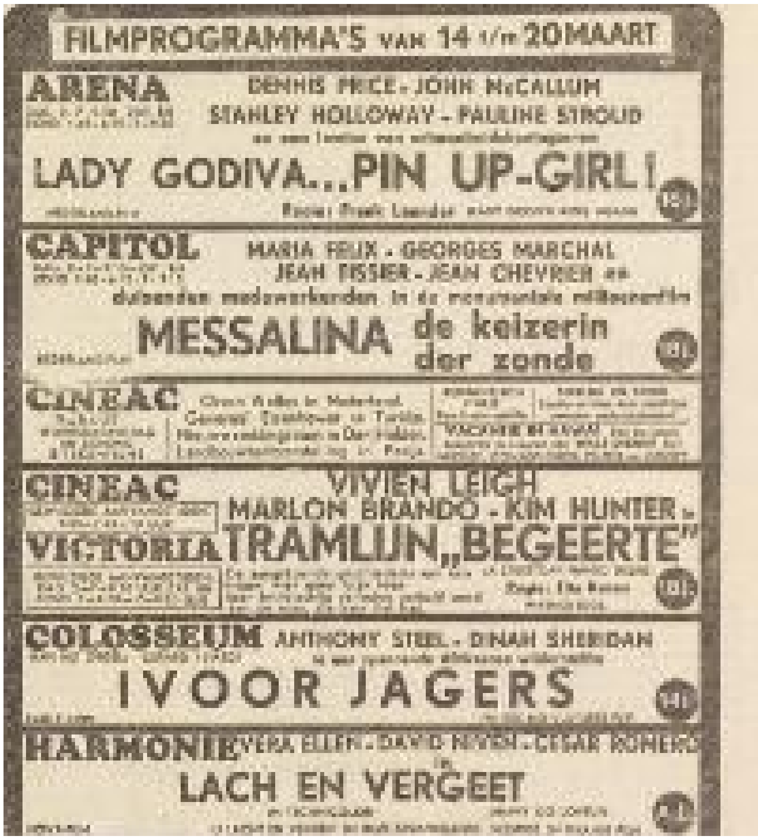
Mining the Movie Landscape

Extracting Film Listings from Digital Newspapers

This presentation reports on the "Digital Film Listings" (DIGIFIL) project, which aimed at automatically extracting, digitizing and publishing film screening data from the weekly *Filmladders* (or films listings) as published in historical newspapers. 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 (Biltereyst et al., 2018).



FILMPROGRAMMA'S VAN 14 t/m 20 MAART

ARENA
DENNIS PRICE - JOHN McCALLUM
STANLEY HOLLOWAY - PAULINE STROUD
LADY GODIVA...PIN UP-GIRL!

CAPITOL
MARIA FELIX - GEORGES MARCHAL
JEAN TISSIER - JEAN CHEVRIER en
duizenden medewerkenden in de monumentale millioenenfilm
MESSALINA de keizerin der zonde

CINEAC

CINEAC
VICTORIA
VIVIEN LEIGH
MARLON BRANDO - KIM HUNTER in
TRAMLIJN „BEGEERTE"

COLOSSEUM
ANTHONY STEEL - DINAH SHERIDAN
IVOOR JAGERS

HARMONIE
VERA ELLEN - DAVID NIVEN - CESAR ROMERO
in
LACH EN VERGEET

Figure 1: Example of a film listing

Methodologically, DIGIFIL demonstrates how to convert implicitly semi-structured text—i.e. text that adheres to a specific pattern that computers can parse (figure 1)—automatically to entries in a database. We focus on the extraction of one type of "micro-events", namely film screenings, but demonstrate how the tools we develop can be applied to harvest other types of information (such as theatre plays and the movements of ships). In this sense, DIGIFIL is an attempt to merge "everyday history" (*Alltagsgeschichte*) and digital humanities. Put differently: we show how to write micro-history with big data.

The first part of the presentation elaborates on the workflow: the collection of data and their transformation from raw text to an enriched, semantically annotated format. The first stage comprised a classical “needle in a haystack” problem, in which we had to retrieve the listings (a very specific type of articles) from the gigantic pile stored in Delpher (a website that provides millions of digitized texts from newspapers, magazines and books in Dutch language).[1] After fishing the Filmladders from the digital archive, we set out to parse the raw text. As the structure of the listings varied considerably across newspapers (and over time) and the OCR of the text was not quite perfect, we had to rely on machine learning for annotating the *ladders* (i.e. label the category of each word in order to identify the listing’s structure). Similar to Part-of-Speech tagging, we trained a sequential model to categorize each word in the listing according to its function within the document (i.e. title, timestamp, cinema name etc.). Then, we identified the listed film titles, linking them to the Internet Movie Database (IMDb). Lastly, we re-parsed the ladders, using the identified titles to estimate reliable versus less reliable sets of movie screening data.

The DIGIFIL pipeline produces an enriched dataset on film screenings. To assess the overall usability of the data we pursued various sanity checks (which we present in the second part) that compare the automatically generated results to a ground truth based on manually collected data.

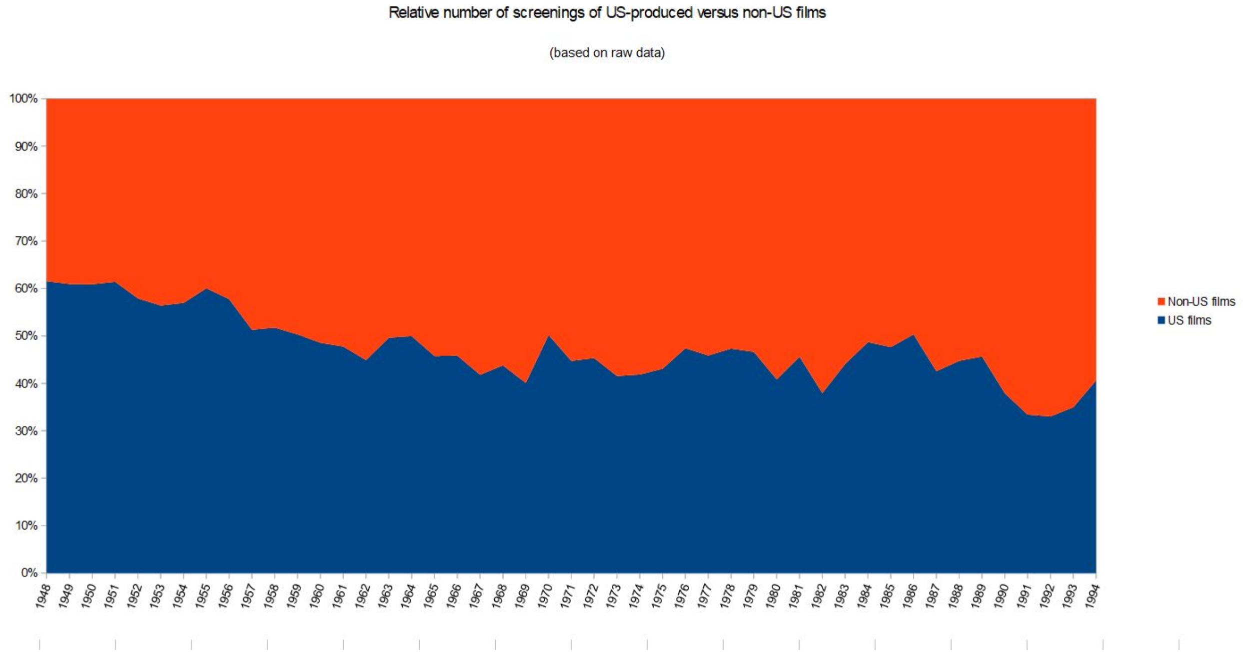


Figure 2: Percentage of US and non-US screenings, based on a total of 566,623 screenings. Unidentified titles and films for which we do not have country data are omitted.

[1] https://delpher.nl

The third part of the presentation discusses the opportunities for digital research created by DIGIFIL: what type of research do we enable? Figure 2, for example, shows the proportion of screenings of US-produced versus non-US produced films over the period 1948 through 1994. The figure indicates a gradual decline of the market share of US film, in terms of screenings, which, if taken at face-value, conflicts with the findings by (Hofstede, 2000), 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%. However, because the sources are rather noisy (given the presence of OCR errors) and each step in the digitization risks adding more errors (by, for example, faulty linkage of the titles or misrecognition of a ladder's structure), the results derived from the DIGIFIL data, should still be taken with a grain of salt, and interrogated further before drawing definite conclusions (foregrounding the topic of data-criticism in humanities research).[2]

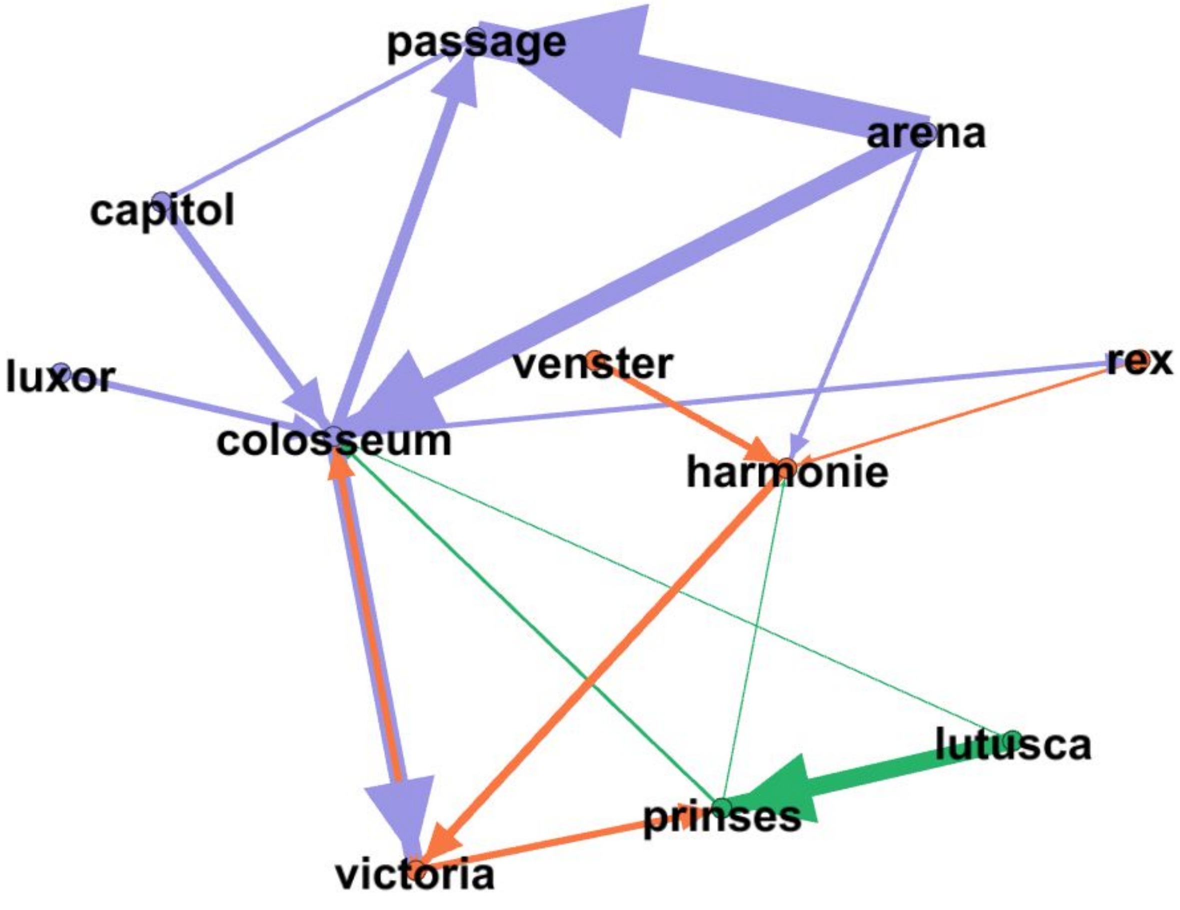


Figure 3: Rotterdam '49 film circulation: the nodes are cinema houses, the edges represent the circulation of films, their thickness indicating the size of the transfer

[2] Another reason for this apparent divergence could be the fact that Hofstede relies on box-office data while figure 2 is based on film screenings.

Another option is to study the circulation of films. Figure 3 tracks the movement 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. Other cinemas--ranked lower in the hierarchy--screened films for the second or later runs, such as Victoria or Colosseum. At the bottom of the food chain appear, i.e. those venues screening only the ‘oldest’ movies, are Prinses or Harmonie. The graph also allows us to measure to what extent cinemas were embedded in the local network (thick lines indicating a high number of shared film titles programmed, suggesting possible cooperations between exhibitors). An art-house cinema such as Venster would hardly participate in the network because it screened a type of niche films with a limited appeal for the ‘regular’ cinemas.

To summarise, DIGIFIL proposes a novel approach to media studies with respect to collecting and interrogating historical data. In previous projects, such a Cinema Context (Noordegraaf et al., 2018), compiling and digitizing records amounted to labour intensive process, requiring considerable time and resources. Besides facilitating data creation, the DIGIFIL dataset opens up innovative mixed-methods approaches in media studies, in which the “macro” can serve as a way to appreciate and contextualize the “micro”. Lastly, DIGIFIL is an (unfinished) exercise in data-criticism; it provides strategies for inspecting the quality of automatically generated data (and incorporating the uncertainty that flows from this process) and lays out perspectives for future improvement.

References

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

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

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