Researching drugs in Dutch cross-media debates. Using and developing the Comparative Search tool in CLARIAH's Media Suite for distant and close reading of cultural heritage "big data"

Berrie van der Molen, Freudenthal Institute, Utrecht University, The Netherlands

Since the digital turn (Nicholson 2013), it has been possible to do "bottom-up" keyword search in large amounts of digitized newspaper data. In CLARIAH research pilot DReAM (Debate Research Across Media) we have contributed to the development of a tool that simultaneously searches the digital Royal Library's newspaper dataset and the digitized television and radio archive of the Netherlands Institute for Sound and Vision. We worked to enable public debate research across both textual and audiovisual cultural heritage documents in a fusion of digital methodology with historical research questions.

My historical research interest is in public debates on drugs and the historical interaction between these debates and drug regulation between 1945 and 1990. In order to understand the history of the public framing of different substances in The Netherlands, I study historical Dutch media debates in newspapers and on radio and television. In this study I combine distant and close reading techniques in what we have called the leveled approach (Van der Molen et al 2017). The benefits of distant reading help to unlock the historical potential of huge datasets, while the relevant material is still subject to historical interpretation ("close reading"). The aim of research pilot DReAM was to accommodate this strategy for both print and audiovisual datasets in the Comparative Search tool. The potential benefits are numerous: it creates opportunities for enriched, fine-grained analysis on a large scale over long periods of time across media; a lot of time could be saved by not having to manually search for all the relevant material; material that may have stayed buried in huge datasets could rise to the surface.

This short presentation consists of two parts. The first part outlines the DReAM proceedings in the co-development of the Comparative Search tool. How did we seek to ensure that using distant reading techniques yields results that can produce a credible (cultural history) narrative? Although the potential of semantic text mining for historians could be huge, there are still hurdles to overcome here (Snelders et al 2017). And how did we deal with the datasets' different (and uneven) access and Intellectual Property Right contexts? The second part offers methodological reflection on simultaneous keyword search of textual and audiovisual data. How does subtitle or ASR (Automatic Speech Recognition) metadata relate to the audiovisual material, and, importantly, how does this type of metadata compare to OCR (Optical Character Recognition) metadata from the newspaper dataset? What kind of disciplinary connections should be forged to approach textual and audiovisual material in a similar and sound way? Although this presentation has a methodological focus, it refers to examples from my drug debate study to clarify the more abstract parts.

References

Nicholson, Bob. “The digital turn. Exploring the methodological possibilities of digital newspaper archives” *Media History* 19.1 (2013).

Snelders, S, Huijnen, P, Verheul, J, de Rijke, M and Pieters. T. “A Digital Humanities Approach to the History of Culture and Science: Drugs and Eugenics Revisited in Early 20th-Century Dutch Newspapers, Using Semantic Text Mining”. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*. London: Ubiquity Press, 2017: 325-336. DOI: https://doi.org/10.5334/bbi.27. License: CC-BY 4.0

Van der Molen, Berrie, Toine Pieters. "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, 2017: 373-390.