Operationalizing "public debates" across digitized heterogeneous mass media datasets in the development and use of the Media Suite

Berrie van der Molen
Freudenthal Institute
Utrecht University, The Netherlands
b.j.vandermolen@uu.nl

Jasmijn van Gorp
Media and Culture Studies
Utrecht University, The Netherlands
j.vangorp@uu.nl

Toine Pieters
Freudenthal Institute
Utrecht University, The Netherlands
t.pieters@uu.nl

Abstract

In this paper, we propose a methodological operationalization of "public debates" as we focus on the research process of CLARIAH research pilot Debate Research Across Media (DReAM). In this pilot, heterogeneous datasets (of digitized print and audiovisual media) were made searchable with tools of the CLARIAH Media Suite, using the leveled research approach that we coined previously (combining distant and close reading) to do historical public debate analysis. The qualitative research interest in public debates on drugs and regulation is historical, but in order to bridge the gap between distant and close reading of the combined digital datasets, a number of insights from media studies is taken into consideration. The natures of the different media, the type of analysis and focus on the source material itself, and the necessity to combine historical expertise with a sensibility towards discursive relations are all considered before we argue that the accommodation of this approach in the Media Suite helps the researcher to gain an improved understanding of historical public debates in mass media.

1 Introduction

In the research pilot Debate Research Across Media (DReAM)[1], we tested and contributed to the development of the Compare tool and related tools in the CLARIAH media research infrastructure Media Suite[2][3]. We worked to accommodate the leveled research approach that we coined earlier (Van der Molen and Pieters 2017) in the Media Suite. This explorative historical research approach assumes that a combination of distant reading techniques (keyword search, word cloud analysis and timeline graph analysis) and historical analysis (close reading) of a thematic subselection can help us to trace and understand public debates in digitized historical material. By combining relevant tools in the Media Suite, we worked to make this possible across two different datasets: the digitized newspaper dataset of the National Library of the Netherlands (KB), and the digital radio and television archive of the Netherlands Institute for Sound and Vision (NISV). The research environment is built on media studies principles; one of the main aims of DReAM was to make the Media Suite equipped for *historical* public debate research. Our historical research interest in drugs and regulation was used to test the usability of the approach. In this paper, we reach a methodological operationalization of *public debate* based on (i) theoretical reflection on the relation between the digitized datasets and public opinion and (ii) reflection on (decisions made in the development of) the research infrastructure in the CLARIAH Media Suite. This gives us a pragmatic methodological framework for use of the leveled approach to

[1] < www.clariah.nl/projecten/research-pilots/dream/dream>.
[2] The MediaSuite (<mediasuite.clariah.nl>) is CLARIAH's online media research environment accessible to all humanities researchers in the Netherlands. The infrastructure consists of different tools and datasets to be combined freely by the researcher. Our research pilot helped to make a combination of tools in this environment suitable for public debate analysis for researchers in the humanities. Some of the questions raised after the presentation of (the short version of) this paper at the CLARIN 2018 conference in Pisa (Italy) regarded the accessibility of the code for others: all of the code is open and can be found at <github.com/CLARIAH/wp5_mediasuite>.
[3] CLARIAH.nl is the Dutch infrastructure related to CLARIN.eu and DARIAH.eu.

Berrie van der Molen, Jasmijn van Gorp and Toine Pieters 2019. Operationalizing "public debates" across digitized heterogeneous mass media datasets in the development and use of the Media Suite. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 205–213.

research public debates in the available heterogeneous digital sources, that works in both the historical and media studies paradigm.

Keyword search has created access to large digital datasets with historical relevance to historians that would be too time-consuming to search manually (Nicholson 2013). In DReAM we wanted to benefit from this for historical public debate research by combining a number of so-called distant reading (Moretti 2013) methods and tools in the Media Suite. The most important of these tools, Compare (Comparative search), is based on a previous CLARIAH cross-media analysis tool called AVResearcherXL (Huurnink et al. 2013; Van Gorp et al. 2015). AVResearcherXL simultaneously searched the previously mentioned KB newspaper and NISV radio and television archive and offered timeline graphs, word clouds and a result viewer.

The development process was iterative: as end users, we set out by outlining our ideas and needs in a so-called Demonstration Scenario; developers then worked on this, after which we then tested the implementations and provided feedback. As such, all developer steps were based directly on our explicit research requirements. Underlying this was our ambition to enable the leveled research approach (Van der Molen et al. 2017). This research approach is based on the assumption that navigation between three levels of reading (macro, meso and micro level; see below in-text) can function as a signposting strategy to find relevant material. The leveled approach itself is based on theoretical assumptions about how the digitized source material can be understood as relating to a public debate, which will be explained in detail below. The accommodation of the leveled approach in the Media Suite, based on our researcher needs but also on pragmatic decisions made in the development process, also frames what we mean exactly when we call this approach public debate analysis.

This paper consists of two parts. In the first part, we reflect on the methodological question of the research pilot, which results in a theoretical connection between the (digitized) source material and a particular conceptualization of public debate based on Jürgen Habermas' writing on the public sphere. The second part completes the methodological operationalization of public debate, by reflecting on how implementations in the research infrastructure lead to further explication of this type of public debate analysis that highlights discursive relations in the relevant cross-media public debates. In this second part, we argue that this approach enables researchers to uncover specific discursive strands in the source material, along with relevant results in need of close reading, resulting in an improved understanding of important themes and power relations in historical debates regarding drugs and regulation in print and audiovisual mass media.

2 Theorizing "public debates"

The methodological question that we aimed to answer in the pilot is "How can public debates on drugs and regulation between 1945 and 1990 be researched across print and audiovisual datasets?". This question means that we set out to safeguard both the historical expressiveness and the methodological soundness of the research infrastructure. Before we get to describe the implementations, it is necessary to explicate the assumption that there is a relation between the relevant datasets and historical public debates.

The qualitative research interest of the research pilot is primarily historical, as it is embedded in historical research project The Imperative of Regulation, in which the postwar drug history of the Netherlands is scrutinized[4]. The historian's primary concern is the careful contextualization of events that does justice to the actors involved. Historical research has a long tradition of source criticism and awareness of the constructive and interpretative role of historians in their efforts to produce an informed understanding of the past. Historians understandably take an ambiguous stance towards digital humanities (DH) techniques. On the one hand, they are sometimes critical towards leaving part of the interpretative process to algorithms, and the quantitative component (word counts and distances) seems to be at odds with the interpretative practice of *understanding* the past. But on the other hand, they embrace the benefits of mass access to historical sources granted by digitization (e.g. Zaagsma 2013), and recent research output continues to highlight the potential of combining historical research with mass access to digital source material (e.g. Klein 2018).

[4] <www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/i/46/13546.html>.

Although media studies is a heterogeneous field of study (Bron et al. 2016, 1536), many strands within it depart from theoretical conceptualization in order to understand the complex role of media and their relation to both media producers and media consumers. Our claim is that drawing on conceptual insights from media studies is one way to critically bridge the gap between distant and close reading of digital media sources to reconstruct historical public debates. Bridging this gap, or fully grasping what goes on between the different distant reading methods and close reading of the material, is essential in order to make the research environment solid for historical research on public debates.

Figure 1. The digitized datasets in the context of the public sphere[5]

Assuming a relation between what is said or written about a particular topic in national or local media and "public debates" of said topic seems obvious. In order to understand public debates on drugs and regulation between 1945 and 1990, it appears natural to research the newspapers and radio/television broadcasts of this time period, as is readily implied in the methodological question of the pilot. But in order to make concrete and meaningful sense of the material in these datasets, an explicit theorization of this relation is required. How *do* these mass media relate to the public debates on a national level? Jürgen Habermas has argued that in modern societies, mass media are part of a public sphere that accommodates a 'society engaged in critical public debate' (Habermas 1989, 52). This perspective, which is rooted in critical theory, is particularly useful for our research aim because of its critical stance towards power relations in society. Habermas' conception of the public sphere and mass media means that the existence of such a public sphere could foster true democratic public opinion, but it could also be a sphere in which the bourgeois class reproduces desirable political thought (Outhwaite

[5] Figures 1 and 3 were developed in cooperation with Frank-Jan van Lunteren <collageboys.nl>.

2008, 251). This can be translated into a necessity to remain observant when it comes to who is given a voice: also on radio, television and in newspapers. This is naturally relevant when it comes to debates on drugs and regulation. Is the public opinion on a particular substance mostly defined by its users, by policymakers or by different actors such as law enforcers or medical specialists? The question is not just *how* public opinion transforms over time, but also *who* gets to be a part of this process.

From a historical perspective, this means that we need to underline that tracing this type of public opinion in mass media is a very specific type of public debate analysis that focuses on the meaning-making process in national and local print and audiovisual mass media. Looking at these statements in national mass media precludes a focus on oral history, on backdoor politics, on non-mainstream media, or even on what mass media producers and consumers *actually* thought about drugs. Those would be equally relevant areas of inquiry that are not covered by this approach.

With the relation between the two datasets and public opinion established (see Figure 1), we can move on to the implementations of the research pilot in order to reach an even more precise methodological operationalization of public debate, based not just on theoretical reflection but also on the infrastructure of the Media Suite. The embedment of the levelled approach in the research infrastructure needs to enable more than the unearthing of *what* has been said about drugs and regulation in the relevant period, it also should allow the researcher to grasp what actors were featured prominently and what actors were excluded from this type of public opinion.

3 Operationalizing "public debates" in the CLARIAH Media Suite

The CLARIAH Media Suite is an online infrastructure that provides media scholars and digital humanists access to datasets from different institutional providers for exploration and mixed-media research (Ordelman et al. 2018)[6]. In order to align concrete researcher needs with the development of this infrastructure, several research pilots comprising scholars and developers tested and contributed to parts of the Media Suite during its development . Our research pilot DReAM aimed to accommodate public debate analysis capable of answering historical research questions. Public debate analysis in the Media Suite can be done by combining the digital tools Collection Inspector, Search and Compare (Comparative search) with the Workspace. Broadly speaking, Collection Inspector is used to gain an understanding of the composition of the different datasets, while Search and Compare are subsequently used to query and analyze the inspected datasets, with the Workspace allowing analysis and annotation of the bookmarked results. Below, we will describe all these methodological steps necessary for public debate analysis with the tools of the Media Suite, thereby also describing the elements of the Media Suite we tested and/or co-developed in the research pilot.

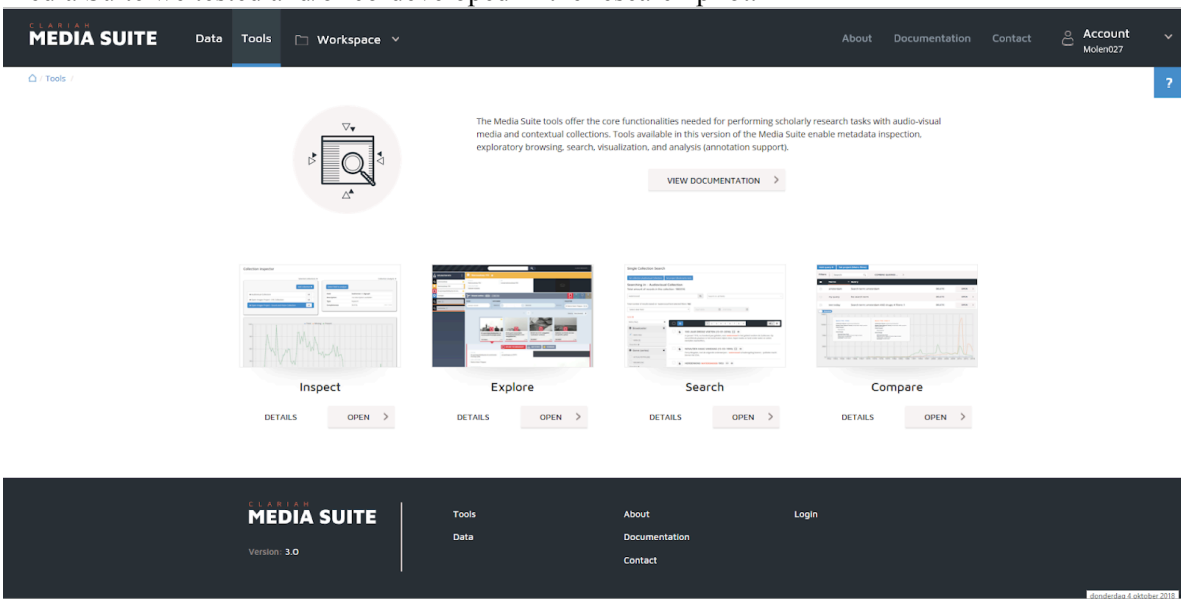

Figure 2. The tools tab of CLARIAH Media Suite Version 3.0 (October 2018)

[6] See also: contribution by Ordelman et al in this volume.

First, each dataset is loaded in the tool Collection Inspector for an assessment of metadata completeness. This allows the researcher to assess the usability of the different datasets. Historical interpretation of the data is only possible with a sufficiently complete date field for both datasets. This is a requirement, because our research question can only be answered if the data can be contextualized historically. A further requirement that needs to be checked in Collection Inspector is whether there are sufficient

a. Optical Character Recognition (OCR) metadata for the newspaper dataset
b. Automatic Speech Recognition (ASR) metadata for the radio and television datasets[7]

This ensures that both datasets are searchable on a similar (textual) level. Newspaper articles without searchable OCR metadata or audiovisual broadcasts without searchable ASR metadata cannot be found using keyword search, which is the first step of the leveled approach. The researcher is then able to send a selected dataset based on specified complete metadata to the next tool: Search[8].

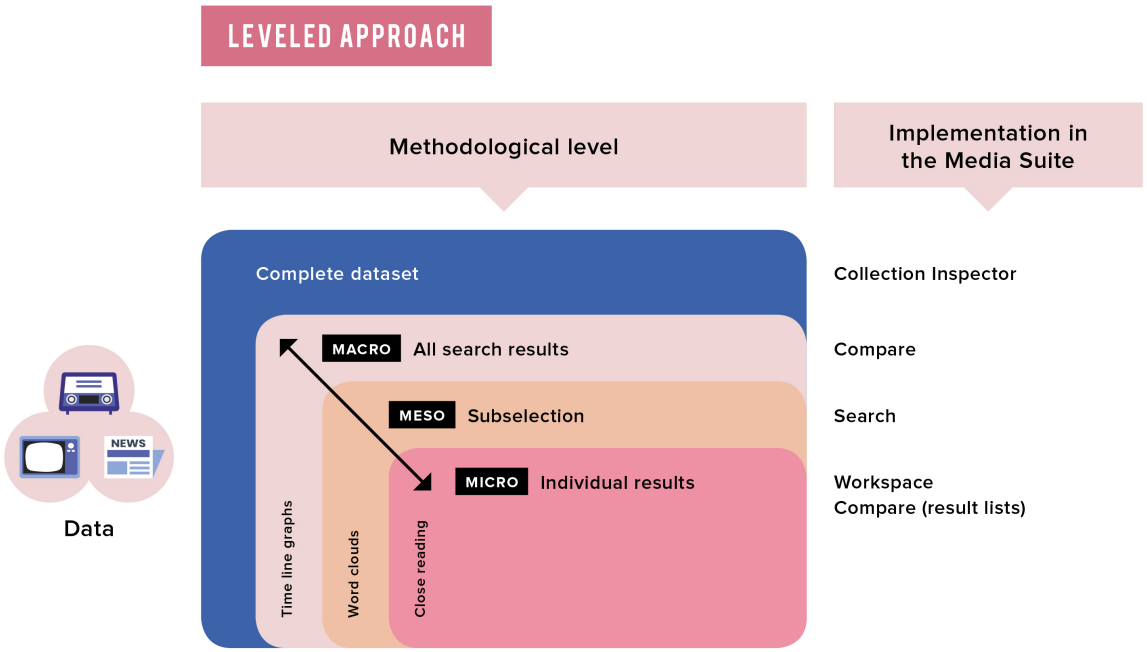

Figure 3. The leveled approach and its embedment in the Media Suite[9]

At this point, the datasets can be analyzed in the tools Search and Compare, using the leveled approach. Now that their metadata have been scrutinized in Collection Inspector, both datasets can be queried by means of specific keyword search queries (macro level) in Search. Combined queries using Boolean operators yield relevant results about particular substances. The party drug ecstasy, for in-

[7] At the time of writing (January 2019) the data integration process for the NISV data is ongoing but not completed yet. Recent status updates regarding the process can be found here: <mediasuitedata.clariah.nl/dataset/nisv-catalogue>.
[8] Depending on the type of research question and approach, the completeness of all metadata fields can be checked in Collection Inspector. The data requirements we describe in-text are necessary for the embedment of the leveled approach in the Media Suite. If, for instance, we would want to do quantitative exploration of different actors, we could search for the relevant metadata fields (e.g. actor/person/author/guest/presenter etc.) in the collections and check whether they are complete enough to proceed with such an approach.
[9] A previous version of this schematic overview appeared in the book chapter we wrote about the leveled approach before the DReAM pilot commenced (Van der Molen and Pieters 2017).

stance, is traced using the search query 'xtc OR ecsta*y OR mdma', comprising the most common spelling variations of its street name and its chemical abbreviation. This query is performed in Search for both the newspaper dataset and the audiovisual dataset, and these queries are subsequently stored in the Workspace. The Compare tool enables the researcher to display these two searches simultaneously by loading them from the relevant project in the Workspace, allowing for comparison of the different searches in the heterogeneous datasets. The search results can be further explored by means of timeline graphs (macro level) and word clouds (meso level)[10]. Furthermore, individual results are listed and can be sorted in several ways. The researcher has an option to bookmark results to the Workspace for structural analysis of the results (micro level). Although this search strategy works as a signposting strategy or funnel, the approach must be performed iteratively. Before the researcher decides to actually analyze the final subselection on the micro level, the query will most likely need to be adjusted a few times. All of these methodological steps along with their implementation location in the Media Suite have been visualized in Figure 3.

When all of these functionalities are combined in a savvy manner, they thus allow for analysis of a cross-media dataset ("public debate") that is thematically and chronologically linked. Pursuing this type of cross-media public debate analysis raises several points of reflection from a media studies perspective. First, we need to reflect on what it means to perceive combined datasets from different media types as public debates, for these media are not just neutral conveyors of messages (e.g. Derrida 1996). Any media, in this case television, radio and print media, function differently and, according to Marshal McLuhan (1964), they even *are* the message (as opposed to what we would traditionally understand as the content): what these media convey is defined to some degree by each medium. In that sense, in order to describe a historical public debate, it is necessary to understand precisely how different media can contribute to a meaningful public debate. The search results in the leveled approach remain clustered in their respective medium-specific datasets, meaning that the researcher can reflect on how the different media contribute differently to public opinion about drugs.

To complicate things more, there are two further layers/media to take into consideration: the digitization processes for both datasets, plus, most importantly, the way digitized datasets are made available and searchable in the Media Suite.[11] The textual data is searchable by means of the OCR data; the audiovisual data is searchable by means of the ASR data. Doing this on this scale is unexplored methodological territory, and it naturally forces reflection on how we can still do justice to the *visual* meanings of the television data. In other words, the distant reading steps of the leveled approach in the Media Suite are currently all based on textual metadata. On the close reading level, this is not the case: with the annotation tool the broadcasts can be annotated (based on whatever visual elements) by means of time-coded tags or comments[12].

Secondly, there are different meaningful focus points when it comes to studying media. Should a public debate analysis based on digitized newspaper, television and radio sources focus on agenda setting points (production history analysis), on what there is *in* the sources (textual analysis), or on how they were likely understood by the public back then (reception research)? All of these meanings are valid angles when it comes to researching the public sphere and public opinion, as has become clear in the previous paragraph. There are many ways to understand and account for the different levels of meaning on this continuum, for instance the encoding/decoding model that claims that audiences decode the media they consume based on their individual backgrounds, meaning that media can have

[10] At the time of writing the word cloud functionality has not been integrated yet. Since Summer 2018 this has been accommodated in a Jupyter notebook. Word cloud functionality is scheduled to be implemented in the Media Suite in April 2019 as part of CLARIAH PLUS, the follow-up to CLARIAH.

[11] It is important to be aware of *which* newspapers or television and radio shows are available in the digitized datasets too, as a reasonable sample (e.g. conservative or progressive titles or broadcasters) is necessary for the approach to yield a narrative that can truly contribute to an improved understanding of a general public opinion on drugs and regulation. A further point to be made is that the datasets primarily comprise news media, which further delineates the meaning of "public debate" here.

[12] We learned an important further lesson regarding digitization and data accessibility methods during the recent CLARIAH Summer School that was organized for Dutch researchers to test the latest version of the Media Suite with sample projects. We led a project on the representation of refugees in Dutch audiovisual media. The effect of some infrastructure design decisions on the representation of refugees was found to be considerable. Speech recognition fails at picking up non-Dutch languages, meaning that everything that has been said by refugees themselves does not become part of the searchable data, making the searchable discourse mostly defined by reporters and politicians. A number of related relevant findings from the summer school are described here: <www.beeldengeluid.nl/kennis/blog/clariah-summer-school-2018>.

different encoded (intended) and decoded (interpreted) meanings (Hall 1980). It is therefore highly problematic to assume that meaning is consistent across these different focus points. All different focus points could lead to meaningful interpretations of the relation between public opinion and mass media. Public debate analysis thus has to be explicit about where on this continuum it locates meaning and, equally importantly, which meanings are then *excluded* from this type of historical narrative. The Media Suite does not directly isolate and contextualize historical events: its reliance on distant reading techniques means that it groups historical sources based on strategies predefined by the infrastructure. The historical meaning is distilled from the digitally combined source material (found in the heterogeneous datasets) itself, which precludes a focus on production and reception. The arrows in Figure 1 could signify each of the meaningful relations (production, text, reception), but in our research approach the scope is limited to textual analysis.

The last related point of reflection is that this more or less artificial nature of public debate requires a theoretical approach. We have already established that we perceive the newspaper as one of the arenas of the public sphere in a society of mass media (as per Habermas), but we also need to be explicit about how the actual infrastructure of the Media Suite implies a particular conceptualization of historical public debates. Since the Media Suite does not isolate, group or contextualize historical events, knowledge of the historical contextualization (the different newspapers, sections, actors), along with a sensibility towards discursive relations, is necessary to signalize meaningful discursive strands within the search results. The grouped data are related in the sense that they are produced by the same society at the same time, meaning they can be understood as part of what Michel Foucault has called discourse: the culturally constructed conditions of truth. These conditions of truth are in dialogue with power over the truth in *all* relations between people (Foucault 1998, 97). This means that discursive conditions enable and restrict what can be said about ecstasy at any given time. Discourse is continuously reproduced in all relations between people, and looking at these moments of reproduction can help us understand how discourse may shift over time. Looking at actors is thus essential. Following Habermas, in modern societies, discursive shifts and its most prominent actors can be traced in the arenas of the public sphere. The digital search method makes it possible to collect and interpret a large number of articles and broadcasts mentioning any particular drug as interrelated in something that can be called a history of the public discourse of drugs in mass media. An awareness of the role of the different actors is important to understand who is most prominent in leading public opinion, which, as we have already seen, is of crucial importance to understand the role of the public sphere itself too. What needs to be researched in order to understand shifts in the discursive formation of drugs and regulation in the public sphere are all the different moments where they are discussed, which is what Foucault called looking at the techniques themselves in a search for patterns (Foucault 2004, 8). This is why the leveled approach ultimately functions as a signposting strategy: understanding historical shifts in public opinion depends on the eventual close reading of the source material.

This brings us back to the importance of researcher expertise that is crucial in the leveled approach to trace and understand the possible connections and different actors in the results. Despite plenty of noise (due to OCR issues or dual word meanings (e.g. XTC as a drug name and XTC as a band name)), sufficient historical contextual knowledge (based on historical expertise, previous research and secondary literature) allows us to recognize meaningful historical relations. The leveled research approach makes it possible to find specific relations with word clouds, and to trace these relations with targeted queries. The query '(xtc OR mdma OR ecsta*y) AND (acid OR house OR acidhouse OR dance)' yields results that relate to ecstasy's reputation of a party drug, whereas the query '(xtc OR mdma OR ecsta*y) AND (politie OR inval OR laboratorium OR onderzoek[13])' yields results about the (prosecution of the) illegal production of ecstasy. Performing the approach iteratively in the Media Suite can help us to define the most important discursive strands, that with extensive close reading can help us understand developments within public debates of or public opinion on drugs. By recognizing how a particular drug undergoes changes in the way it is framed over time (e.g. how use of the substance is either normalized or "othered") across the different results, very specific nuances can be applied to our historical understanding of the socio-cultural context of the drug, or any topic with historical relevance.

[13] The second half of this query translates as 'police OR raid OR laboratory OR investigation'.

4 Conclusion

In this paper, we proposed a methodological operationalization of "public debates" based on theoretical reflection and the resulting pragmatic development decisions we made. This approach is not aimed towards re-constructing particular debates as they happened; instead, it focuses on discursive processes and is a result of critical reflection on the CLARIAH Media Suite infrastructure, grounded in historical research and safeguarded with media studies sensibilities. By searching and analyzing the relevant datasets with the leveled approach in the Media Suite, it is possible to become aware of shifts in the discursive formation of particular topics. Although this is a fundamentally constructive exercise, reliance on historical contextual expertise makes it possible to improve our understanding of historical relations and discursive dynamics of public debates across media and the roles of the different media in this process. For our qualitative research interest in drugs and regulation, this means that tracing and following different substances in the national print and audiovisual media enables us to answer historical questions about the dynamics of public debates in mass media and about the interaction between regulation and public debates, based on fine-grained reading of the digitized source material.

References

[Bron et al 2016] Marc Bron, Jasmijn van Gorp, Maarten de Rijke. 2016. "Media studies in the data-driven age. How research questions evolve." *Journal of the Association for Information Science and Technology*. 67(7), 1535-1554.

[Derrida 1996] Jacques Derrida. 1996. *Archive fever. A Freudian impression*. Chicago: University of Chicago Press.

[Foucault 1998] Michel Foucault. 1998. *The will to knowledge. The history of sexuality: 1*. London: Penguin Books.

[Foucault 2004] Michel Foucault. 2004. *Security, Territory, Population. Lectures at the Collège de France 1977-1978*. New York: Picador.

[Habermas 1989] Jürgen Habermas. 1989. *The structural transformation of the public sphere. An inquiry into a category of bourgeois society*. Cambridge: MIT Press.

[Hall 1980] Stuart Hall. 1980. "Encoding/decoding." In: Stuart Hall, Dorothy Hobson, Andrew Love and Paul Willis (eds.) *Culture, Media Language*. London: Hutchinson.

[Huurnink et al 2013] B. Huurnink, A. Bronner, M. Bron, J. van Gorp, B. de Goede, J. van Wees. 2013. AVResearcher: Exploring Audiovisual Metadata. DIR 2013: Dutch-Belgian Information Retrieval Conference Delft: DIR.

[Klein 2018] Wouter Klein. 2018. *New Drugs for the Dutch Republic. The Commodification of Fever Remedies in the Netherlands (c. 1650-1800)*. Utrecht: Freudenthal Institute, FI Scientific Library no. 101.

[McLuhan 1964] Marshal McLuhan. 1964. *Understanding Media*. London: Routledge.

[Moretti 2013] Franco Moretti. 2013. *Distant Reading*. London: Verso.

[Nicholson 2013] Bob Nicholson. 2013. "The digital turn. Exploring the methodological possibilities of digital newspaper archives" *Media History* 19.1

[Ordelman et al 2018] Roeland Ordelman, Liliana Melgar, Carlos Martinez-Ortiz, Julia Noordegraaf. (2018) "Media Suite. Unlocking archives for mixed media scholarly research." In: Inguna Skadina, Maria Eskevich (eds.). *CLARIN Annual Conference 2018. Proceedings.* <office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf>. 21-25.

[Outhwaite 2008] William Outhwaite. 2008. "Jurgen Habermas". In: Rob Stones (ed.) *Key Sociological Thinkers*. Second Edition. New York: Palgrave MacMillan, (251-260): 254.

[Van Gorp et al 2015] J. van Gorp, J.S. de Leeuw, J. van Wees, B. Huurnink. 2015. "Digital media archaeology. Digging into the digital tool AVResearcherXL." *VIEW. Journal of European Television History and Culture*. 4.7, 38-53.

[Van der Molen et al 2017] Berrie van der Molen, Lars Buitinck, Toine Pieters. 2017. "The leveled approach. Using and evaluating text mining tools AVResearcherXL and Texcavator for historical research on public perceptions of drugs." arXiv:1701.00487.

[Van der Molen and Pieters 2017] Berrie van der Molen, Toine Pieters. 2017. "Distant and close reading of Dutch drug debates in historical newspapers. Possibilities and challenges of big data research in historical public debate research." In: Arun K. Somani, Ganesh Chandra Deka (eds.). *Big Data Analytics. Tools and Technology for Effective Planning*. Boca Raton: CRC Press, 373-390.

[Zaagsma 2013] Gerben Zaagsma. 2013. "On Digital History." *BMGN - Low Countries Historical Review*, 128.4, 3–29.