Christian Gosvig Olesen and Ivan Kisjes

From Text Mining to Visual Classification: Rethinking Computational New Cinema History with Jean Desmet's Digitised Business Archive[1]

Abstract

Focusing on the specific case of cinema owner and film distributor Jean Desmet's digitised business archive, this article discusses how computational approaches may facilitate the archive's unlocking for researchers in the Dutch national research infrastructure – CLARIAH – Media Suite. To this end, the article considers previous computational approaches to film-related sources in New Cinema History research in a historical perspective, suggesting a novel approach which combines text mining and visual classification. The article argues that such a combination is necessary to yield results which reflect the archive's material heterogeneity and complexity, and that it offers a new direction for computational approaches in New Cinema History and their conceptualisations of film-related materials as historical sources.

Keywords: New Cinema History; digital humanities; Jean Desmet; CLARIAH Media Suite; quantitative history

Introduction

Since the 1970s, film scholars have increasingly looked to a variety of film-related sources – rather than films primarily – to study cinema culture, cinema-going and film distribution. In the process, they have contributed to the emerging interdisciplinary research field of New Cinema History.[2] As cinema historian Deb Verhoeven highlighted in 2012, there are numerous film-related sources with great historical significance for New Cinema History:

> government reports, ordinances, building or police records, regulatory legislation, tax files, oral histories, marketing materials, industry archives, maps, box-office data, phone books, ticket stubs, newspaper advertisements just to name a small few.[3]

In the past few decades, research centring on such sources has increasingly involved databases, but also quantitative digital humanities methods that allowed for the detection of patterns in the distilled data. To unlock data, methods for automatic transcription have become

increasingly popular. Through the use of optical character recognition (OCR) and data mining methods, large corpora of trade journals, fan magazines and newspapers become machine-readable, and subsequently searchable for scholars. For instance, the Media Digital History Library's Lantern project (2011–), developed at the University of Wisconsin-Madison and, in the Netherlands, the Royal Library's Delpher project, used OCR and text mining to unlock collections and allow researchers to find and visualise patterns in them.

The principles behind such methods are not new. They find hermeneutic antecedents in, among others, the discipline of history, and the roots of such practices can be traced back to the emergence of the computational humanities some sixty years ago. They emanate from Cliometrics and *Annales*' serial history of the 1960s and 1970s, which introduced processes of labelling and extracting keywords from historical documents in order to process and analyse sources computationally, as a new type of historical inquiry. Such methods have proven extremely valuable to historians – both in the past and today – because they facilitate the searchability and combined quantitative analysis of large, heterogeneous and hitherto inaccessible paper collections. Anticipating contemporary discussions on big data, computerised methods were praised at the time for their ability to process large-scale datasets – for example, as historian Edward Shorter pointed out, for how they enabled scholars to handle a 'gigantic quantity of information'.[4]

Our recent project *MIMEHIST: Annotating EYE's Jean Desmet Collection* took inspiration from the aforementioned developments in New Cinema History, and sought to test the potential of computational methods for unlocking and analysing a significant collection of film-related documents. *MIMEHIST* was one of six pilot projects carried out in 2017–2018 within the framework of the CLARIAH Media Suite (the media studies work package of the Dutch national research infrastructure).[5] One of *MIMEHIST*'s key objectives was to unlock the EYE Filmmuseum's Jean Desmet business archive – which we discuss in more detail below – to media historians, facilitating exploratory browsing, pattern detection and network analysis. In terms of the different document types it holds, the archive's approximately 127,000 items are materially extremely heterogeneous and reflect the great diversity of sources with which cinema historians work. It contains personal and professional correspondence, postcards and telegrams, leasing contracts, intertitles, photos, bills and brochures, among many other types of documents.

On a basic level, our key research question was how, and to what extent, various digital humanities tools can be used to transcribe data from the Desmet Collection's heterogeneous paper archive in order to improve its searchability and exploration in future research. However, in seeking an answer to this question, we encountered problems that made us question the epistemology of current computational methods in light of what they enable us to study in historical paper sources. In particular, we realised that current methods do not do justice to the scope, complexity and material heterogeneity of Desmet's archive. Because current approaches are primarily text-centred, they neglect the significant information contained in paper documents' graphic and visual features – for instance, handwriting, letter layouts or logos. This prompted us to try other kinds of (visual) analysis, and we considered whether it might be useful for scholars to complement text mining approaches with computational approaches to visual classification, thus avoiding the pitfall of certain reductionisms.

Our efforts have historical precedents. In the past, tool-critical debates in quantitative research on paper sources have highlighted reductionisms and limitations. In the 1960s and 1970s, the computerised, punched-card-based approaches of Cliometrics and serial history were met with the critique that they produced evidence by extracting and performing statistics on a limited set of keywords and named entities. Among others, philosopher and historian Paul Ricoeur contended in the late 1970s that 'contemporary historiography, with its data banks, its use of computers and information theory, its constituting of series (using the model of serial history)', failed to reflect the complexity and contingencies of historical documents and consequently limited critical analysis.[6] Moreover, historians today express a strong desire for continued tool criticism as a way to understand the limitations of computational approaches and to avoid privileging certain types of analyses and assumptions over others. As digital historian Andreas Fickers contended in 2013, in the digital age, 'doing history is the "art of not being too sure"', and the tools we work with should reflect this.[7]

Thinking from the specific case of Jean Desmet's business archive, we felt an urgency to try to open new directions for computational New Cinema History by combining textual with visual analysis software as a means for unlocking data from historical film-related sources. Taking our cue from this research experience, our article argues for a combination of text mining and visual classification approaches, while at the same time reconsidering the historical foundations of current approaches in order to reorient them.

Our argument here is organised into two parts. In the first part, we discuss computational New Cinema History's methodological antecedents in serial history in order to elucidate how its text-centred focus forms the basis for current approaches. In the second part, we discuss in more detail the *MIMEHIST* project's work on the Desmet paper archive and the results it has yielded so far, so as to then finally reflect on the research avenues this may open for future New Cinema History research.

New Cinema History and Text Mining: Hermeneutic Antecedents and Current Approaches

Computational analysis of paper sources as a type of historical inquiry involves a great variety of procedures, which have developed in recent decades and moved from the discipline of history to that of media studies, specifically New Cinema History. To prepare and process film-related source materials for analysis with text mining methods, new cinema historians transcribe and combine data from paper sources in order to subsequently be able to analyse them quantitatively. Traditionally, such approaches seek to identify important keywords and concepts in documents in order to build databases and perform quantitative analytical tasks on the transcribed materials. In performing such analytical work, new cinema historians lean heavily on the methodological developments in computerised, quantitative historical research of the 1960s and 1970s. In particular, as cinema historian Richard Maltby has highlighted, New Cinema History's application of quantitative methods is indebted to the methodological foundations of *Annales*' serial approach to socio-economic, cultural history.[8] In order to elucidate New Cinema History's

current textual focus in the computational analysis of film-related sources, it is helpful to consider the research field's epistemological underpinnings in computational methods at the time.

Throughout the 1960s and 1970s, computational punched-card methods increasingly became the basis for quantitative historical approaches. Although access to computers was limited to a small number of individuals, they gained a strong presence in university research environments, especially in the US and France.[9] As such, they profoundly impacted discussions about historical methodology, and as mentioned above, they were initially embraced, much like in today's discussions on big data, for their ability to process large-scale datasets. The emergence of computerised methods in historiography is exemplified by the research field of Cliometrics (a denominator which combines the name of the muse of history in Greek mythology, Clio, with 'metrics').[10] The work of economic historians Robert William Fogel and Stanley Engerman are paramount in this regard, in particular their key work *Time on the Cross: The Economics of American Slavery* (1974), which resulted from years of experimentation with computerised data processing.[11] As a basis for computational analysis, historians would transcribe data from historical sources and order them into taxonomies of variables (for instance, people's political affiliations) represented by different numbers and grouped into different fields, which would be explained in accompanying codebooks and entered onto punched cards.[12] In practice, such methods would entail manually transcribing words and data from potentially thousands of different sources held in (different) institutional archives, in order to then process the data with computers (in those first years, primarily IBM equipment).[13]

Methods like these opened new research avenues in different fields. In political history, they allowed scholars to structure transcribed biographical data or named entities in such a way that they could go beyond the meticulous study of single significant individuals to create biographies of entire elites and discern shared features in backgrounds and social profiles.[14] In sociological history, they supported the comparative analysis of demographic data in studies of mass movements through a wide array of variables.[15] For instance, for strikes, it was possible to compare different historical cases using variables such as date, location and magnitude and to perform more advanced tasks by linking and classifying protest statements in order to then to quantify and compare their associated sentiments and discourses.[16]

Inspired by those developments, *Annales* historians became leaders in the use of computerised methods.[17] Emmanuel Leroy Ladurie's key work *The Territory of the Historian* (first published in French as *Le territoire de l'historien*, 1973) epitomises this. Its first part, 'Learning to Live with Computers: The Quantitative Revolution in History', contended that 'the computer as an instrument of historical discovery' had become 'taken for granted'.[18] Like Edward Shorter, Ladurie emphasised computational methods as a powerful way to deal with the 'bulk' of larger corpora and as a promising step forward, especially for historical demography.[19] Ladurie championed this development as a methodological emancipation, arguing that historians had hitherto been 'prisoners of their unsophisticated methods' and would need to become 'historio-metrician[s]', concluding (not unproblematically) that 'history which is not quantifiable cannot claim to be scientific.'[20] Embracing this development, Ladurie and other *Annales* historians propagated a computer-assisted 'serial history' (*histoire sérielle*). A term suggested by historian Pierre Chaunu and inspired by Fernand Braudel's notion of *longue durée*, serial history

aimed at discerning macroscopic changes in societal structures over long time periods, ideally several centuries, by de-emphasising the centrality of historical events and their ordering into linear accounts.[21] As a practitioner of serial history with the help of computers, Ladurie, for instance, analysed the developments of Parisian rents from the Middle Ages to the eighteenth century by combining and processing data from sources from a multitude of long-lived institutions such as universities, hospitals and factories.[22]

New Cinema History research does not operate with an acute historical consciousness spanning several centuries, nor does it orthodoxically follow serial history in all its assumptions. Even so, its computational approaches are fundamentally informed by those traditions' quantitative, structural approaches.[23] In what follows, we highlight those historiographic underpinnings by looking more closely at two prominent New Cinema History database projects: Cinema Context and Lantern.

In 2006, film historian Karel Dibbets launched Cinema Context, a public, online database for the analysis of film distribution and exhibition networks in the Netherlands. Over time, it has become widely known, nationally and internationally, as a pioneering New Cinema History database project, facilitating quantitative historical research, in particular network analysis.[24] In its empirical fundaments, Cinema Context's creation dates back to the late 1970s, when Dibbets began researching the development of film distribution networks in the Netherlands, primarily in the silent period.[25] Not unlike serial history in its study of the interaction among members of the elites, Dibbets' research analysed each cinema and its steering board as a node within a distribution network in order to understand their mutual relations. More specifically, he did this by studying 'interlocking directorates', focusing on how cinema board members were part of different corporations simultaneously.[26] As the basis for his network analysis, Dibbets created a dataset using the name and address lists from the Nederlandse Bioscoopbond (NBB, the Dutch Federation of Cinemas) from the late 1920s in particular.[27] More recently, this dataset was enriched with data from the files of the Dutch Board of Film Censors for the period between 1928 and 1960, and, for the early 1910s, news ads from both national and local Dutch newspapers.[28]

In development since the late 1970s, Dibbets' research was located at the cusp of punched-card-based serial history and a later wave of personal computer-based, socio-economic history. Dibbets has recounted how, when making his dataset in the late 1970s, he initially created and processed punched cards from his source materials, using the University of Amsterdam's 'super computer' facilities.[29] In more recent years, Cinema Context has used Microsoft's Access software to organise datasets about films, cinemas, people and companies, allowing users to contribute data from a wider array of sources.[30] The latter development is exemplary of procedures with widely available personal microcomputers, as those emerged during the 1980s and 1990s.[31] While tying in with the organisational modes of punched-card-based procedures, those approaches were also more flexible, particularly in terms of how historical documents were coded.[32]

Dibbets' research can be taken to reflect the epistemological underpinnings of serial history in two fundamental ways: first, in terms of how it combined text data from different collections to establish a quantitative, macro-perspective on New Cinema History, situating the historian as a mediator of data between otherwise unconnected collections, in order to combine and analyse

them in scholarly research;[33] second, as in serial history, it conformed key words from variegated document types and from different collections into a unifying coding scheme or taxonomy so as to facilitate quantitative network analysis of film-related sources on the historian's own terms.

New Cinema History projects emerging since Cinema Context increasingly rely on automated analysis of typewritten text in order to move beyond the possibilities of manual transcription and to facilitate analysis of a greater set of text features aside from named entities. A case that illustrates this particularly well is the Lantern project.[34] The Lantern search tool was developed by media scholar Eric Hoyt and his team for the analysis of primarily North American technical and fan journals digitised for the online Media History Digital Library.[35] As in the case of punched-card-based analytical procedures, the key affordance of Lantern is that it can process large amounts of data while also allowing for automation of data transcription and analysis.[36] Whereas historians hitherto manually extracted and coded words from documents to punch cards, digitisation allows for automated text analysis using either exploratory approaches, or machine learning and topic modelling.

For many document formats that are closed for editing – PDF being the most pervasive example – it is possible to use OCR to extract text and intervene analytically in entire documents rather than only a manual selection. For this reason, OCR is associated with greater comprehensiveness, in that bulks of largely unread or neglected digitised sources can be explored and analysed – in the manner of Franco Moretti's 'distant reading'.[37] Results achieved with OCR depend on a piece of software's capacity to recognise a given typography and on scan quality, and also involves coding in the process of defining and/or selecting keywords. In spite of those difficulties, it produces larger textual corpora for computational analysis than hitherto possible, and allows for visualising and linking results in a more inductive manner, because it can work with unstructured data.[38] Compared to the reliance of punched-card methods on codebooks of manually defined top-down categories, topic modelling can nurture exploratory word analysis by automatically coding and counting select numbers of topics in large datasets.[39]

Hoyt's Lantern project deployed such methods by digitising and OCR'ing 900,000 pages from public domain trade journals and making them accessible in the online Media Digital History Library.[40] The Lantern search tool, added later, allows for data mining and visualisation in combination with simpler, standard search queries to complement traditional archival research.[41] In the time it has been available, it has furthered research on silent-era periodicals in two fundamental ways: first, it allowed users to explore 'the great unread' beyond 'the canon of trade papers and fan magazines' by bringing scholarly attention to journals that had previously only been sparsely cited or neglected;[42] second, it has recast the study of silent-era publications by allowing for analysis with data mining and word-cloud visualisations created through topic modelling.[43] For instance, a select number of topics have been listed and prioritised in order to visualise how frequently they appear and in which periods they trend, so as to understand when, where and how they became prominent within the film industry or among fans.[44] By enabling the analysis of linguistic patterns within journals, Lantern's data mining tools have also facilitated the identification of relevant individuals and groups, making it possible, for instance, to discern the networks of popular cultures rather than elites.

In sum, in its approaches to data-driven analysis, New Cinema History – as exemplified here by Cinema Context and Lantern's pioneering efforts – closely aligns with computational history as it has developed from punched cards to data mining. It relies on database organisation to prepare transcribed text, at a time when data mining methods are gaining prominence because larger corpora of film-related sources are digitised and require automated methods in order to be explored. New Cinema History methods remain fundamentally based on the foundations of serial history, in particular in their focus on named entities and network analysis. Recently, however, it has seen the gradual emergence of more dynamic, exploratory approaches that are less reliant on predefined top-down categories and coding schemes.

Such developments have fundamentally shaped how media historians identify meaningful structures, topics and debates within film-related sources. However, the methods discussed ultimately remain focused on typewritten text. In this respect, they exclude potentially significant information held by the visual features of film-related sources – for instance, in logos or images – while at the same time privileging typewritten material over handwritten text. As we shall argue in the next section, which zooms in on the case of Jean Desmet's business archive, current computational New Cinema History research could be productively advanced towards greater analytical complexity by forging composite analytical approaches that combine text mining and visual classification with computer vision tools. In what follows, we discuss how this can help facilitate new types of New Cinema History research.

MIMEHIST's Combined Approach to Desmet's Heterogeneous Business Archive

The project *MIMEHIST: Annotating EYE's Jean Desmet Collection* (2017–2018) aimed at unlocking the EYE Filmmuseum's Jean Desmet Collection in the Media Suite of the Dutch national research infrastructure CLARIAH (Common Lab Research Infrastructure for the Arts and Humanities). The Desmet Collection contains the archives of film distributor and cinema owner Jean Desmet (1875–1956). It holds sub-collections of approximately 950 films produced between 1907 and 1916, a business archive containing approximately 127,000 documents, around 1,050 posters and around 1,500 photos. Parts of the collection were acquired by the Filmmuseum in 1957, shortly after Desmet's death, and then gradually expanded throughout the years with additional acquisitions.[45] The collection is unique because of its large amount of rare films from the transitional years of silent cinema, and because of the richness of its business archive, which holds extensive documentation of early film circulation in the 1910s and beyond. The documents which make up Desmet's business archive form a highly comprehensive record of all of his business activities, as well as his personal interactions. The documents date back to 1875, with Desmet's birth certificate, and continue on until the time of his passing in 1956.[46] As media historian Ivo Blom's extensive research has shown, Desmet's business archive is a rich resource for understanding film trade and exhibition in the Netherlands and its neighbouring countries.[47] This is one of the reasons why the collection was inscribed, in 2011, on UNESCO's Memory of the World Register.[48]

While Desmet's business archive was the sub-collection most central to the *MIMEHIST* project, two other sub-collections mentioned above – the films and posters – were also processed. Overall, *MIMEHIST* can be characterised as a video annotation project, in that it sought to allow scholars to create hypervideo editions of archival films, drawing inspiration from the sort of enhanced digital edition formats developed by media scholars throughout the 1990s and 2000s. Also referred to as 'historical-critical' editions, such hypervideo formats allowed commentary on films in annotations containing film-related materials.[49] Taking its cues from enhanced electronic text editions in literary studies and text-critical principles in philology, historical-critical film editions allow scholars to contextualise the style, content and circulation of digitised archival films by making use of shot segmentation, annotation and hyperlinking – for instance, for the comparative study of archival versions of films.[50]

A particularly inspirational example of such scholarship for *MIMEHIST* is the DVD series of annotated Russian and Soviet classics Hyperkino/KinoAcademia launched in 2008 (Absolutmedien and RUSCICO), created by film scholars and archivists Natascha Drubek and Nikolai Izvolov.[51] Hyperkino allowed exploration of film as a hypertext, or a form of 'hyperkino', which enables navigation between film segments and contextual materials.[52] We were inspired by this embedding of the Desmet Collection into the CLARIAH Media Suite, as this format enables scholars to build a corpus consisting of films and/or related materials using the environment's search and bookmark functionalities to subsequently link and comment on the materials in annotations.

Desmet's 127,000 business documents became central to the project, primarily because their unlocking represented a particularly complex challenge. Because of the size of the archive, it has been difficult for scholars to process it single-handedly through close reading and manual transcription. Previous efforts at transcribing data from or relating data to the Desmet Collection to create databases and apply text mining approaches have covered a very limited portion of the collection's contents, while involving extremely labour-intensive research.

In 2005, Rixt Jonkman, who was then a film student at the Vrije Universiteit in Amsterdam and is currently registrar at the EYE Filmmuseum, manually transcribed handwritten distribution and rental information from a small part of Desmet's business archive, organising it into an Access database.[53] Jonkman transcribed data on the rental, exhibition and distribution of 771 films purchased from two German distribution companies – Westdeutsche Film-Börse (WFB) and Deutsche Film Geselschaft (DFG) – between 1910 and 1912 and mentioned in Desmet's account books. The films were typically shown as part of programme packages put together and distributed by Jean Desmet. Jonkman's transcriptions contain data on named entities such as locations, screening venues, and exhibitors, and in this way, they allow network analysis to be conducted that can show who Desmet dealt with most frequently during the period based on this particular corpus of films. Cinema Context contains additional data on Desmet's distribution, not originating directly from Desmet's business archive but from Dutch newspaper clippings and advertisements.[54] Focusing on named entities such as films, cinemas, names and companies – with information on one screening per week, per title – it allows the user to study where and in which cinemas many of Desmet's films were screened.

Since the pioneering research conducted by Jonkman, Blom, and Dibbets, the business archive's status and organisation at the EYE Filmmuseum have changed fundamentally. In 2007–2008, the archive was almost entirely digitised, with subsidies from the Metamorfoze programme dedicated to the conservation and digitisation of Dutch paper heritage collections. Since then, these materials have been made available through the Filmmuseum in JPEG and TIFF formats.[55] In addition, the contents of each folder in the business archive are now described in the archival finding aid completed in 2011 by EYE archivist Piet Dirkx. In this regard, one of the great qualities of Dirkx's finding aid is that it has followed the archival principle of *respect des fonds*, re-establishing, to the greatest possible degree, the archive's original order as it appeared when first acquired by the Filmmuseum, which had changed over the decades. This allows the researcher to more easily distinguish between personal and film-related documents, as well as to find documents relating to specific companies, for instance, or to see whether documents relate to film acquisition or rental. In this sense, Desmet's business archive has become a much more readily searchable resource at the folder level, while not containing metadata at the file level, other than a file name and technical details.

In addition, and because of its digitisation, it now also lends itself particularly well to computational analysis, which allows processing of its large bulk of documents in order to facilitate more varied and complex research. It had long been a wish of the Filmmuseum to undertake such work in order to improve the papers' searchability – specifically with the help of OCR. However, the opportunity had never presented itself. This situation offered a strong incentive for *MIMEHIST* to take on the task as part of the work of embedding the documents in the Media Suite. Thus, *MIMEHIST* set out to enrich the archive with new data by experimenting with various pieces of software.

In preparing this analytical intervention, we were acutely aware of the existing database and text mining approaches in New Cinema History research, their historiographic underpinnings, and previous work on transcribing data from the Desmet collection which we discussed above. Initially, our starting point was to follow the approaches they exemplified. Yet, we quickly felt a need to go beyond them and reconsider whether they were in fact most suited for the types of documents with which we were working. In particular, we realised that a classic, text-centred OCR approach would exclude a large number of documents from Desmet's business archive, especially handwritten documents. Leaving out these documents would significantly limit the archive's searchability, as well as the temporal scope covered by the enrichments. For instance, most documents in the archive from before 1912 are handwritten and would be excluded if one was only analysing typewritten documents. Consequently, it would become difficult for researchers to rely on the enrichments to understand the developments in Desmet's activities from their very beginnings.

A great methodological challenge in this regard was that Desmet's business archive is profoundly heterogeneous in the material sense: it holds a large variety of document types. For instance, it contains telegrams, lease contracts, postcards, identity cards, brochures, promotional materials, paper clippings and lists of intertitles – all of which also contain a plethora of graphical features – and often several document types appear in one folder. Furthermore,

the archive contains a significant number of handwritten documents authored by a large number of individuals, with different styles and vocabularies, with whom Desmet corresponded – and as yet, a full overview of those individuals does not exist. This observation led us to experiment with a combined approach, which, on the one hand, involved a classic, text-centred OCR procedure, while, on the other, also explored the possibilities of visual document classification through clustering. At the time of writing, this work has resulted in an ordering of 27,758 documents into 815 clusters, ranging from particular individuals to specific institutions, cinemas, production companies, government institutions, grocery stores, electricians, tax forms, rental contracts, and film lists, among several others (see Figure 1).

Making these clusters available to researchers is useful for various reasons. First, because it allows them to identify documents pertaining to specific institutions, themes, companies or types of correspondence more easily, across folders. This may significantly aid the development of more comprehensive or clearly delimited case studies based on Desmet's archive. Second, in addition to making document retrieval and labelling easier and quicker, it may also, for instance, nurture topic analysis in the vein of the Lantern search tool by enabling the identification of word frequencies and the occurrence of words over time. However, our results, since they are not limited to the outcomes of OCR of typewritten text (as is the case in Lantern) and derive from a varied range of document types, may also ease the contextualisation of word uses.

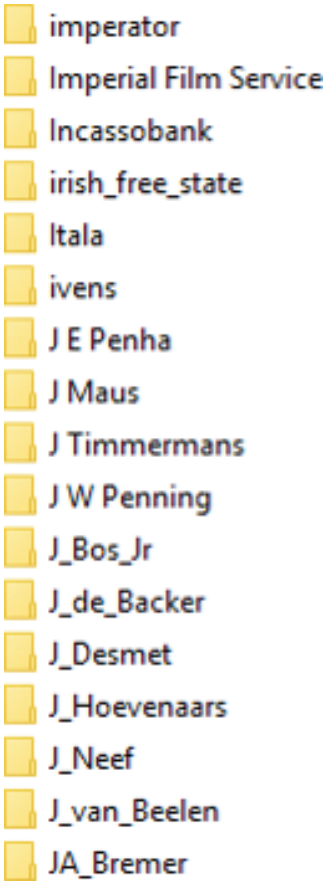

Figure 1. Details from the list of 815 clusters into which 27,758 documents from the Desmet business archive have currently been ordered.

In particular, it could be interesting to use the clustering of document types for analysing and contextualising how specific word uses related to historical events (say, World War I) differ in various document types.

In the remaining part of this section, we discuss the issues we encountered in relation to the OCR and the visual classification of the documents, respectively, before offering a reflection on the possible wider applications in research on the Desmet Collection and New Cinema History. This discussion covers the technical details of our work in quite some depth. We feel this is necessary to make our intervention as transparent as possible to other researchers interested in exploring a similar approach and to make explicit how it differs from current approaches.

The OCR of the typewritten documents in the Desmet archive raised issues concerning language, vocabulary and typography. With respect to the first topic, a key issue is that the Desmet business archive contains letters and writing from a large number of individuals, but also in different languages. While Dutch is the archive's main language of communication, and therefore our primary focus, it also has a significant share of French, English, German and Italian texts, because Desmet acquired and/or received catalogues from all over Europe. Typically, one may assume that letters only contain one language. However, as Silent Film Collection Specialist Elif Rongen-Kaynakçi pointed out, many of the letters in the Desmet collection exist in different language versions, because they were first written in Dutch as draft versions before going out in a foreign language.[56] Since we did not have a precise idea of the extent of this practice, we had to try to recognise several languages, while focusing mainly on Dutch. Furthermore, there were also unique language issues to consider. For instance, from 1912 onwards, foreign correspondence would primarily become the task of Desmet's assistant, George de Vrée, whose English and French skills, as pointed out by Ivo Blom, left a lot to be desired.[57] In addition to this, and as we have mentioned, the archive holds letters from a great variety of individuals, and thus contains many distinct voices and language styles. These issues posed a challenge to the OCR because together they create a vocabulary which is not uniform, or in some cases (as in de Vrée's writings) include mistakes that make them difficult to analyse automatically. Finally, it was evident that for letters written by Desmet and his assistant, the OCR would only work well on documents from after around November 1912, which we can infer is the time at which Desmet acquired a typewriter.

With these obstacles in mind, we first performed named entity recognition on the Dutch documents, using the Java implementation of Stanford University's Named Entity Recognizer (NER) tool.[58] Early results quickly indicated that about 70% of standard places and personal names could be recognised, as long as the first letter of the OCR'ed words was correctly read as a capital. In subsequent months, we made a special effort to build on those results to extract cinema names, places, film titles and personal names. To achieve this, we parsed a set of documents known to have lists of films, and semi-automatically annotated the film titles in them. In the summer of 2018, we had achieved an accuracy of 25% for cinema names, places and film titles, 38% for dates, but only 5% for personal names. Through iterative re-training and correction of the results produced by our classifier, the results' accuracy will likely be improved for a later data export, which will be integrated into the Media Suite.

For the purposes of distinguishing between individuals and document types within the business archive, word frequency-based document classification also proved productive. For instance, this method supported the discernment of letters from bills, based on the specific vocabularies used. Personal letters which express sympathy can, for instance, be recognised through such words as *medeleven* or *deelneming* ('sympathy') and *verlies* ('loss'). In contrast, business letters have other vocabularies associated with them. For example, words like *snoezig* ('cuddly') are unlikely to appear in a bill – unless perhaps it concerns a specific film title – while closing formulas such as *Hoogachtend* ('Yours sincerely') are more likely to be business-related. Other examples of words often associated with bills are *verplicht* ('required'), *rekening* ('bill'), *bedrag* ('amount'), *gulden* ('guilder'), *mark*, and so forth. The efficacy of different classifier types for various document types varied, but even when weaker results were achieved, they allowed for making basic distinctions. For instance, they provided a nice angle for distinguishing typewritten letters from handwritten ones: in some cases, there was both typed and handwritten information in a letter, and the OCR results could help determine whether such a document was mainly hand- or typewritten. If the OCR results were weaker, this helped indicate that it might be handwritten, and vice versa. In this way, we could, to some degree, classify document types automatically in folders (which was not possible using the existing metadata). This was beneficial for our work, because it enabled us to determine which documents could be analysed with different approaches in order to achieve better results.

In order to go beyond OCR approaches for typewritten text and to find ways for analysing the great variety of materials and handwriting in Desmet's business archive, we also experimented with computer vision methods – specifically, an HSV colour histogram classification and SIFT classification to perform image recognition for the analysis of various image features (for instance, texture). Our efforts showed that those methods could be fruitfully applied in a mutually complementary fashion with the OCR, especially for the purposes of document classification. In addition to handwritten documents, other kinds of documents also do not produce any text when applying OCR to them – for instance, photographs, various covers and empty pages. Here, computer vision methods could also be used to classify them into document types – for instance, on the basis of colour. This way, we could distinguish personal notes with addresses written on them, for example, or archive folder covers. A conclusion we drew in this regard was that if we attended to the material features of the documents in the collection beyond the content level with the use of computer vision methods, we could again add classifications to the documents which allowed us to make distinctions which the existing metadata did not. This was useful both for training our algorithms to achieve better results and as way to facilitate different ways of ordering the business archive across the folders.

In an early phase of our research, we tried to see how far we could advance our document classification in a temporary, purpose-built interface, which allowed (tentative) filtering of documents based both on the amount of OCR'ed words and on colour values such as hue and saturation (see Figure 2).

Furthermore, we conducted handwriting recognition experiments in an effort to identify authors and writing styles based on texture, using local binary pattern histograms. At an early

stage in our experimentation, our classifier succeeded in identifying some authors (primarily family members) correctly, based on training samples of a minimum of ten handwritten letters. Building on those tests, we improved the results by indicating which ones were accurate, which then allowed us to detect similarities in handwriting.[59] Even so, recognising the archive's many different handwriting styles arguably constitutes the biggest challenge and requires further research and experimentation to produce stronger results. Within the context of our project so far, we have focused primarily on Jean Desmet's own handwriting and that of his assistant George de Vrée in the hope of producing key corpora with more useful data enrichments in the upcoming months.

As a further measure to tackle the complex issue of numerous handwriting styles, we also experimented with identifying document logos. For this work, we tested various algorithms, ranging from a very simple structural similarity index comparison to a more computationally intensive scale-invariant feature transform (SIFT) approach. These approaches detect a great variety of image features, including specific objects. A homography finder using SIFT features to detect similarities proved to be the most successful, and its results could be refined in combination with the OCR. For instance, letterheads and logos from Desmet's business correspondences could be found, allowing us to take initial steps for creating an overview of Desmet's business relations and network. Textual analysis of these matches could help further refine the data and the results obtained, while also contributing to creating an overview of individuals represented in the archive, as a way to improve the basis of the OCR engine and creating training sets for further handwriting recognition. In particular, iterative retraining of both visual and text-based clustering run consecutively helped to find documents pertaining to sets of clusters that would otherwise not have been found (and that were then manually evaluated and confirmed by *MIMEHIST* researcher Kathleen Lotze). In this process, the algorithms underlying the text – respectively, visual clusterings – would find documents which the other would not, but which could be used for further training of both. For instance, a company bill with a logo which

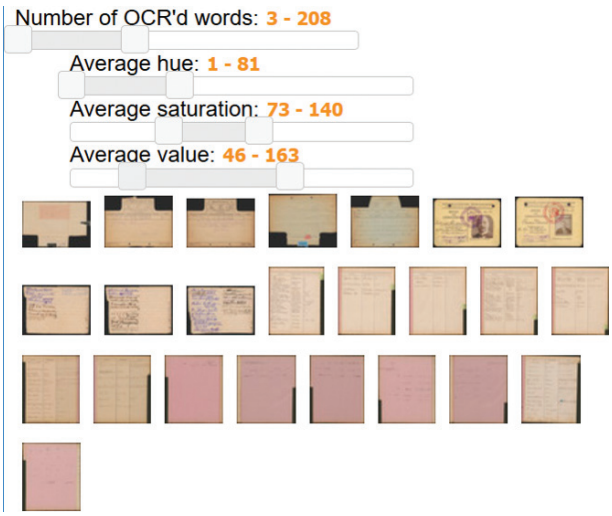

Figure 2. Detail from temporary interface used for filtering Desmet's business archive documents based on number of OCR'ed words and on colour features.

differed from the company's most known logo might not be picked up by the SIFT classifier, but the vocabulary would classify it as pertaining to that company based on the words in the text. In this way, the SIFT classifier could be asked to look for several types of logos for the same company in the subsequent training, based on the OCR results. This intervention was particularly useful for facilitating a combined analysis of different document types.

In the summer of 2018, we began to integrate the results we achieved into the CLARIAH Media Suite, starting with the OCR results (to be followed by the visual classifications later in 2018). In the Media Suite, the results appear as a separate set of annotations in addition to data from the EYE Filmmuseum's metadata and Dirkx's finding aid. Testing the results of the OCR after their integration into the research environment made it clear to us that they allow retrieval of large numbers of documents across the archive's folders in ways that were not hitherto possible. For instance, we tested the OCR with a group of media historians during a media studies summer school organised by CLARIAH, trying to find out how many documents we could retrieve that make reference to a specific word or theme – in our case, all colour-related. Using a combined Boolean query consisting of relevant, related keywords, we managed to retrieve 1,668 documents located in 136 folders.[60] Once the data resulting from the visual classification is imported, the archive's searchability will improve even further, allowing researchers to search for topics occurring in different document types while also giving the possibility to search different document types separately more easily.

By way of reflection, we would argue that the results of our experiments align with and complement existing approaches in New Cinema History research. First off, we built on methods which we know from serial history – for instance, named entity recognition, as a basis for network analysis – while also incorporating more recent text extraction methods to enable exploratory searches. Once our results are integrated into the Media Suite, media historians will be able to identify word frequencies and topics as a basis for research on early film distribution and exhibition, taking a distant reading approach to discern networks and topics on their own. Second, and most importantly, by trying out computer vision methods for the analysis of visual features, we have tried to open up an avenue for an approach to Desmet's business archive which also considers the documents as visual resources. In doing so, we seek to offer an approximation to the Desmet documents which acknowledges their visual heterogeneity and materially composite nature. Taking this approach has enabled new ways of clustering document types based on visual classification, and has contributed significantly to discerning correspondences between individuals, companies and institutions within the archive regardless of document type. These clusterings may ultimately support traditional network analysis and, thus, continue a tradition, while also offering new entry points into the archive.

The material heterogeneity we were confronted with is, of course, not unique to Desmet's business archive. In terms of the materials it contains, it reflects the wide array of document types which Verhoeven considers relevant for cinema historians – and with which cinema historians generally work. For this reason, we would argue that our approach holds a wider relevance for the field. Ultimately, the results of our research should invite a reconsideration of

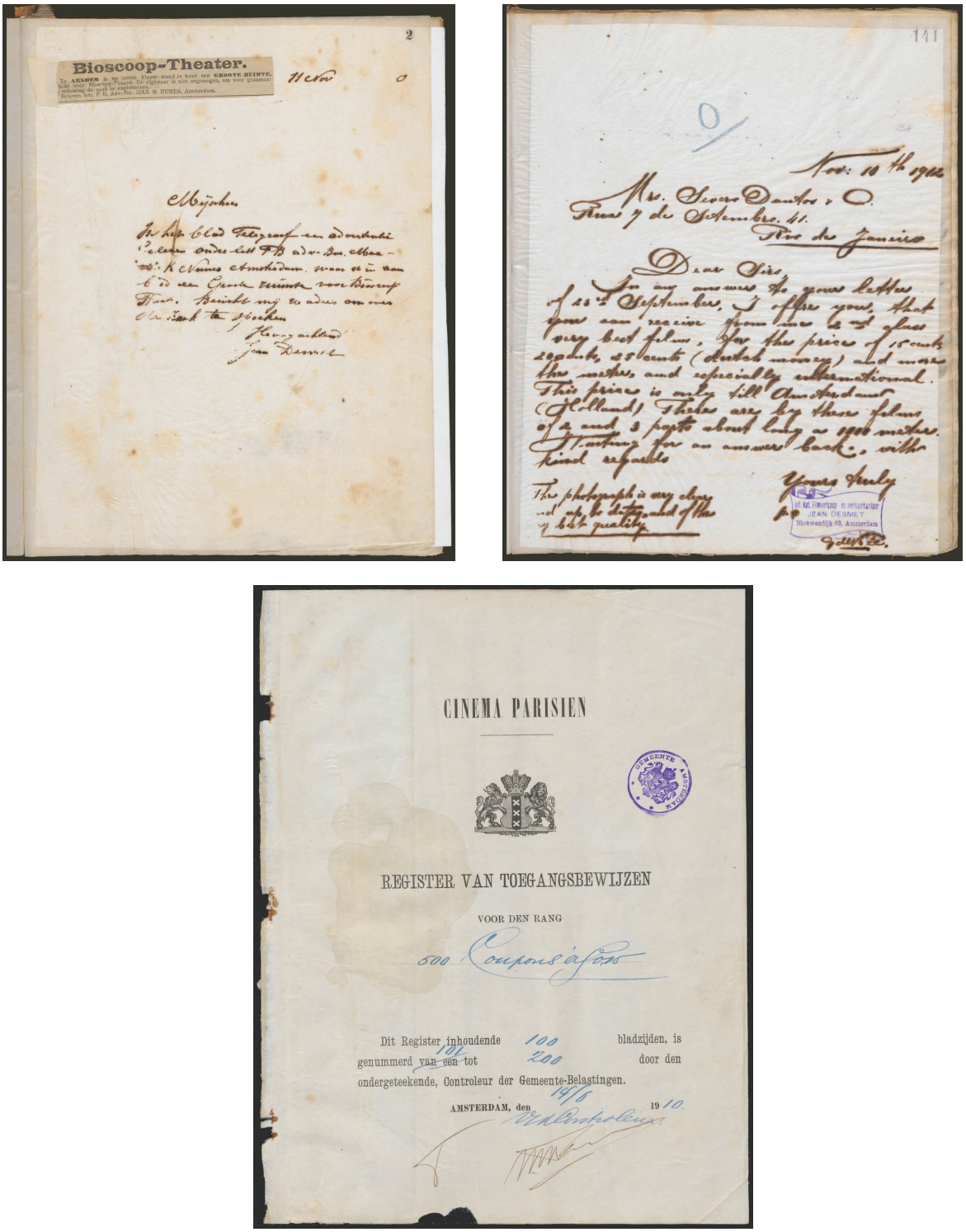

Figure 3. Three examples of the heterogeneity of materials in the Desmet business archive. Courtesy Eye Filmmuseum. Top: Carbon copy of handwritten letter in Dutch with paper clipping concerning film advertising written by Jean Desmet. Dated 11 November, 1910. Digital file MMFMA01_AF_100461. Middle: Carbon copy of handwritten letter with Jean Desmet's letter logo in English written by George de Vrée logo. Dated 10, November 1912. Digital file MMFMA01_AF_097554. Bottom: Tax registration from the Municipality of Amsterdam concerning cinema tickets dated 14 June, 1910. Digital file MMFMA01_AF_026138.

current New Cinema History approaches and their text-centred focus to facilitate automated analysis of multiple types of sources in combination. In our view, this could help advance computational New Cinema History towards acknowledging, to a greater degree, the material complexity of the sources being processed with computational methods. In this regard, we consider these experiments potentially significant first steps in a new direction for this area of research and its conceptualisations of film-related documents as historical sources.

Conclusion

After a preamble on the historical antecedents of 'big data' research in (media) history, we have suggested a new approach to the automated analysis of heterogeneous sets of film-related documents in New Cinema History research that combines text mining with computer vision methods for visual classification. Discussing the *MIMEHIST* project's efforts to unlock Jean Desmet's business archive, we have argued that the predominantly text-centred computational approaches which have hitherto informed it need to be reconsidered and complemented with methods for visual classification. As we argued, such combinations may simultaneously facilitate the searchability of a wider array of different document types by unlocking both visual and textual data. This is because the application of computer vision tools helps to add classifications which, in our case, could not be produced with the OCR, and, in combination with it, adds data which complements the existing archival finding aid significantly.

With regard to Desmet's business archive, our results constitute an extensive new dataset with which researchers can work, supporting different analytical foci by using the OCR results and clusters produced in the visual classification. For instance, it enables researchers to group together documents pertaining to specific companies, institutions or types of correspondence across folders. By the same token, the results also offer researchers an approximation to the archive's material heterogeneity and complex graphic components, which OCR alone cannot produce. Based on these results, we want to encourage other new cinema historians to experiment to a greater degree with image recognition and computer vision tools in addition to, or in combination with, OCR – in order to be able to extract and facilitate analysis of new types of data from film-related sources. While the approaches we discussed are not yet common in this field, we strongly believe that they do have a place in it in the future.

If this is the case, the interdisciplinary tradition of *Annales* historiography, which New Cinema History draws on, can be reactivated in productive ways. On the one hand, as we have argued, the text-centred quantitative procedures for processing large document collections, which New Cinema History has inherited from this tradition, need to be reconsidered, partly because of their inherent reductionisms. On the other hand, *Annales*' interdisciplinary ethos may still provide a particularly fertile ground for nurturing new types of curiosity-driven experimentation and productive encounters with the approaches of other research fields – not unlike when computational analysis became integrated into the discipline of history in the 1960s and 1970s.

Notes

1. The research for this paper was made possible by the CLARIAH-CORE project financed by NWO (www.clariah.nl). The section "New Cinema History and Text Mining: Hermeneutic Antecedents and Current Approaches" is partly based on research for first author Christian Olesen's PhD dissertation entitled "Film History in the Making: Film Historiography, Digitised Archives and Digital Research Dispositifs" (University of Amsterdam, 2017).
2. An early piece that posited the need for such research was Thomas Elsaesser's 'The New Film History,' *Sight & Sound*, 55, no. 4 (1986): 248. As Elsaesser wrote: 'To do film history today, one has to become an economic historian, a legal expert, a sociologist, an architectural historian, know about censorship and fiscal policy, read trade papers and fan magazines, even study Lloyds Lists of ships sunk during World War One to calculate how much of the film footage exported to Europe actually reached its destination.' For a recent introduction to the field of New Cinema History, we also refer to the special issue of *Tijdschrift voor Mediageschiedenis*, 21, no. 1, on "New Cinema History in the Low Countries and Beyond" (2018), edited by Clara Pafort-Overduin and Thunnis van Oort.
3. Deb Verhoeven, "New Cinema History and the Computational Turn" (paper presented at the World Congress of Communication and the Arts in Guimarães, Portugal, 2012), available online at http://dro.deakin.edu.au/eserv/DU:30044939/verhoeven-newcinema-2012.pdf (last accessed 22 February 2018).
4. Edward Shorter, *The Historian and the Computer: A Practical Guide* (Englewood Cliffs: Prentice-Hall, 1971), 22.
5. See http://mediasuite.clariah.nl (last accessed 14 August 2018).
6. Paul Ricoeur, "Archives, Documents, Traces," in Paul Ricoeur, *Time and Narrative*, volume 3, trans. Kathleen Blamey and David Pellauer (Chicago: Chicago University Press, 1988 [1978]), 116–119.
7. Andreas Fickers, "Veins Filled with the Diluted Sap of Rationality: A Critical Reply to Rens Bod," *BMGN – Low Countries Historical Review*, vol. 128, no. 4 (2013): 161. Fickers here cites historian Achim Landwehr.
8. Richard Maltby, "New Cinema Histories," in *Explorations in New Cinema History: Approaches and Case Studies*, ed. Richard Maltby, Daniel Biltereyst and Philippe Meers (Chichester: Blackwell Publishing, 2011), 32.
9. Shawn Graham, Ian Milligan and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (London: Imperial College Press, 2016), 6, and Edward Shorter, *The Historian and the Computer: A Practical Guide* (Englewood Cliffs, Prentice-Hall, 1971), 63.
10. The term's coinage is widely attributed to economic historian Stanley Reiter and dates back to 1960; see the "About" page of the Cliometric Society, at http://cliometrics.org/about.htm (last accessed 1 March 2018).
11. Lawrence J. McCrank, *Historical Information Science: An Emerging Unidiscipline* (Medford, NJ: Information Today, Inc., 2001), 62.
12. Shorter, *The Historian and the Computer*, 28–30 and 36.
13. Ibid.
14. Ibid., 16.
15. Ibid., 19.
16. Ibid., 21.
17. Ibid., 25.
18 Emmanuel Leroy Ladurie, *The Territory of the Historian*, trans. Ben and Siân Reynolds (Chicago: University of Chicago Press, 1979 [1973]), viii.
19. Leroy Ladurie, *The Territory of the Historian*, 3 and 4.
20. Ibid., 7, 61 and 15.
21. Pierre Chaunu, "Histoire quantitative ou histoire sérielle," *Cahiers Vilfredo Pareto*, vol. 2, no. 3 (1964): 166–169.
22. Ladurie, *The Territory*, 61.
23. For an in-depth discussion of cinema history's relation to *Annales* historiography, we refer to Richard Maltby, "New Cinema Histories," in *Explorations in New Cinema History: Approaches and Case Studies*, ed. Richard Maltby, Daniel Biltereyst and Philippe Meers (Chichester: Wiley-Blackwell, 2011), 3–40.
24. Karel Dibbets, "Cinema Context and the Genes of Film History," *New Review of Film and Television Studies*, vol. 8, no. 3 (2010): 335. See also: http://cinemacontext.nl/ (last accessed 22 February 2018).
25. Ibid., 334.
26. Ibid.
27. Karel Dibbets, "Bioscoopketens in Nederland: Economische concentratie en geografische spreiding van een bedrijfstak, 1927–1977" (MA thesis, University of Amsterdam, 2012 [1980]), 8–9.

28. Karel Dibbets, via e-mail of July 2, 2014, to Julia Noordegraaf, with the first author added in cc. See also "Colophon," *Cinema Context*, http://cinemacontext.nl/cgi/b/bib/bib-idx?c=cccfilm;sid=3ead0c910549a143f-f293b4fef46cb5b;tpl=colophon.tpl;lang=da (accessed 12 August 2018). This concerns, among others, *Algemeen Handelsblad* (today *NRC Handelsblaad*) and *Nieuws van den Dag*.
29. Karel Dibbets, *Bioscoopketens in Nederland*, II. As Dibbets wrote: 'Tegenwoordig is zo'n netwerkanalyse makkelijker uit te voeren dan in 1979, toen de thuiscomputer nog niet bestond. De universiteit bezat één zogenaamde supercomputer. De data en de programmatuur moest je op ponskaarten inleveren bij de balie van het Rekencentrum, waar ze 's nachts verwerkt werden. De volgende dag mocht je de resultaten komen ophalen, afgedrukt op een dik pak papier, want beeldschermen waren er nog niet.'
30. "About," *Cinema Context*, http://cinemacontext.nl/cgi/b/bib/bib-idx?c=cccfilm;sid=550f42d8bf50a542381a68245f1ac256;tpl=about.tpl;lang=da (last accessed 12 August 2018).
31. Evan Mawdsley and Thomas Munck, *Computing for Historians: An Introductory Guide* (Manchester and New York: Manchester University Press, 1993), 3 and 8.
32. Ibid., 3–4.
33. Dibbets, "Cinema Context and the Genes," 332.
34. See http://lantern.mediahist.org/ (last accessed 27 February 2018).
35. Eric Hoyt, "Lenses for Lantern: Data Mining, Visualization, and Excavating Film History's Neglected Sources," *Film History: An International Journal*, vol. 26, no. 2 (2014): 155. See http://mediahistoryproject.org (last accessed 12 August 2018).
36. Graham, Milligan and Weingart, *The Historian's Macroscope*, 2.
37. Hoyt, "Lenses for Lantern," 158.
38. Ibid.
39. John W. Mohr and Petko Bogdanov, "Introduction – Topic Models: What They Are and Why They Matter," *Poetics*, vol. 41, no. 6 (2013): 546.
40. Hoyt, "Lenses for Lantern," 146. As Hoyt explains, the OCR is 'dirty', meaning that the results (here) were not checked by human researchers afterwards.
41. Ibid., 148. See also Eric Hoyt, Kit Hughes, Derek Long and Anthony Tran, "Scaled Entity Search: A Method for Media Historiography and Response to Critiques of Big Humanities Data Research," *Proceedings of IEEE Big Humanities Data* (online), *http://doi.ieeecomputersociety.org/10.1109/BigData.2014.7004453*http://dx.doi.org/10.1109/BigData.2014.7004453
42. Hoyt, "Lenses for Lantern," 152 and 159.
43. Ibid., 148.
44. Ibid., 164.
45. Ivo Blom, *Pionierswerk: Jean Desmet en de vroege Nederlands filmhandel en bioscoopexploitatie (1907–1916)* (PhD diss., University of Amsterdam, 2000), 309.
46. See the archival finding aid made by EYE archivist Piet Dirkx for Jean Desmet's business archive, titled *144. Jean Desmet. Archief 1896–1956* (EYE Film Instituut Nederland, 2011), 3.
47. See Ivo Blom, *Jean Desmet and the Early Dutch Film Trade* (Amsterdam: Amsterdam University Press, 2003).
48. See https://www.unesco.nl/collectie-desmet (last accessed 8 July 2018).
49. For an early discussion of the CD-ROM's potential for film studies and an overview of significant editions in the 1990s, see Ben Singer, "Hypermedia as a Scholarly Tool," *Cinema Journal*, vol. 34, no. 3 (1995): 86–91.
50. Kurt Gärtner, "Philological Requirements for Digital Historical-Critical Text Editions and Their Application to Critical Editions of Films," in *Celluloid Goes Digital. Historical-Critical Editions of Films on DVD and the Internet*, ed. Martin Loiperdinger (Trier: Wissenschaftlicher Verlag Trier, 2003), 51–53.
51. See http://hyperkino.net/hyperkino/What-is-HYPERKINO (last accessed 22 February 2018).
52. Natascha Drubek-Meyer and Nikolai Izvolov, "Critical Editions of Films on Digital Formats," *Cinema & Cie*, no. 8 (2006): 209.
53. See Rixt Jonkman, "*De distributeur als programmeur*: Distributie en Programmering van films door Jean Desmet tussen 1910–1912" (BA thesis, Vrije Universiteit Amsterdam, 2005).
54. Ivo Blom and Wanda Strauven, "Cinema in context: Het einde van filmstudies?," *Tijdschrift voor Mediageschiedenis*, vol. 9, no. 2 (2006): 7.
55. Dirkx, *144. Jean Desmet*, 6. For information on Metamorfoze see https://www.metamorfoze.nl/digitalisering (last accessed 22 February 2018).

56. Interview with Elif Rongen-Kaynakçi conducted by Christian Olesen at the EYE Collection Centrum in Amsterdam, 15 June 2017.
57. Blom, *Pionierswerk*, 137.
58. See: https://nlp.stanford.edu/software/CRF-NER.html (last accessed 28 February 2018).
59. Timo Ojala, Matti Pietikäinen and David Harwood, "Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions," *Proceedings of the 12th IAPR International Conference on Pattern Recognition, vol. 1* (1994): 582–585.
60. The query consisted of the following keywords: kleur* OR color* OR colour* OR farb* OR couleur* OR tint* OR toning OR virage OR g*kleur* OR g*färb*. The query was put together by media scholars Bregt Lameris, of the University of Zurich, and Eef Masson and Christian Olesen, of the University of Amsterdam, during the CLARIAH Media Studies Summer School at the Netherlands Institute for Sound and Vision in Hilversum, 2–6 July 2018.

Biographies

Christian Gosvig Olesen is a postdoctoral researcher at the University of Amsterdam's Media Studies Department. In the research project *The Sensory Moving Image Archive (2017–2019)*, led by Professor of Film Heritage and Digital Film Culture Giovanna Fossati, he is involved in developing a search interface which enables artistic researchers to source digitised audiovisual collections based on visual features. Currently, he is also principal investigator in the NWO-funded project *MIMEHIST: Annotating EYE's Jean Desmet Collection (2017–2018)*. The project aims at embedding the Desmet Collection in the Dutch digital national research infrastructure CLARIAH and at developing an annotation environment for the collection's film and paper collections. In the academic year 2017–2018, Olesen has also been invited by the EYE Filmmuseum as the first scholar in the museum's new researcher-in-residence programme.

Ivan Kisjes studied archaeology in Leiden and is currently working as a scientific programmer at the CREATE programme of the University of Amsterdam, investigating the urban historic creative environment. He has worked on various archaeological research projects at the University of Amsterdam and as a programmer for commercial websites. His current work is in digital research support in various fields, including urban history, art history, theatre studies, media studies, and musicology.