volume 4 issue 7/2015
VIEW
JOURNAL OF EUROPEAN TELEVISION HISTORY & CULTURE

DIGITAL MEDIA ARCHAEOLOGY

DIGGING INTO THE DIGITAL TOOL *AVRESEARCHERXL*

Jasmijn Van Gorp
Utrecht University
Muntstraat 2a
3512 EV Utrecht
The Netherlands
j.vangorp@uu.nl

Sonja de Leeuw
Utrecht University
Muntstraat 2a
3512 EV Utrecht
The Netherlands
j.s.deleeuw@uu.nl

Justin van Wees
Dispectu
Nieuwezijds Voorburgwal 130 C
1012 SH Amsterdam
The Netherlands
justin@dispectu.com

Bouke Huurnink
The Netherlands Institute for Sound and Vision
Postbus 1060
1200 BB Hilversum
The Netherlands
bhuurnink@beeldengeluid.nl

Abstract: Recently, scholarly works started to turn their interest to the epistemological and methodological challenges that research with new digital tools and technologies do pose. In this article, we would like to contribute to this methodological discussion and to shed light on the role of digital tools for media studies, by taking the tool *AVResearcherXL* as case in point. *AVResearcherXL* is a new exploratory tool for media studies research, enabling users to search across, compare and visualize both the metadata of Dutch public television and radio programmes, and a selection of Dutch newspaper articles of the Dutch Royal Library. By tracing the word 'television' with the use of the tool, we provide a practical use case of doing media archaeology with digital tools for media archives. Our deconstruction shows the importance of a media archaeological approach to look into the materiality of digital technology as well as the relevance of studying the deep material structure of media technology. *AVResearcherXL* thus could be seen as an archaeological site in which the user or 'archaeologist' decides where to dig and which search lights to use. Using *AVResearcherXL* to do media

(historical) research is not about finding the 'right' answers, but about contextualising results, and about finding new, sometimes unexpected, pathways and questions.

Keywords: digital humanities, methodology, digital tools, television archives, newspapers

1 Introduction

Over the past decade, interdisciplinary teams developed several new tools to search across large, diverse and dispersed digital and digitized collections. The development of these new digital tools has been largely approached as 'a practical revolution: it has made research faster, easier, more convenient and more productive.'[1] Scholars working in computer sciences often write 'demo papers' in which digital tools and their technicalities are showcased.[2] Other scholarship on digital tools takes a social sciences perspective, mainly aiming at describing and studying 'users' and 'user behaviour.'[3] User behaviour is logged, tracked, and measured to gain insight in the working of the digital tools.

Recently, scholarly works started to turn their interest to the epistemological and methodological challenges that research with new digital tools and technologies pose. The challenge at stake is well formulated by American media scholar Tara McPherson: 'The role of computation in the humanities is about much more than building robust archives that scholars then write about in traditional ways (…); it is also about navigating new pathways through scholarly material that can transform the questions scholarship might ask.'[4] In this article, we would like to contribute to this methodological discussion and to shed light on the role of digital tools for media studies, by taking a tool which we developed ourselves as case in point: *AVResearcherXL*. It is a new exploratory tool for media studies research, enabling users to search across, compare and visualize both the metadata of Dutch public television and radio programmes, and a selection of Dutch newspaper articles of the **Dutch Royal Library**.[5]

We approach the tool by combining insights of digital humanities with media archaeology. Digital humanities is largely a practical discipline or a 'generative' enterprise: it is about *making* things, such as texts, software and platforms.[6] Digital humanities is booming nowadays, after a first wave in the late 1980s. Entangled with linguistics and literature, it focused mainly on textual corpora and cataloguing, linguistic features, learning environments and structured data,[7] and it still does to a large extent. Linguistics and literature have a longer discursive relation with the concept of digital humanities than media studies.[8] Scholarship on audiovisual data and audiovisual archives is only recently more explicitly present in digital humanities journals such as *Digital Humanities Quarterly* and *Journal of Digital Humanities*, and at *DH*, the major conference of the Alliance of Digital Humanities Organizations.[9]

1 Bob Nicholson, 'The Digital Turn: Exploring the methodological possibilities of digital newspaper archives,' *Media History*, vol 19, (1), 2013, p. 59–73, p. 61.

2 For an example of a 'demo paper', see Bouke Huurnink, Amit Bronner, Marc Bron, Jasmijn Van Gorp, Bart de Goede, and Justin van Wees, '**AVResearcher: Exploring Audiovisual Metadata**', *DIR2013:Dutch-Belgian Information Retrieval Conference*, 2013.

3 As also outlined by Michael Goddard for the Anglo-American context, Michael Goddard, 'Opening up the Black Boxes: media archaeology, 'anarchaeology' and media materiality,' *New Media & Society*, April 2014, p. 1–16, p. 2.

4 Tara McPherson, 'Introduction: Media Studies and the Digital Humanities', *Cinema Journal*, 48, 2, 2009, p. 119–123, p. 122.

5 *AVResearcherXL* is developed by the Netherlands Institute for Sound and Vision, Centre for Television in Transition at Utrecht University, and ILPS at University of Amsterdam, the Netherlands.

6 Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, Jeffrey Schnapp, *Digital_Humanities*, MIT Press, p. 10.

7 Ibid., p. 8.

8 It is difficult to pinpoint, but 2009 seems to be a turning point for a more discursive connection between the terms 'Digital Humanities' and 'Media Studies' in publications.

9 The latter recently approved a Special Interest Group in Audiovisual Material for Digital Humanities, see **https://avindhsig.wordpress.com/**. The relative scarcity of Audiovisual Data research in Digital Humanities might be related to the fact that it is easier to digitize and make available written sources than audiovisual sources, both in terms of effort and in terms of right issues.

Media archaeology is a theoretical and philosophical (un)discipline, which has become a common reference in debates on digital archives. Media archaeology, though, has neither a permanent home, nor a clear-cut methodology. There are variable approaches to media archaeology as Huhtamo and Parikka discuss, and this variety indeed seems to particularly represent the core of it.[10] Even more so, every new media archaeology text increases the discipline's complexity.[11] What seems to be a recurrent pattern in the discourse on media archaeology is the notion of looking for *alternative histories*, an act of reading against the grain, "a hermeneutic reading of the 'new' against the grain of the past, rather than telling the histories of technologies from past to present."[12] In the same vein, Zielinski argues how media are spaces of action for constructed attempts to connect what is separated. He calls this 'anarchaeology' and argues how in the longer term, "the body of individual anarchaeological studies should form a 'variantology of the media." What he suggests is that instead of "looking for obligatory trends (…) one should be able to discover individual variations."[13] Parikka discusses the archive as a key site where media archaeology takes place. As the archive has increasingly become a digital archive, media archaeology cuts across digital humanities.[14]

Digital humanities and media archaeology apparently share a core interest in (the relativity of) 'the new.' The alternative paths or counter-histories of media archaeology methodologically touch upon the key question of digital humanities: does our research (radically) change by using (new) digital tools? Scheinfeldt pinpointed the key question of digital humanities to "where is the beef?"[15] Or what do we learn what we could not know before?[16] The parallel between digital humanities and media archaeology not only lies in (potentially) bringing to the fore *alternative histories*, but also in fundamentally raising new methodological questions. As a consequence, this involves the practice of doing media history and encourages reflections on how this practice might change. We argue in this article that a humanities approach to digital tools is not so much about providing an empiricist 'proof' that the tools are different from (and implicitly better) than the 'old' standard tools, but rather about providing a better understanding of how these tools could be used and what they mean for media studies research with a focus on the tools' ambiguities and shortcomings. To put it differently, by considering digital humanities through media archaeology, we aim to encourage a critical approach to using digital tools for media historical research.

A dialogue between digital humanities and media archaeology helps to understand the digital tool as media technology and to take its particular material nature into consideration. According to Ernst, a scholar in the digital age needs competence in informatics to reach the sub-semantic strata of media culture as well as the non-cultural dimensions of the technological regime making cultural analysis calculable.[17] This speaks to the ecological turn in media archaeology and to a geological (not only a material) approach to media as put forward by the British media culture scholar Goddard. Goddard sees the value of media archaeology in its insistence on materiality, on what he calls "material ecologies of media objects, systems and processes."[18] He argues for an opening up of the black box of technology, paying attention "to the material ecologies of human, non-human and machinic entities, the inorganic, organic and (…) geological strata that underlie technical media systems and networks."[19] This is in line with what Parikka suggests as studying the

10 Erkki Huhtamo and Jussi Parikka, eds, *Media archaeology: Approaches, applications, and implications*, University of California Press, 2011. They refer to early examples of Walter Benjamin's unfinished *Das Passagen-Werk* (1927–1940) and to Michel Foucault's, *The Archaeology of Knowledge and the Discourse on Language*, Pantheon Books, 1972.

11 Michael Goddard, 'Opening up the Black Boxes: media archaeology, 'anarchaeology' and media materiality,' *New Media & Society*, April 2014, p. 1–16.

12 Erkki Huhtamo and Jussi Parikka, eds, *Media archaeology: Approaches, applications, and implications*, University of California Press, 2011.

13 Siegfried Zielinski, *Deep Time of the Media. Toward an Archaeology of Hearing and Seeing by Technical Means*, MIT Press, 2006, p. 7.

14 Jussi Parikka, *What is Media Archaeology?*, Polity Press, 2012, p. 15. See also his blog ***Machinology***; and Jussi Parikka, 'Archives in Media Theory: Material Media Archaeology and Digital Humanities', in David Berry, ed, *Understanding Digital Humanities*, Palgrave MacMillan, 2012, p. 85–10.

15 Tom Scheinfeldt, 'Where's the Beef? Does Digital Humanities Have to Answer Questions?', in Matthew Gold, ed, *Debates in the Digital Humanities*, University of Minnesota Press, p. 56–58.

16 Ibid.

17 Wolfgang Ernst, 'Media Archaeography: Method and Machine versus History and Narrative of Media', in Erkki Huhtamo and Jussi Parikka, eds, *Media archaeology: Approaches, applications, and implications*, University of California Press, 2011, p. 249.

18 Michael Goddard, 'Opening up the Black Boxes: media archaeology, 'anarchaeology' and media materiality,' *New Media & Society*, April 2014, p. 2.

19 Ibid.

'materialities of materials,' approaching media technology through the various materials, minerals, components, signs, meanings and attractions.[20] We follow this media archaeological approach by deconstructing *AVResearcherXL* as material structure, focusing on its components, and its materials (as to be found in its layers).

In order to demonstrate the methodological potential of the tool, we take a use case to explore the archive. In this respect, we follow the approach used by media historian Bob Nicholson as discussed in the journal *Media History*, in which he demonstrates the value of digital newspaper archives for media historical research by tracing the word 'America' in online newspapers.[21] He combines a hands-on showcase of the search term 'America,' drawn from his own research into the late-Victorian transatlantic press, with critical accounts on how new methodologies created by keyword search might be applied. Instead of 'America', we chose the term 'television' as we wanted to find out what kinds of representations and research questions about television are triggered by the tool. Tracing the word 'television' in the digital archives adds a second, meta-layer to this article as it sheds light on discourses about television in television and radio programmes, metadata, subtitles and in transcripts of newspapers and thus potentially makes visible how television, radio and other media represent 'television.' Any representations are expected to support an understanding of television's role in mediating history and eventually identity. We explicitly invite the reader to explore the tool together with us. We provide hyperlinks to all retrieved documents and programmes, and also videos when available. Taken together, the article provides methodological strategies to cope with a digital tool such as *AVResearcherXL* and thus aims to further enhance an understanding of digital media archaeology as an opening to media historical inquiry.[22]

We proceed by first giving an introduction to *AVResearcherXL* and its material structure, followed by a use case demonstration and discussion of the tool by tracing the word 'television,' and finally concluding insights about the meaning of our (re)search for digital media archaeology.

2 *AVResearcherXL*'s Material Structure

AVResearcherXL is a digital tool used for comparing two sets of items in the Dutch public radio and television archive and the newspaper archive of the Dutch Royal Library. *AVResearcherXL* is an extended version of *MeRDES*[23] and *AVResearcher*,[24] the tools developed in 2012 and 2013 by the project BRIDGE and the Netherlands Institute for Sound and Vision. Standard search tools are typically supporting searches by professionals, who already know what they would like to find, while media researchers prefer to dive into archives, explore and grasp pathways for their own research projects.[25] BRIDGE wanted to develop new type of tools, conceived to be 'exploratory search systems,' supporting the exploration of media archives by media researchers.[26] *MeRDES* was the first prototype aiming to

20 Jussi Parikka, 'New Materialism as Media Theory: Medianatures and Dirty Matter,' *Communication and Critical/Cultural Studies*, vol 9 (1), March 2012, 95–100, p. 97. For Parikka this would also involve a more political analysis of media culture, including the cheap labour in media technology factories.

21 Bob Nicholson, 'The Digital Turn: Exploring the methodological possibilities of digital newspaper archives,' *Media History*, 19 (1), 2013, p. 59–73.

22 This is also argued for by Jussi Parikka on his blog *Machinology*, ibid.

23 Marc Bron, Jasmijn Van Gorp, Frank F. Nack, Maarten de Rijke, Andrei Vishneuski, and Sonja de Leeuw, '**A Subjunctive Exploratory Search Interface to Support Media Studies Researchers**,' *SIGIR 2012: 35th international ACM SIGIR conference on research and development in information retrieval*, Portland: ACM, 2012.

24 Bouke Huurnink, Amit Bronner, Marc Bron, Jasmijn Van Gorp, Bart de Goede, and Justin van Wees, '**AVResearcher: Exploring Audiovisual Metadata**', *DIR2013: Dutch-Belgian Information Retrieval Conference*, 2013.

25 For accounts on media studies research with 'standard' search tools, see Marc Bron, Jasmijn Van Gorp, Frank F. Nack, Maarten de Rijke, '**Exploratory Search in an Audio-Visual Archive: Evaluating a Professional Search Tool for Non-Professional Users**', *EuroHCIR 2011: 1st European Workshop on Human-Computer Interaction and Information Retrieval*, Newcastle, 2011 and Jasmijn Van Gorp, '**Looking for what you are looking for: a media researcher's first search in a television archive**', *VIEW: Journal of European Television History and Culture, 1*(3), 2013.

26 Exploration can be considered as the first phase in Media Studies research, followed by contextualization and presentation. See Mark Bron, Jasmijn Van Gorp and Maarten de Rijke, 'Media studies research in the data-driven age: How research questions evolve,' *Journal of the American Society for Information Science and Technology*, 66 (12), 2015, **DOI: 10.1002/asi.23458**.

facilitate media researchers in exploring the Dutch public television archive. In *AVResearcher* subtitles and tweets about television programmes have been added.

The third prototype, *AVResearcherXL*, was launched in late 2014 and incorporates new collections and functionalities. It contains the metadata of programmes of all Dutch public broadcasters (title, date, genre, broadcaster, people, etc.) from the 1920s for radio and the 1950s for television, up until 26 October 2013. Annotators provided parts of these metadata, such as general tags (keywords) and a summary. For some programmes, they also described what they saw and heard in an elaborate programme description, even indicating time slots. Since 2012, subtitles for the deaf and hearing impaired were added to the metadata descriptions of all television programmes, and speech recognition files were (sparsely) added to the metadata descriptions of radio and television broadcasts. As for the newspapers, the tool searches across metadata (newspaper and article titles, date, type of article) and the full transcripts of the **OCR'd scans** of the newspaper articles as provided by the Dutch Royal Library, from 1 January 1900 to 30 November 1994. To date, the tool searches across the metadata of about 932.035 public radio and television broadcasts (of which 18.124 have subtitles) and the transcripts of 25 million newspaper articles.

The front-end of the interface is double-sided (see Figure 1). It consists of two identical search boxes, both containing options to search within television/radio or newspaper databases, and two time sliders to adjust the time range. The comparison between two search terms in *AVResearcherXL* can be done in terms of time (by means of a timeline), related words (by means of word clouds and bar charts) and snippets of individual programmes/newspaper articles (by means of a result list). When clicking on words in the word clouds, search terms are added to the search box, using an AND-boolean operator.

AVResearcherXL uses the indexes at the back-end to calculate frequencies of words, which are related to the user's search terms. It plots the frequencies in the above mentioned data visualizations, e.g. the bar charts, word clouds and timeline. Each document that contains the search terms stands for one hit, regardless of the number of times the search terms occur in the document. In this respect, *AVResearcherXL* differs from the well-known Google Books **NGRAM-viewer**, which plots the frequency of all hits within each document.[27] Next, the tool provides a ranked result list of matching documents, which also appears when clicking on a data point on the timeline. To get to know the individual programme, the user needs to click on a title in the result list, which leads him/her to the external catalogue of the Netherlands Institute for Sound and Vision: **in.beeldengeluid.nl**. And, when available, the metadata description also contains a link to the educational website **academia.nl** where the video of the television broadcast can be viewed. When clicking on a title in the result list of the newspapers, the user is lead to the newspaper archive **delpher.nl**, which contains full and browsable OCR'd scans of the newspapers. An elaborate user manual is available **here** (registration required).

In summary, as Figure 2 shows, different material representational processes make up the different 'strata' or layers of *AVResearcherXL*: (1) the *front-end* of the interface with search boxes, time sliders, bar charts, word clouds, a timeline, result lists, and the user manual, (2) indexes at the *back-end* consisting of metadata, subtitles, speech recognition files, transcripts of OCR'd newspapers, (3) linked individual document descriptions of radio and television broadcasts, newspaper scans at *other websites and portals*, which eventually also leads to video websites such as academia.nl, and the source code on Github, and (4) the *broadcasts and newspapers* which are annotated, and OCR'd. These representations and the relation between these representations have to be considered when using *AVResearcherXL*, as we show in the next section.

27 For those who are interested in the technical workings of the tool, the source code is available on Github under an open source license: **https://github.com/beeldengeluid/AVResearcherXL**.

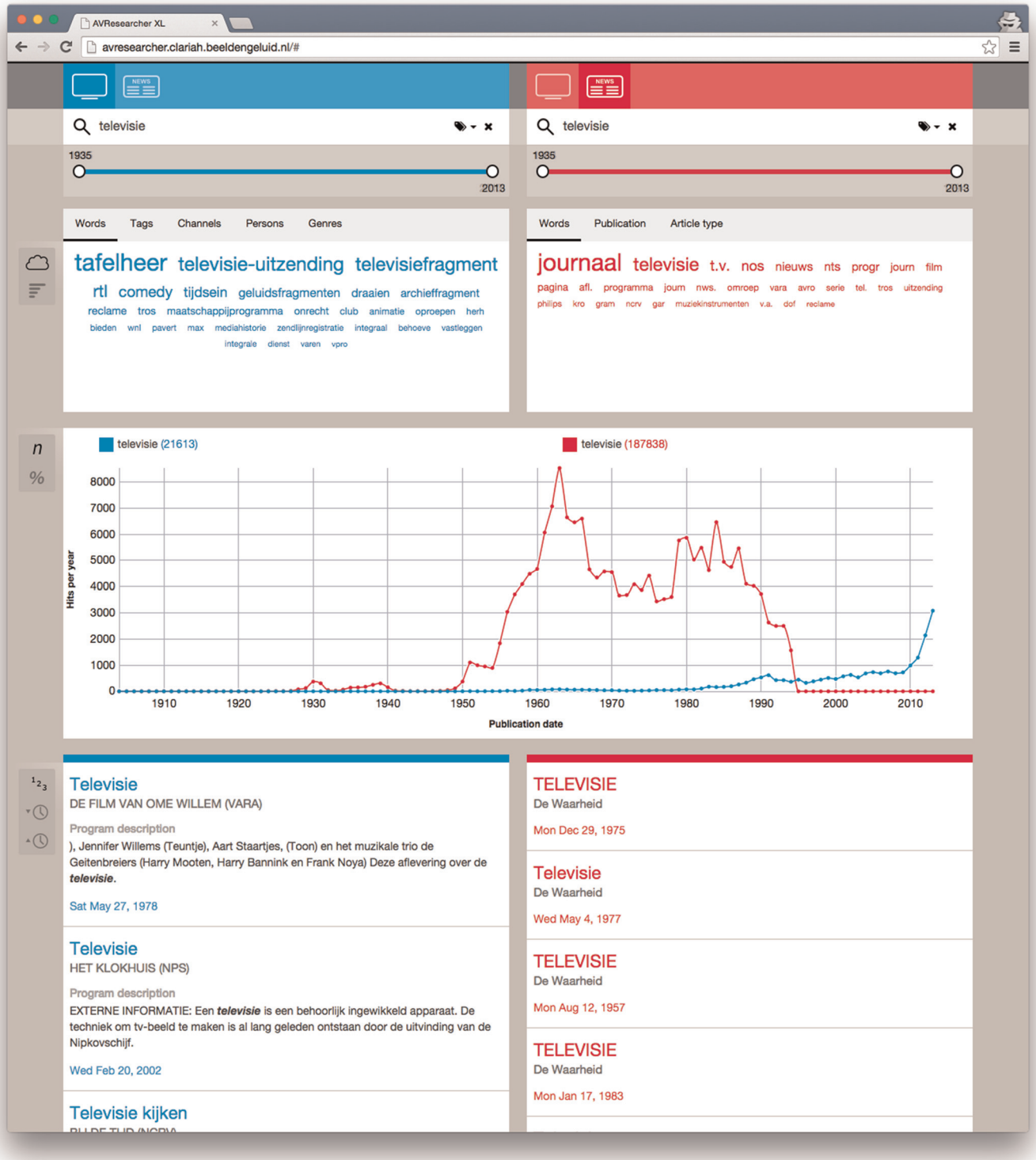


Fig. 1 Front-end of *AVResearcherXL*: search box, time slider, word clouds, time line and result lists.

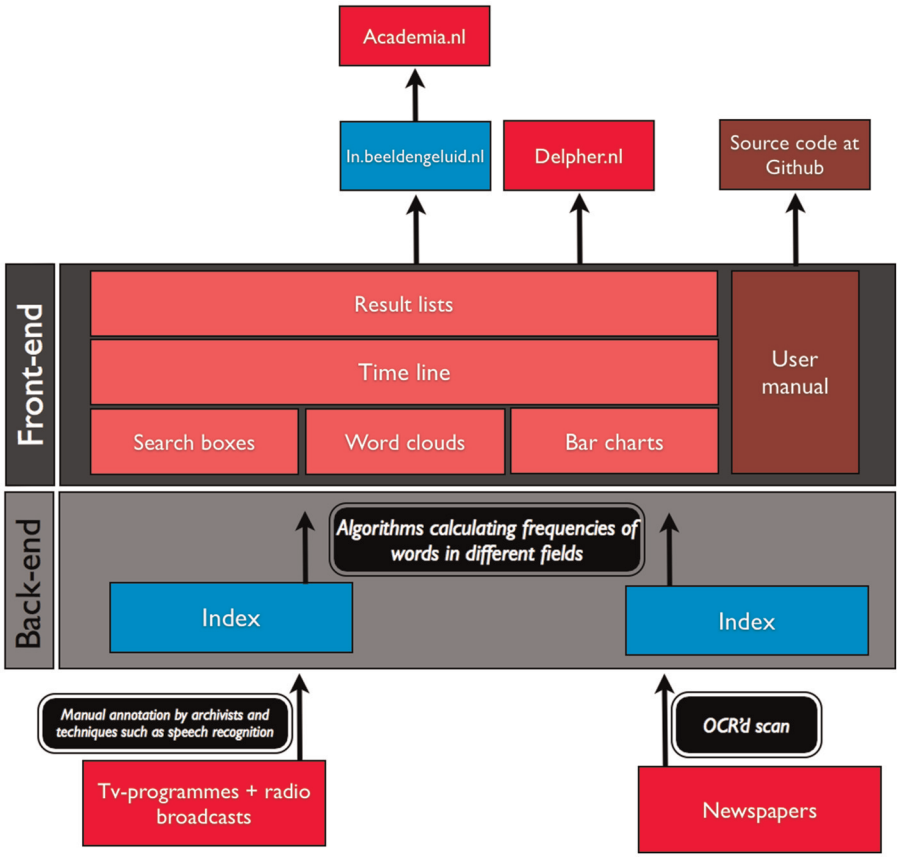


Fig. 2 Different material representations and layers of *AVResearcherXL*.

3 Deconstructing *AVResearcherXL*

3.1 The Timeline Paradox

We start our exploration by typing the word 'television' in the left search box, selecting the radio and television archive, and typing the same word in the right search box, selecting the newspaper archive. The central element in the front-end of the tool, which immediately catches our attention, is the timeline. *AVResearcherXL*'s timeline is evidently linear: it plots the frequency of the words over time. This linearity contradicts the core idea of Media Archaeology of 'reading against the grain.' The latter draws on Foucault's archaeology as a method of historical analysis and on Zielinski's critique of chronology as the dominant time mode. Foucault emphasizes rupture and discontinuity, which he discusses in terms of threshold, break, mutation, and transformation.[28] Yet, tracing the word 'television' with *AVResearcherXL*, the result list shows how, paradoxically, the timeline should not be considered as successive or temporal, but rather as indicative, leading to new queries, searches and questions.

When comparing the timeline of 'television' in the television and radio archive with the newspaper archive (see Figure 3), we observe two peaks in the timeline of the newspaper archives: one around 1960 and one around 1990. The first peak is built up from 1953 (the launch of television in the Netherlands) with a peak in 1960 showing 4672 hits. Closer inspection tells us that the newspaper *De Waarheid* starts publishing the broadcast schedules in 1953 in the

28 Michel Foucault, *The Archaeology of Knowledge and the Discourse on Language*, Pantheon Books, 1972, p. 4–5; Siegfried Zielinski, *Deep Time of the Media. Toward an Archaeology of Hearing and Seeing by Technical Means*, MIT Press, 2006. The very title of Zielinski's book emphasizes the importance of time and temporality.

section '**About both channels**' ('Over beide zenders'), while another newspaper *De Telegraaf* does it from 1955 onwards in the section '**Programmes of domestic and international channels**.' This is an interesting difference in titles and in the publishing of broadcast schedules in newspapers in the 1950s, which leads us to raise the question of why the one newspaper already starts in 1953 and the other only in 1955. *AVResearcherXL* triggers this research question, but answers to this question can only be found when digging further into all articles in the 1950s, and by looking for contextual information on other websites. Such result underscores the exploratory character of the tool, raising new questions, which evidently cannot be answered without further investigating and contextualizing its material representations.

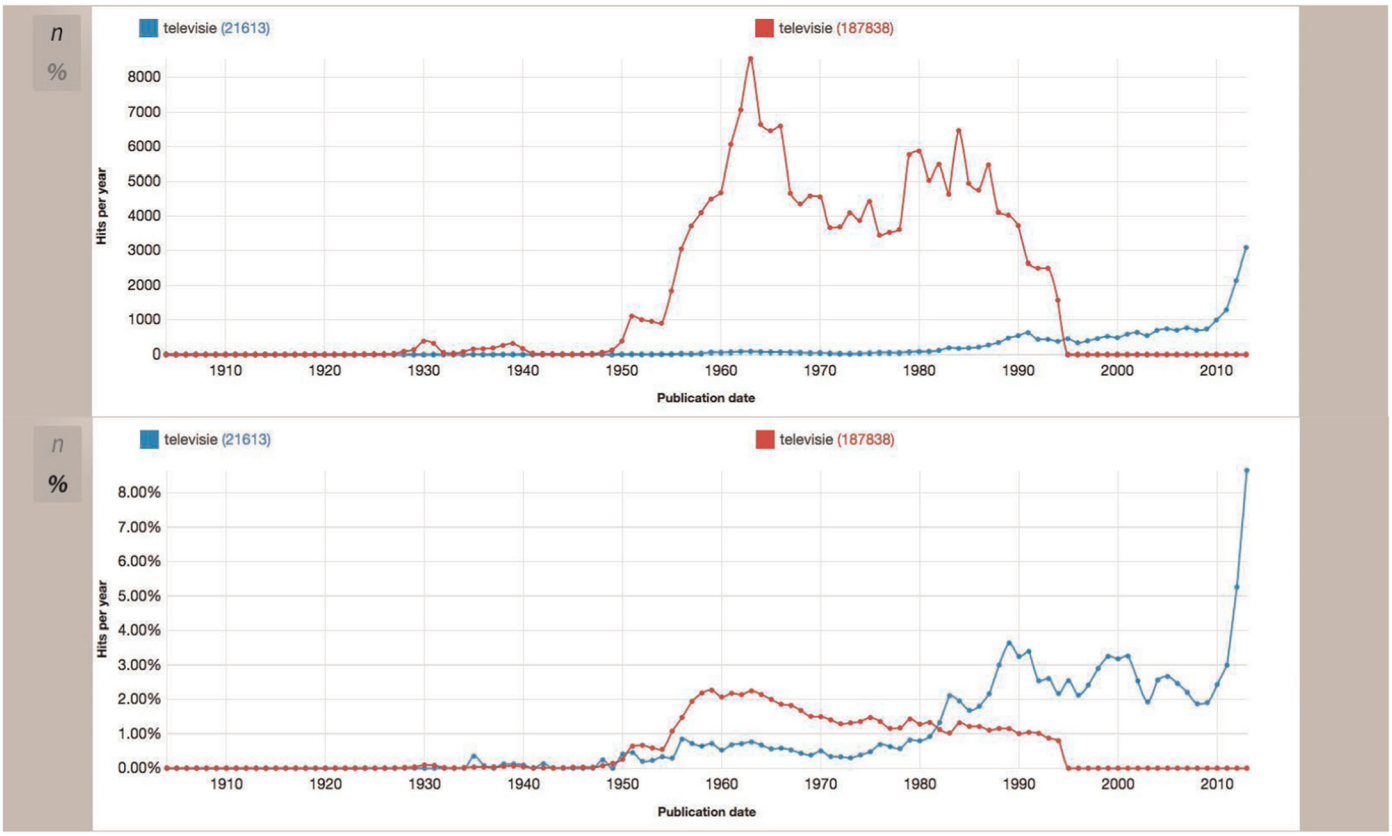
The broadcast schedules are not the only television-related items in the newspapers. A few examples indicate what one could discover and how discoveries could be valued. *De Telegraaf* starts to write extensively about television from 1955 onwards, for instance in the article '**This kind of difficult beautiful work for television**' ('Dit moeilijke mooie werk voor televisie') - followed by the subheading 'A lot will happen' - about the anxieties of the arrival of the new medium of television in the Netherlands. The first data point on the timeline directs us to *De Telegraaf* in 1926: **Baird's first experiments with television** ('Baird's proeven met televisie') in the section 'Radio World' ('Radio Wereld'). We also notice a hit in 1904, which is remarkable, but it appears to be an OCR mistake: 'War about Television' ('Televisie oorlog') says the transcript, but the **OCR'd scan of the newspaper article** at the Delpher website shows that it is titled 'War about Tariffs' ('Tarieven oorlog'). OCR mistakes are blurring the data, and show how important it is to go and check particularities in the timeline.

The timeline is also depending on the collections, the amount of items in the collection and the composition of the collection. From 1995 onwards, the timeline flattens for the newspaper archive because there are currently no recent newspapers in the Dutch Royal Library archive. In 2010, the timeline increases dramatically, because the subtitles are connected to the broadcasts from this year onwards. Broadcasts enriched with subtitles have a higher chance of appearing in the search results because each word that was spoken during the broadcast is also considered. Having knowledge on the composition of the collections, therefore, is necessary to interpret the data visualizations. One way to get a better insight in the composition is normalizing the timelines, as it enables us to compare the relative size of collections and selections made. Relative frequencies are calculated by dividing the number of hits for each year/moment by the total amount of documents of each year/moment.

If we normalize the timeline in Figure 4, e.g. visualize the relative instead of the absolute frequencies, we notice a totally different pattern compared with the timelines in Figure 3. The newspaper timeline is flattened. In the television and radio archive, there is now a large peak in 1989, which is almost 4% of all broadcasts, a quite large percentage. When we look at the result list, we notice that the 1989 peak is caused by '**Integral recording of Dutch Television for Media Historical Purposes**,' which started according to *AVResearcherXL* on May 19, 1987. Apparently, historians in 1986 wondered how they could have a representative image of Dutch television in the future if only individual broadcasts would be archived. To address this question, the **Foundation for Film and Research** at the time started recording integral broadcasts of two full weeks a year.[29] Our search therefore, leads us to a precious object for television historical research and research on scheduling: historical recordings of full days of broadcasts, including announcers, commercial breaks, and technical interruptions.

Instead of contradicting the core idea of media archaeology, the timeline in *AVResearcherXL* underlines the importance of contextualization and source criticism. While timelines are explicitly putting forward causation and succession, our exploratory study shows that these trends should not be taken at face value. The timelines in themselves can be tweaked: changing the time period, adding or removing search terms, generating absolute versus relative counts, or including subtitles or not. The comparison between the two timelines can point to gaps and errors in the composition of the collections, which helps grasping the particularities of the specific collection at hand. It is in this combination of representations that the tool can be used. Each action, each variation renders multiple visualizations and readings, contributing to a better understanding of the working, composition and construction of the digital archive.

29 See Dutch blogpost at **http://www.beeldengeluid.nl/blogs/collecties/201210/unesco-werelddag-voor-audiovisueel-erfgoed** and English language version at **http://www.iasa-web.org/netherlands-institute-sound-and-vision-holds-its-bi-annual-week-dutch-television**.



Figs. 3 and 4 Number of hits for 'television' in metadata descriptions of television/radio programmes (blue) and in newspaper articles (red) visualized on a timeline in absolute counts (see Figure 3, top) and percentages (see Figure 4, bottom).

3.2 Combining Searchlights: Words versus Tags

We continue our (re)search by looking at the other data visualizations: the bar charts or histograms and word clouds. Bar charts and word clouds visualize the words, which occur most frequently together with the search term in the same document, also known as 'related terms.' They can be faceted by words, tags, channel, people and genre (for radio and television) and words, publication and type (for newspapers).

First, we compare the bar charts for 'television' in terms of words and in terms of tags, in the television and radio collections. 'Words' in this case refers to the 'most descriptive words' in the descriptions and subtitles as calculated by an algorithm. This algorithm selects the most unique words for each document by comparison with all the words in all other documents. When we would not use this algorithm, all word clouds would only display common words such as 'the,' 'a,' 'is,' 'in,' 'for' and 'are,' as these are often the most frequent words in a text. The 'most descriptive word' algorithm manages to highlight words that are descriptive for the texts, such as names, sentiments, places etc., hence better pointing to the content of the programme/newspaper article. 'Tags' are the keywords attached to the television and radio programmes by (mainly) archivists. This shows that it is necessary to dig further into the tool, and especially to go to the individual metadata descriptions at **in.beeldengeluid.nl** to make sense of the data visualizations.

If we compare 'television' in words and tags for the full period (see Figure 5), we see that the related 'words' differ completely from the 'tags,' indicating that it makes a big difference searching by words in descriptions versus searching by tags provided by annotators. The 'words' are all television-related, such as 'TV host' ('tafelheer'), 'broadcast' ('televisie-uitzending'), 'clip' ('televisie-fragment'), 'comedy' ('komedie') and 'shooting a TV programme' ('draaien'). The tags, on the other hand, only show two television-related words, namely 'television' ('televisie') and 'television

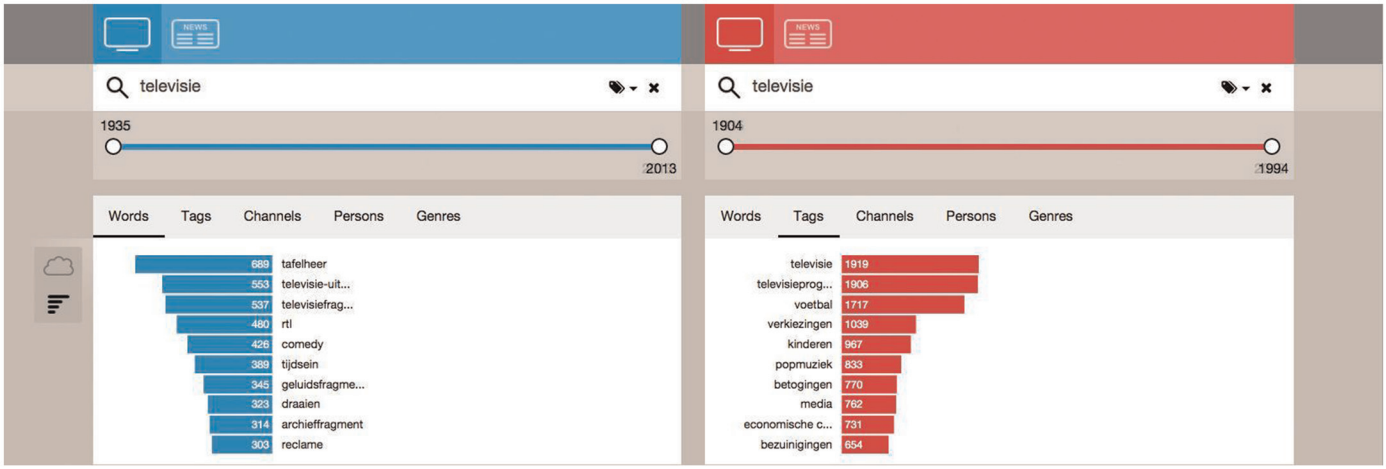


Fig. 5 Bar charts of words related to 'television' in metadata descriptions [blue] and tags provided by annotators [red] of television and radio broadcasts.

programme' ('televisie programma'), and for the rest they shed light on the themes of the programmes: 'soccer' ('voetbal'), 'elections' ('verkiezingen'), 'pop music' ('pop muziek'), 'children' ('kinderen'), 'media', 'protests' ('betogingen'), 'jubilees' ('jubilea') and 'politics' ('politiek'). In other words, the tags provide an overview of the main topics on television in the archive, mostly related to news events, but also to genres such as sports programmes (soccer) and children's television. Tags provide an insight into the vocabulary of documentalists, which are also constructs and evolving throughout time. By contrasting words and tags, it becomes visible to what extent annotations of documentalists define search results. This illustrates the relevance of knowledge about the construction of the archive, particularly about the provenance of metadata. The data visualizations provide different 'slices of' and 'searchlights on' the metadata, thus contributing to a "variantology of the media, leading away from the obligatory trends."[30] Individual documents (close reading) are as important as macro-views (distant reading). It is in comparison and contrast that material representations become meaningful.

3.3 Digging Deeper: Streamed Materials

Now, we limit our search to the recent years 2012 and 2013, the years in which subtitles were structurally added to the archive. For the metadata descriptions of broadcasts in 2012 and 2013, we witness a similar trend for the full period: it contains primarily television-related words such as 'archival clip' ('archief fragment'). Interestingly, it also displays four words which do not immediately ring a bell: 'pavert,' 'phone call' ('opbellen'), 'max' and 'in the past' ('vroeger'). We decide to select these words one by one in the word clouds, in order to gain more understanding of their context. The three words 'max,' 'phone call' and 'past' appear to be in the same television programme: ***Tijd voor Max***. If we look at the **metadata description at in.beeldengeluid.nl**, this programme turns out to have fixed television features in its shows, among which the feature of 'a friend or family member making an unexpected phone call to the guest.' Furthermore, it contains the feature 'Television from the film canister: reviving remarkable television moments from the past' ('Televisie uit blik: herleven van opmerkelijke televisie-momenten van toen') presented by Koos Postema from a viewing room of the Netherlands Institute for Sound and Vision.

The broadcast itself can be watched on **academia.nl with institutional access**, and also on the free portal **Uitzending Gemist** (access it **here**). The example shows that one fixed feature about television in the description causes the appearance of all kinds of non-related words of the other features mentioned in the same description. This can be considered as a 'failure' (the word 'phone' appearing), but it is actually also pointing to the importance of layers within digital archives as archaeological sites: a data visualization (including a statistical-like bar chart with the word 'phone')

30 Siegfried Zielinski, *Deep Time of the Media. Toward an Archaeology of Hearing and Seeing by Technical Means*, MIT Press, 2006, p. 7.

brings us to a metadata-description on another portal (*Tijd voor Max* at **in.beeldengeluid.nl**), which is then connected to an audiovisual broadcast about virtual media objects (film canisters at **academia.nl**). Again, this example illustrates the workings and construction of the archive, particularly regarding the material representations of television.

Fig. 6 Kees Postema presents archival material in *Tijd voor Max: Television from the film canister* (MAX, 28.11.2012). Watch the full video **here**.
(Source: **npo.nl/uitzending-gemist**)

Digging into the different layers of the digital archive opens the possibility of exploring its 'geological strata'[31] further and of discovering new discursive constructions (not just material representations, but also narratives) related to television, such as one about television and immigration, as demonstrated by the clip of *Tijd voor Max*: *Television from the film canister*. The very subject of the film roll is a story of immigration after the Second World War from the Netherlands to Australia, jumping from 1948 to 1957. The canister thus contains images from different time periods, with the help of which the program's presenter (in 2012) narrates a televisual story about immigration. Incorporating historical knowledge about the real experiences of immigrants, his narrative contrasts the images from the canister and addresses the question of the reliability of visual images. Another discursive construction that might be further investigated surfaces around television and nostalgia as the presenter of the programme walks around in the archive carrying a canister, not only literally finding the past, yet also mediating with authority between past and present. This visual representation of the archive implicitly addresses television's role in constructing historical knowledge. The clip thus illustrates how the tool generates discoveries, which the researcher could interpret further.

When digging deeper into the tool up to the layer of streamed material, other glimpses of the subject are offered. This underlines that *AVResearcherXL* is not one interface that can be considered in isolation, but rather one 'node'[32] in a web of interfaces, relying heavily on other interfaces such as **in.beeldengeluid.nl**, **delpher.nl** and **academia.nl**.

31 Michael Goddard, 'Opening up the Black Boxes: media archaeology, 'anarchaeology' and media materiality,' *New Media & Society*, April 2014, p. 2.

32 Jussi Parikka, 'Operative Media Archaeology: Wolfgang Ernst's Materialist Media Diagrammatics', *Theory, Culture and Society*, vol. 28 (5), 2011, p. 64.

3.4 Subtitles as Rematerialized Broadcasts

Finally, we look at words appearing in subtitles for the hearing impaired. As radio broadcasts are not accompanied by manually created subtitles, we filter out the radio broadcasts by typing *television -radio* in the search box.[33] Television broadcasts are returned as a result.


```
{
  "date": "2013-06-28T18:45:59",
  "source": "http://zoeken.beeldengeluid.nl/internet/index.aspx?chapterid=1164&contentid=7&verityID=_16502_16542_16542_4153992@expressies",
  "title": "SPORTJOURNAAL",
  "meta": {
    "roles": [
      {
        "playerName": "Peperstraten, Toine van",
        "key": "maker",
        "value": "Peperstraten, Toine van",
        "playerFunction": "presentatie"
      },
      {
        "playerName": "NOS",
        "key": "producent",
        "value": "NOS",
        "playerFunction": null
      },
      ...
    ],
    "categories": [
      {
        "key": "genre",
        "value": "sportprogramma"
      },
      {
        "key": "person",
        "value": "Armstrong, Lance"
      },
      ...
    ],
    "descriptions": [
      "De Amerikaan, die in de periode 1999 tot 2005 zevenmaal de Franse ronde won maar uit de uitslagen...",
      "Nederland speelt twee oefenduels in Nederland, tegen Australië en Noord-Ierland, alvorens af...",
      ...
    ],
    "subtitles": "... middelen heeftgebruikt. De Franse televisie vroeg oud-Tourwinnaar Bernard Hinault...
        die lid is van het organisatiecomite van de Tour de France, om een reactie.",
    "subtitles_descriptive_terms": [
      "monde",
      "lorenzo",
      "dopinggebruik",
      "yamaha-rijder",
      ...
    ]
  }
}
```


Fig. 7 An example of data format of a single document as indexed by *AVResearcherXL*. The example is slightly simplified and shortened […] to enhance readability. The full document can be found **here**.

33 Please note that this search string also filters all documents, which contain both words. The tool does not have a facet for radio yet.

The result list shows that people talked a lot about television in television programmes in 2012 and 2013 (1519 hits): 'on television,' 'about television,' 'known from television' etcetera, are phrases in the subtitles as displayed in the snippets in the result lists. If we look at the related genres, the word 'television' is mostly mentioned in news (417 hits) and talk shows (275 hits). If we look at the related persons, Princess Beatrix of the Netherlands ranks first with 37 hits, followed by King Willem-Alexander with 32 hits and Queen Maxima with 29 hits. If we look at the result lists, it immediately becomes clear why: the change of the throne in the Netherlands in April 2013 was highly mediatized. This is not unexpected, and coincides with the key role of royal events in the history of television. This example conversely points to the role of television in mediating events and eventually notions of (cultural) identity to be further investigated by the researcher. At another point the example again illustrates the workings and construction of the archive and consequently its limits and possibilities.

Interestingly, the word 'live' occurs often in combination with the word 'television' in the subtitles. If we look at the result list, we do not see both words in the snippet. Nor can we see more context for the phrase, as the **metadata description at in.beeldengeluid.nl** does not display the full subtitle-files due to copyright restrictions. If we look at the index (hidden for users) at the back-end (see Figure 6 for an example), it turns out that the connection between 'television' and 'live' is mainly caused by the phrase 'This programme is live subtitled' at the very end of the subtitle-file, which is not said but shown on television. An additional Google search teaches us that there is a new technology using speech recognition to automatically generate subtitles during live broadcasts (see Video 1). Again, this discovery points to the workings of the archive: our search teaches us something about the history of the archiving of subtitles and new technologies. However, as the rather low frequencies show, not all television programmes are enriched with subtitle-files yet.

The subtitles and speech recognition files underline the importance of royal and sports events on television, but it is difficult to really draw conclusions and dig deeper into the subtitles because only a small percentage of the television and radio broadcasts are enriched with subtitles and speech-recognition files.[34] Moreover, subtitles and speech-recognition files are not visible for the user at **in.beeldengeluid.nl**, which prohibits the television historian from conducting a close reading of the context, an important step to interpret the data-driven visualizations rendered by the tool. The missing information, as the example of the subtitles shows, is actually pointing to a very important feature of the tool. The tool visualizes frequencies of words occurring in the indexes at the back-end of the tool, but these indexes do not necessarily match what is available as separate publicly available documents on the different portals.

Fig. 8 Screenshot from the information video about the project NEON that conducted research on ways to make subtitling less labour-intensive. Watch the full video **here**.

34 According to *AVResearcherXL* it is roughly 4,5% for 2012–2013. The actual percentage for 2013 is higher: 25%. In 2013, 13.655 out of 52.823 ingested video files had subtitles.

The subtitles are rematerialized sound files of television programmes, and also contain additional information of phrases shown on television. Subtitles are valuable as they offer a different representation (e.g. what has been said on television). The user can compare this representation with the description as provided by annotators and the streamed broadcasts on other portals. The act of interpretation, however, is complicated when data is situated for users within 'hidden' layers, such as the index. This points to the importance of accessible and well-documented 'strata' of interfaces. The user manual, which is available at **the front-end of *AVResearcherXL*** is a helpful instrument to gain more insight in the different strata, as in the source code at Github.

4 Conclusion: Towards a Digital Media Archaeology?

By deconstructing *AVResearcherXL*, we provided a practical use case of doing media archaeology with digital tools for television, radio and newspaper archives. Our deconstruction shows the importance of the media archaeological approach for looking into the *materiality* of digital technology (the components as discussed above) as well as the relevance of studying the deep material structure (the minerals as to be found in layers as discussed above) of media technology.[35] *AVResearcherXL* thus, could be seen as an archaeological site in which the user or 'archaeologist' decides where to dig and which search lights to use. Depending on what searchlights are used and strata are visited s/he comes across different results.

The tool enables us to shed different searchlights on material representations of the word 'television' in television and radio programmes and newspapers, but none of these lights are straightforward and can be taken for granted. In this article, we argued that it is necessary to combine different representations, different visualizations to use the tool in a meaningful way. Every action provides another perspective, thus enabling the user to construct alternative television histories and individual variations. The tool, therefore, provides a way to deconstruct the archive, to dig deeper into trends, to discover new objects, and – importantly – to raise additional research questions. As such, *AVResearcherXL* is what Walter Benjamin describes as a database in his *Passagen-Werk*: bringing different collections and aspects of collections together without presenting a pre-organized narrative.[36] To interpret these relations of representations or re-materializations, users may apply several strategies. What seems to be a precondition for working with the tool in a meaningful way is to rethink its materiality by getting into its black box and to understand the particular material nature of it.

Our (re)search also shows the limits and ambiguities of the tool. As Ernst argues, tools themselves can also become active 'archaeologists' of knowledge.[37] They are programmed to help us, to support our research, but they also pre-define what we can find. Only 1.9% of the broadcasts have subtitles, and the tool only searches across six newspaper titles up to 1995, containing exclusively metadata of the Dutch Public Broadcasters. As such, the tool in itself only provides one slice of historical knowledge. Also, some issues remain covered and hidden. For instance, we do not know exactly what information is missing, which procedures were used by the documentalists to annotate the television and radio programmes throughout the decades of archiving, and how many OCR'd scans of newspapers contain mistakes. Moreover, what proved to be a difficult matter is that the indexes are not one-to-one matching the available documents on public portals. The indexes contain more information, such as subtitles, while the user has only access to the annotated descriptions at **in.beeldengeluid.nl**. In other words, it would be helpful to know how the metadata is compiled (e.g. by conducting interviews with documentalists), to get full access to indexes and datasets, to have techniques to improve the reliability of OCR'd scans and to incorporate feedback of users directly into the tool.

35 Jussi Parikka, *What is Media Archaeology?*, Polity Press, 2012, p. 15. See also his blog ***Machinology***. And for media archaeology as ecology, see Goddard, p. 2.

36 See Walter Benjamin's unfinished *Das Passagen-Werk* (1927–1940). For Benjamin collecting a huge database consisting of a great variety of sources was to be able to presenting new ways of looking at 19th century history.

37 Wolfgang Ernst, 'Media Archaeography: Method and Machine versus History and Narrative of Media', in Erkki Huhtamo and Jussi Parikka, eds, *Media archaeology: Approaches, applications, and implications*, University of California Press, 2011, p. 239.

The methods within the 'traditional' and the digital humanities, as such, remain the same: by comparing and contrasting, one of the main techniques of the television historian, we are able to interpret the results. Yet, the question remains: what is 'new' about digital tools? As we have shown, using the tool raised many other questions related to television history research. The more search actions we conducted, the more additional questions and uncertainties arose. For instance, why exactly do sports programmes pop up together with television in the newspaper archive of the 1980s? How and to what extent is the treasure of 'integral broadcasts of Dutch television' used in scholarly research and what additional information do these integral broadcasts provide in contrast standard, singular recordings? Why does one newspaper only publish the schedule of domestic channels, while another newspaper provides both domestic and international broadcast schedules in the 1950s? What exactly is the relation between the word 'live' in subtitles in relation to royal events and Dutch television history in general?

Using *AVResearcherXL* to do media (historical) research is not about finding the 'right' answers, but about contextualising results, about finding new, and sometimes unexpected pathways. Plus it is about raising even more questions, which might be solved with other tools and methods. *AVResearcherXL*, therefore, is not an endpoint, but a starting point for new discoveries in the archive, based on which, researchers could define a (new) historical question. It also illustrates how important it is to put uncertainty and ambiguity to the fore. "It could have been otherwise" as Elsaesser states, attributing Noël Burch.[38] Every search leads to a new investigation and new research questions, for which *AVResearcherXL* provides one possible way of support. We might even go so far as to postulate that in order to become meaningful for media (historical) inquiry, digital humanities as a discipline needs media archaeology: getting into the black box, uncovering the material structure of the digital tools that have been constructed and are being used to explore the digital archive. Only then, so it seems, looking for alternative histories or reading against the grain of the past (the dominant historical narrative) - one main objective of Media Archaeology - becomes a viable possibility.

The tool will be further developed, the data will be updated, and – hopefully – also new collections will be connected. After publication of this article, for instance, a new 'live' version of *AVResearcherXL* will be available, in which the metadata of the broadcasts will feed daily into the indexes. If *AVResearcherXL* is connected live to the catalogue, the visualizations will change on a daily basis, leading to even more 'variantologies'.

AVResearcherXL is available **here**. *Registration with a university e-mail address is required. AVResearcherXL is financially supported by CLARIN-NL and CLARIAH-SEED.*

Biography

Jasmijn Van Gorp is Assistant Professor in television studies at Utrecht University. She received her PhD in Social Sciences from Antwerp University (2008) and has been a visiting scholar at the Russian Film Institute in Moscow and at the Comparative Media Studies program at MIT. She is specialized in the development and testing of digital tools for media archives and is the project leader of *AVResearcherXL*.

Sonja de Leeuw is Professor at the Department of Media and Culture Studies at Utrecht University. Her research and teaching interests are: Dutch television culture in an international context (history and theory genres and productions practices) and media and cultural diversity (diasporic media, representation of ethnicity). She published on television culture in the broadest sense, on diasporic media and on children's media. Sonja de Leeuw participated in the EU funded research project *CHICAM, Children in Communication about Migration* (2001–2004) and coordinated the EU funded projects *Video Active, Creating Access to Europe's Television Heritage* (2006–2009) and *EUscreen, Exploring Europe's Television Heritage in Changing Contexts* (October 2009–2012). She is also co-leader of a research project *The Power of Satire: Cultural Boundaries Contested*. She co-founded and coordinates the European Television History Network (with dr. A. Fickers, University of Luxembourg). She participated in the project AVResearcherXL as a humanities partner and user.

38 Thomas Elsaesser, 'The New Film History as Media Archaeology', *CINéMAS*, 14(2–3), 2004, p. 71–117, p. 81.

Justin van Wees is co-founder of the Dutch software company Dispectu, which specializes in development of search systems and interfaces that provide access to large datasets. He regularly participates in academic projects to develop tools such as the *AVResearcherXL*.

Bouke Huurnink is Development Manager at the Netherlands Institute for Sound and Vision. He received a PhD in Information Retrieval with a dissertation on *Search in Audiovisual Archives* (University of Amsterdam, 2010). He participated in *AVResearcherXL* as a cultural heritage partner.

VIEW Journal of European Television History and Culture Vol. 4, 7, 2015
DOI: 10.18146/2213-0969.2015.jethc080
Publisher: Netherlands Institute for Sound and Vision in collaboration with Utrecht University, University of Luxembourg and Royal Holloway University of London.