JOURNAL FOR MEDIA HISTORY

Marcel Broersma, Frank Harbers and Mark Vallinga

# Comparative Search in Digital Newspaper and Audiovisual Archives: Six Methodological Issues and Some Practical Solutions

Coverage of news events in different media is intrinsically related. News organizations build upon and respond to each other's reporting; they include facts that have been revealed by others, add new information and use different news angles while stories evolve over time. Moreover, different media compete with each other and play to their own strength in terms of their affordances. They all have their own functions in the media ecology. While radio is quick to deliver news coverage and conveys spoken text in a lively manner, television adds moving images and visual detail, making it easier to imagine what is going on. In turn, newspapers and magazines have more space for analysis and detailed background information. For media researchers it is therefore crucial to study media in relation to one another to understand how and why news stories develop as they do, and why different media present them the specific way they do.

The availability of digital media archives opens up the opportunity to search in various collections and do such research. However, that archives are usually not linked and therefore function as silos that can only be searched separately is problematic. Moreover, the institutional genesis of how such collections have been formed and made accessible, as well as how they can be searched makes linkage of archives challenging. While newspaper and magazine archives allow, for example, full text search options through keywords, searching audiovisual collections is dependent on the metadata that respective archivists have added to items. Even when oral text has been made available through speech recognition software and can be subjected to full text search, the lack of accompanying video content raises issues.

The CLARIAH Media Suite is a unique initiative to make digital collections from various media and cultural heritage institutions accessible in one research environment. Our research pilot 'Remediation of Sports News' (ReSpoNs) explored the research functionalities of this infrastructure to study how sports journalism in newspapers has changed in response to the rise of television between 1959 and 1989. Claims about how remediation took shape when television entered as the new kid on the block and newspapers as 'old media' had to 'refashion themselves to answer the challenges of new media',[1] have been asserted and theorized, but never been substantiated through systematic empirical research of news texts.[2] So far, scholars either grounded their arguments in a limited number of interviews with journalists who look back at the rise of television or, even more common, just take this causal relation for granted.

Media history often refers to sports reporting as one of the most important examples of the process of remediation. Researchers have assumed that newspapers have shifted their focus from the detailed description of competitions to a more analytical and reflective way of writing. They reasoned that this was because, first of all, matches had already been shown on television and newspaper readers already knew the outcomes. And secondly, because the visual and semi-live nature of television reports ensured that they already received a vivid picture of the match. This way viewers got the feeling they were actually live present in the stadium. Together this resulted in newspapers realizing that they could not add much in that respect anymore, and no longer needed to offer a detailed description of the complete course of the game. Instead, they shifted their focus to writing more background stories and interviews, offering analysis and opting for more human-interest-like articles in the newspaper.[3]

This narrative is compelling, but has never been tested. The lack of empirical research into remediation within the existing historiography of journalism is partly due to the time-consuming nature of archival research. Because for a long time collections were hard to access, not linked, and methodologies for analysis differed considerably, a systematic textual analysis of one medium was already an extensive and challenging project, let alone comparing two different media.[4] The comparative search 'recipe' in the Clariah Media Suite aims to enable researchers to access and search different media archives simultaneously.[5] This shows great potential to study processes of remediation as these inherently deal with material from different media, which are generally part of different archives or collections. Additionally, the availability of these datasets makes the use of computational methods possible to study developments on a large scale.[6]

Yet, the initial enthusiasm should not obscure that digitized archives and computational tools also come with a new set of issues and problems.[7] We developed a demonstration scenario to show the challenges researchers, studying remediation, encountered when exploring these collections and how these possibly could be overcome. In this short contribution we will outline more concretely both the potential of such an approach as well as the methodological issues researchers face in setting up this type of crossmedia historical research.

## Methodological issues and opportunities

1. As our research project aims at comparing sports coverage on the television and in the newspapers, our first step was to get an idea of the nature and extent of the material available, while simultaneously getting acquainted with the search functionalities of the media suite.
One problem to study the remediation of sports coverage turned out to be that while the selected newspapers have been archived integrally and digitized in full, the remaining television footage of sports matches turned out to be very scarce. This meant that when we found potentially interesting articles in the papers, we were often lacking a reference point in the audiovisual collections, even when we limited our research design to football - arguably, the most covered sport in the Netherlands. Moreover, when footage from football matches was found, it was usually limited to just the visual content of the match itself without any of the accompanying commentary. Moreover, the introduction of the coverage by a presenter in the studio and all other content from the broadcast was mostly missing. One explanation for this is that the studio presentation and the voice-over commentary were done live and have not been recorded. An alternative explanation is that these programs did get taped, but the footage got lost due to the reuse of costly film tape. In any case, the context in which football matches were discussed and appropriated is lost - which is a major disadvantage for research into remediation.

2. Next to working with differences in the availability and degree of preservation of archival material, researchers need to deal with the important and time-consuming issue of the pollution of search results. This can imply multiple copies of the same article in different newspapers that pop up in your search results, or ‘false positive’ results that do contain a search term, but are actually not relevant for a researcher’s project. Adding additional search features as well as more metadata would be useful to increase the likelihood of retrieving relevant results. One fairly easy improvement would be linking dates to an eternal calendar, which would enable searching for content that was published or broadcasted only on specific days of the week. News reporting has a certain rhythm. For example, sports coverage will often be published on Mondays, while book reviews commonly appear on Fridays, and specific television programs have a dedicated air time. Another improvement would be to add the genre of newspaper articles and television programs to the metadata - preferably as part of a uniform and consistent categorization. It would be convenient for researchers to only search in book reviews,

or in parliamentary reports, news reels or talk shows. Another feature helps to quickly scan the results: the interface of newspaper archive Delpher shows three sentences close to a keyword in the results list. This gives the researcher the opportunity to determine on that basis whether it is not interesting at all or still worth clicking on.

3. Another related issue pertains to the diverging search options for digitized newspaper archives and the audiovisual collections, which differ considerably. While newspapers have only recently been digitized, in longitudinal batches that run from the start of the paper to either its closure or – due to copyright laws – 1995, they have uniform full text search functions through keyword search. This is very different from the audiovisual archives which are searchable mostly through metadata. These descriptions of the content are often very brief and sometimes incomplete, making it much harder to find material than in the newspaper archives.

The most important effect of only being able to search the metadata is that one can only search on a far more general level, limiting the search options considerably. One way to improve searchability, which archives are actively pursuing, is using speech recognition techniques to transcribe spoken text in audiovisual content. To a certain extent, this enables similar possibilities as a full text search. Yet, even when spoken text is transcribed, the options remain far more limited than searching full text in the newspaper archives because the context of the spoken text is missing. Therefore, results are harder to interpret. And of course, the visual features, which in themselves also provide much information, are still not searchable. Finally, given these considerable differences in search possibilities, it is also hard to present search results from both collections in one interface and in a meaningful and comparable way.

4. Apart from the generally limited search possibilities of metadata compared to full-text search, the consistently differing nature and quality of metadata between collections also poses problems. While in the newspaper archives the metadata that have been added are limited to the kind of article (advertisement, personal announcements, illustration, and editorial articles), distribution area (national, local, or one of the former Dutch colonies) and place of publication, these are *post hoc* allocated in a logical, uniform and consistent way. In contrast, the audiovisual archives have a long historical genesis with different curatorial strategies and systems of categorizing material. This

means that, depending on the moment the material was added to the archive, different systems for assigning metadata with their own set of terms and categories, have been applied. This makes the results of search queries rather unreliable, and it requires quite a bit of knowledge about the structure and development of the archive to get (all) relevant results - far more than is necessary for searching newspaper archives. This also implies that while a researcher can be quite confident to get all results when doing a search in the latter archive, they can be less sure about the hits generated in the audiovisual collection. This makes a quantitative exploration of the data - for example by listing numbers of articles and television items on a timeline - a rather treacherous, if not impossible, endeavor which is hard to substantiate methodologically.

5. Storage functions are very important for an investigator to keep an overview and to get back to relevant articles and videos at a later stage of the investigation. The Media Suite offers good possibilities for this. Newspaper pages can be saved in a list of favorites and downloaded, and articles can be cut from the page and downloaded as a jpg-file. Moreover, the permalink and a complete reference are easy to copy. Similarly, different items can be cut from the televisions program they were part of and stored. It is convenient to be able to create different personal folders in the interface so articles and items can be orderly stored and easily retrieved. Furthermore, after (or during) the compilation and storage of research material, the Clariah Media Suite provides (video and textual) annotation tools, which make it possible for researchers to add metadata themselves. This helps them to keep track of their results, but also to analyze the data according to specific research questions. For audiovisual archives, to determine whether a program or item is interesting for research, it is useful when the metadata contains a minute-to-minute description of the content. The catalogue of the Netherlands Institute for Sound and Vision, for example, helps the researcher to quickly slide to a certain moment in the broadcast. If the detailed descriptions of the material would be displayed clearly (preferably next to the video), this would be a big plus for viewing the visual material.

6. In order to compare television and newspapers both databases must be accessible in one interface. In the Media Suite and in a previous tool, AV Researcher XL, newspapers and audiovisual material can be searched simultaneously and plotted on a graph that visualizes the number of hits in a certain timeframe. This makes it much easier for the researcher to match and compare relevant hits. Yet,

such a comparative search raises a serious issue in terms of segmentation. In newspaper archives, articles are sometimes incorrectly segmented on the page, i.e. multiple articles are considered as one, or one article has been split in various segments that are considered separate pieces. In contrast, in the audiovisual archive whole television programs count as separate units in the search results. In terms of the newspaper, this would mean that the entire page or even the entire newspaper would be included in the hits instead of an article. Similar to searching in newspapers, as a researcher you would prefer to search and count different items in a broadcast. For example, if a newsreel consists of ten different items about different topics, you would like to be able to search, count, and analyze only the ones you are interested in. This structural difference between the level on which the content in newspaper and audiovisual archives is segmented results in a severe impediment of making a systematic comparison between, in our case, the development of sports coverage on television and in the newspaper. In order not to compare apples with oranges, it is important for a comparative analysis between newspapers and television to equalize the measurement level between both collections as much as possible. This means that segmentation must also take place at item level in audiovisual collections.

## Final Remarks

Our research pilot took place in the initial phase of developing the CLARIAH Media Suite. Developing such a complex research environment comes with challenges. One challenge such an ambitious endeavor faces is copyright issues related to the collections involved. It is incredibly challenging for individual archival institutions to reach agreements with publishers, broadcasters and media makers such as journalists and photographers to make collections available either on-site or through their website. Bringing them all together in an encompassing research environment and making them searchable in one interface, brings up new issues that need to be solved both legally and technically (for instance by approaching collections from a central research environment but keep them stored on the servers of the individual archival institutions). Solving such issues takes time, which meant that not all of these issues were solved yet during our project. This meant that comparative search could not be implemented yet in the Media Suite, so we did our pilot research through parallel searches in the individual collections, and by using a previous version of the comparative tool on a much smaller part of the collections.

A major benefit from the Dutch situation compared to other countries, however, is that newspaper and audiovisual collections are quite complete at the Royal Library and the Netherlands Institute for Sound and Vision respectively. While in other countries especially newspapers are part of different public and commercial archives and settings - which makes it even more complicated to link them in a research infrastructure - the Dutch collections are mostly publicly available at a limited number of institutions, which makes it easier to link them and accomplish comparative search options. When this is achieved and the research tools within the Media Suite as well as opportunities to work with the raw data are added, fascinating opportunities for doing intermedia and comparative research are opened up that allow for research on a far larger scale and deeper level than is currently possible.[8]

What is crucial, though, is keeping in mind that in the construction of such an extensive research infrastructure making collections digitally accessible and (simultaneously) exploring (various) collections is not the ultimate goal. While enabling digital access and facilitating various search opportunities is important in itself - and a major leap forward in comparison with what was possible before digitizing collections - the next step in media and digital humanities research should be to use the infrastructure to do (digital) research that answers pressing research questions in the field of media history and media studies at large. During our pilot project the Clariah Media Suite, and the comparative search recipe in particular, were still under development. But working with the infrastructure and the individual digital archives has shown the potential for further grounding remediation as a theoretical concept and answering research questions related to it that were earlier hard to study empirically.

## Notes

[1] J.D. Bolterand and R. Grusin, *Remediation. Understanding New Media* (Cambridge: MIT Press, 1998), 6.

[2] Cf. Marcel Broersma, "Journalism as Performative Discourse. The Importance of Form and Style in Journalism," in *Journalism and Meaning-Making: Reading the Newspaper*, ed. Verica Rupar (Cresskill, N.J.: Hampton Press, 2010), 15–35.

[3] Ruud Stokvis, "Een genre in beweging. De ongemakkelijke verhouding tussen sport en journalistiek," in: *Journalistieke Cultuur in Nederland*, ed. Jo Bardoel, Chris Vo, Frank van Vree en Huub Wijfjes (Amsterdam: Amsterdam University Press, 2002), 191–207; Frank Harbers, *Between Personal Experience and Detached*

*Information. The development of reporting and the reportage in Great Britain, the Netherlands and France, 1880–2005* (Groningen: s.i, 2014).

[4] Marcel Broersma, "Nooit meer bladeren. Digitale krantenarchieven en onderzoek naar journalistiek," *TMG Journal for Media History*, Special Issue Digital Archives, 14, no. 2 (2011): 29–55; Sonja de Leeuw, "Het archief als netwerk. Perspectieven op de studie van online televisie-erfgoed," *TMG Journal for Media History*, Special Issue Digital Archives, 14, no. 2 (2011): 10–28.

[5] Jasmijn van Gorp, Sonja de Leeuw, Justin van Wees and Bouke Huurnink, "Digital Media Archaeology: Uncovering the Digital Tool AVResearcherXL," *VIEW Journal of European Television History and Culture* 4, no. 7 (2015): 38–53.

[6] Marcel Broersma and Frank Harbers, "Exploring Machine Learning to Study the Long-Term Transformation of News," *Digital Journalism* 6, no. 9 (2018): 1150–1164.

[7] Huub Wijfjes, "Digital Humanities and Media History: A Challenge for Historical Newspaper Research," *TMG Journal for Media History* 20, no. 1 (2017): 4–24.

[8] Cf. Broersma and Harbers, "Exploring Machine Learning".

## Biographies

**Marcel Broersma** is a full professor and director of the Centre for Media and Journalism Studies at the University of Groningen. He is also the academic director of the Dutch Research School for Media Studies (RMeS) and coordinator of the national VSNU Digital Society research program. His research focuses on the current and historical transformation of journalism, changing media use and digital literacy, and digital humanities. Broersma published numerous articles in peer-reviewed journals, chapters, monographs, edited volumes and special journal issues on media history, social media, transformations in journalism and political communication, among which *Redefining Journalism in the Era of the Mass Press, 1880–1920* (2017; edited with John Steel).

**Frank Harbers** is as an assistant-professor at the Centre for Media and Journalism Studies at the University of Groningen. He received his PhD in 2014, which focused on the development of journalism in Great Britain, the Netherlands and France since the second half of the 19th century. In 2016 he was researcher-in-residence at the National Library of the Netherlands for which he conducted a digital humanities project into automatically classifying the genre of historical newspaper articles. In 2018, together with Huub Wijfjes, he published an edited volume on the history of the press in the Netherlands. His research interests focus on (comparative) journalism history, digital humanities approaches to journalism history, narrative forms of journalism, and journalistic innovation. Harbers has published several articles about all of these themes in refereed journals and edited volumes.

**Mark Vallinga** holds a BA in history and an MA in journalism from the University of Groningen. After finishing his MA in 2017, he was involved as junior researcher within the Clariah pilot study *Remediation in Sports News* (ReSpoNs) between 2017 and 2018, for which he dove into the media archives to uncover historical research material, and he explored opportunities of digital methods and tools for research into sports journalism history. He currently works as a reporter for *Dagblad van het Noorden*.

**TMG Journal for Media History**
Volume 24 No (1/2)/2021

**DOI**
https://dx.doi.org/10.18146/tmg.813

**PUBLISHER**
Netherlands Institute for Sound and Vision