TS •> Tijdschrift voor tijdschriftstudies
ISSN 1386-5870

Publisher: Igitur Publishing. Website: www.tijdschriftstudies.nl

URN:NBN:NL:UI:10-1-114173. TS •> # 36, December 2014, p. 139-146.

TS Tools: Problems and possibilities of digital newspaper and periodical archives

THOMAS SMITS

t.smits@let.ru.nl

ABSTRACT

Digital newspaper archives are the most widely used resource for digital humanities research. The basic functionality of these archives – the ability to perform keyword searches across multiple titles – is unlikely to change. This means that researchers can safely develop a digital methodology without having to fear that new technology will make their efforts obsolete. On the basis of this observation, this article discusses three methodological problems of the digital newspaper archive. First, it provides unequal access to both different scholars and historical periods. Second, it tends to ignore the historical form of media landscapes. Third, the digital mediation of newspapers and periodicals results in a problematic loss of context. This article proposes that the solution to these problems lies in the assertion of agency over the digital newspaper archive and in the debates surrounding them by researchers. They should see digital newspaper archives as any other paper archive: with its own possibilities and limitations, which can be adjusted by the researcher.

KEY WORDS

Digital newspaper archives, digital methodology, Trove, Delpher, PaperPast

INTRODUCTION

'The rapid digitisation of newspapers and periodicals has transformed even the recent past into a foreign country.'[1] With this observation Bob Nicholson starts his article 'The Digital Turn' (2013) about the transformation of historical research since 2002. We are currently experiencing a shift in focus in the humanities enabled by a 'practical revolution': the advent of searchable digital newspaper and periodical archives. Nicholson notes that 'one of the core features of digital newspaper archives – the ability

[1] B. Nicholson, 'The Digital Turn.' *Media History* 19.1, 2013, 59-73 (59).

to perform basic keyword searches across multiple titles – is unlikely to change.'[2] This means that methodologies that are developed for it today will not be rendered obsolete by the technology of tomorrow.

This last assertion of Nicholson is the basis on which this article discusses three fundamental problems of the digital newspaper archive. First, it provides unequal access to both different kind of scholars and to different historical periods. Second, it tends to ignore the shape of historical media landscapes. Therefore, digital newspaper archives are, mainly as a result of their institutional benefactors, often biased towards the present form of the nation state and the publications of majority groups. Finally, digital newspaper archives provide a new point of access to information that is stored in historical newspapers and periodicals. Digital mediation results in a significant lack of context and, as I argue, it is precisely this context which has to be reinserted. I came across these problems during my own research in the digital newspapers archives of the Netherlands (Delpher), Australia (Trove) and New Zealand (PaperPast) concerning the reception of *Illustrated London News,* a renowned illustrated newspaper from the nineteenth century.

Discussions concerning digital newspaper archives are often centered on the question what digitization does, or will be doing, to research in the humanities. Proponents like Rens Bod present digital humanities – the move to what he calls 'Humanities 3.0' – as a direct result of technological innovation.[3] Nicholson laments this presentation of technology as an 'uncontrollable force', which is mainly used to rob scholars of agency in the debate. The primacy of technology results in the fact that the methodological implications of the use of digital newspapers archives have 'remained frustratingly absent from the debate.'[4] To avoid the pitfall of only raising questions without giving answers, this article aspires to present three possible solutions to the problems it discusses: all based on a reassertion of agency by researchers. It hopes to contribute to a sound methodological basis on which a turn to the digital can be safely made.

UNEQUAL ACCESS

Every archive, be it digital or not, provides unequal access on many levels. Before starting the section, it is important to note that digital newspaper archives in general have tremendously improved access to the historical sources they contain. These sources are no longer bound to a physical space and it is much easier to navigate their contents. However, this access is far from universal, as it is often presented and perceived. This section discusses two forms of unequally distributed access, which are problematic for social and methodological reasons.

In his influential article 'The Digitization of Newspaper Archives' (2010) Adrian Bingham makes an interesting observation: digitization has provided historians with access to a wide range of sources – 'or rather, they do if they are members of institutions

[2] *Ibid.* (62).
[3] R. Bod, 'Who's Afraid of Patterns? The Particular versus the Universal and the Meaning of Humanities 3.0.' *BMGN-Low Countries Historical Review* 128.4, 2013, 171-180.
[4] Nicholson 2013 (62).

willing to buy subscriptions, or if they are willing to pay on a individual basis.'[5] Bingham does not elaborate this point, but his side-note has won in importance over the last four years. While some digital newspaper archives, like Delpher (The Netherlands), Trove (Australia) and PaperPast (New Zealand), remain open-access, others like the British Newspaper Archive started to charge both institutions and individual users. Commercial publishers such as Gale Engage offer expensive institutional subscriptions to famous publications like the *Times*, the *Illustrated London News* and *The Economist*. Publications from the eighteenth and nineteenth century, which were in the public domain due to the lapse of copyright, have been reintroduced to the market as digital products.

Researchers outside Academia, or those from institutions with limited budgets, cannot make the same use of digital newspapers archives as their richer, mostly Anglo-American, colleagues. This social divide seems to be a problem of the digital humanities at large. Only affluent Western institutions can afford the investments in both the hard- and software, which are needed to kick-start the digital turn. This social divide was much less felt in the previous 'cultural' turn in the humanities, in which researchers from developing countries, such as the Subaltern Studies Group, played a very significant role.[6]

Digital newspaper archives also provide unequal access to different historical periods. This becomes evident when Delpher's selection criteria are studied. On its website Delpher presents its corpus as a unity: 'One million daily newspapers from the seventeenth, eighteenth, nineteenth and twentieth century.'[7] However, the corpus is actually divided into six periods – 1618-1800, 1800-1814, 1814-1869, 1869-1940, 1940-1945, 1945-1990 – while publications from the former colonies of Surinam, the Dutch Antilles and Indonesia are selected with still other criteria.[8] Nicholson argues that a tailor-made methodology has to be devised for three basic historical periods that can be observed in almost all digital newspaper archives. The periods can be very roughly described as pre-nineteenth century, the nineteenth century to 1910-1940 (depending on the archive) and the remaining part of the twentieth century until the present, which can be divided into two periods itself.[9]

The methodological issues plaguing digital newspaper archives in the first period have to do with both quality and quantity of the corpus. James Tierney identifies two basic problems for the seventeenth and eighteenth century British digital collection. First, inconsistencies in typography reduce the accuracy of OCR-scanning software significantly. Second, digital archives are often centered on the collections of contemporary collectors, like the Burney Collection of the British Library.[10] It can be added that collections of a single title are often incomplete, which makes the corpus especially unstable

[5] A. Bingham, 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians.' *Twentieth Century British History* 21.2, 2010, 225-231 (226).
[6] P. Burke, *What is Cultural History?* Cambridge: Polity 2008 (106-107).
[7] '1 miljoen dagbladen uit de 17e, 18e, 19e en 20e eeuw'. www.delpher.nl.
[8] 'Selectie van titels: 1618-1800', www.kb.nl.
[9] Nicholson 2013 (60).
[10] J. Tierney, 'The State of Electronic Resources for the Study of Eighteenth-Century British Periodicals: The Role of Scholars, Librarians and Commercial Vendors.' *Age of Johnson* 21, 2012, 309-338.

for quantitative research. One of the selection criteria of Delpher for the period 1618-1800 makes this very clear: 'A sufficient number of issues of a single title should be preserved, in order to make an analysis of content, technology and distribution for *a somewhat longer period of time* possible. [my italics, TS]'[11]

The second period, the nineteenth century, is often seen as the most complete and reliable. Absence of copyright restrictions, systematically compiled archives rather that incomplete private collections and improvements in nineteenth-century typography all lead to better and more stable results for both quantitative and qualitative methods.[12] However, this rosy picture deserves some adjustment. Copyright related concerns still play a major role since commercial publishers restrict access to the digital versions of out-of-copyright publications. Furthermore, archives in the second period may be systematically compiled, but they are 'collected' all the same: not by individuals but by national institutions (see next section).

The sheer quantity of titles and issues in the first half of the third period – 1900 to 1940 in Delpher – is often greater than the two before. However, the corpus is less complete and, as a result, not stable. Only a fraction of the entire media landscape, all the titles that were published in a certain period, is digitally accessible. In contrast, copyright issues plague the second half of the third period, which roughly coincides with the latter part of the twentieth century. In addition, Nicole Maurantonio identifies a very specific problem of digitized publications in this period. Digital archives, like LexisNexis, offer only text without the accompanying photographs. This means that a central aspect of twentieth century media, the interplay between text and image, is lost.[13]

It can be concluded that the corpus of digital newspaper archives is unstable for the entire period that they claim to span. The presented unity of the corpus is deceptive. Therefore, a long-term analysis that crosses different periods, one of the central promises of digital humanities, is highly problematic. Offering researchers more possibilities to select their own corpus would partly solve the problem. However, a major improvement could also be made if scholars became more perceptive to the unstable corpus and adjust their projects accordingly.

THE NATION STILL MATTERS

The digitization of newspapers and periodicals is a relatively costly enterprise. Therefore, it is not surprising that the main open access archives are all financed by national libraries and thus in the end by national governments. As Gerben Zaagsma remarks: '(…) the nation still matters, and it matters a lot.'[14] But how does the nation state exactly assert its influence over the digital corpus and how can this influence be reduced? This section will

[11] 'Er moeten van een krant voldoende exemplaren bewaard zijn gebleven om inhoudelijke, technische of distributieve ontwikkelingen op wat langere termijn te kunnen duiden.' 'Selectie van titels: 1618-1800,' www.kb.nl.

[12] Nicholson 2013 (60).

[13] N. Maurantonio, 'Archiving the Visual. The Promises and Pitfalls of Digital Newspapers.' *Media History* 20.1, 2014, 88-102.

[14] G. Zaagsma, 'On Digital History.' *BMGN – Low Countries Historical Review* 128.4, 2013, 3-29 (20).

argue that digital newspaper archives transplant the current, mostly national, media landscape to earlier periods without taking the specific historical form of these landscapes into account. This not only leads to biased research but also to a further underrepresentation of transnational and minority groups: a problem plaguing historical practice since its conception in the early nineteenth century.

This point can be illustrated with a few examples. Trove has historical newspapers from the whole of present-day Australia, including Tasmania, in its collection. However, in the late 1850s present-day Australia consisted of five separate colonies: New South Wales, Victoria, South Australia, Queensland and Van Diemen's land (see figure 1). The first four of these colonies were created by a partition of the older colony of New South Wales. The colony of New Zealand, established in 1841, was also a result of this partition. For the second half of the nineteenth century it can be argued that inhabitants of Van Diemen's Land and New Zealand equally depended on the colonial capitals of Melbourne and Sydney for their news. Therefore, New Zealand and Tasmania should be seen as the peripheries of a media landscape dominated by Sydney and Melbourne. However, newspapers from New Zealand cannot be searched in Trove because New Zealand is now a separate country with its own digital newspaper archive, whereas Tasmania became a part of Australia.

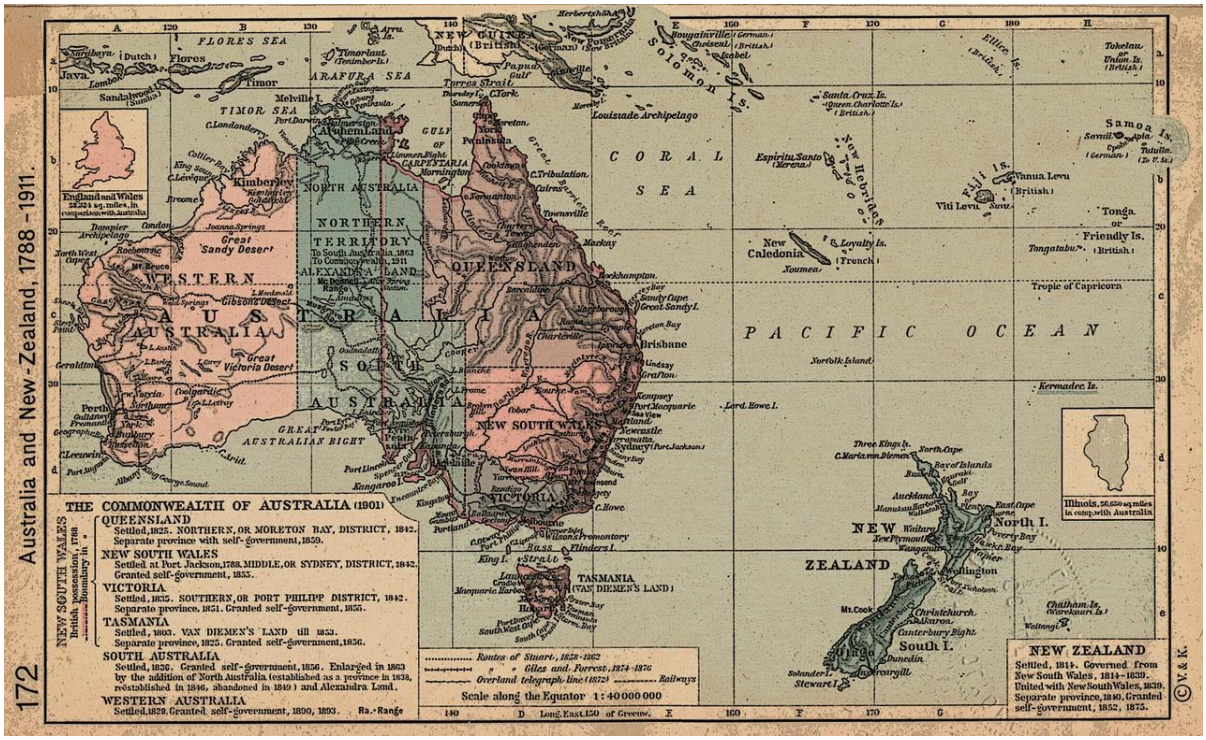

Fig. 1: Australian colonies in the nineteenth century. William Sheperd, *Historical Atlas*. New York: Henry Holt and Company, 1911 (172).

The same is true for ethnic and other minorities who have disappeared or are completely integrated nowadays. Note for example that the national library of New Zealand has a special digital collection of Maori newspapers – the indigenous people of New Zealand –

next to its regular collection of English publications. The Maori are placed outside the national narrative of New Zealand.[15] However, this actually is a positive exception: at least Maori newspapers are digitized. In general, publications of all sorts of minority groups, be they ethnic, religious or social, remain largely confided to their paper form. The same goes, as Zaagsma remarks, for publications of a transnational nature.[16] It follows that digital newspaper archives are generally biased towards liberal, middle class publications that were important in shaping the current nation state. By using the archives uncritically, scholars in the humanities run the risk of further obscuring marginal and/or transnational groups.

The historical form of media landscapes should be a part of methodological considerations of research that uses digital newspaper archives. It is, in this case as well, important that scholars recognize their agency over the corpus. Historical media landscapes can be reconstructed: newspapers from New Zealand should for example be included in projects that concern the Australian colonies in the nineteenth century, or New Zealand, Tasmania and other peripheries should be excluded altogether.

DIGITAL MEDIATION, LACK OF CONTEXT

Can datasets, like digital newspaper archives, be large enough so that researchers can ignore the context of publication? To clarify this question: proponents of digital humanities are fond of showing so-called Ngram graphs. These graphs show the frequency of a search term in relation to all the words in the corpus (see figure 2).

Fig. 2: the search term 'Illustrated London News' is related to the entire Google books database.

A number of things can be deduced from these kinds of images. For example, in figure 2 we can see that the *Illustrated London News*, the object of my own research, was first published in 1842 and quickly became popular. The enormous peak, or 'burst' as it is often called, around 1890 is much more difficult to explain. Does it correspond to the fiftieth anniversary of the publication in 1892, or is it a result of the corpus?

[15] See the Māori Niupepa Collection: www.nzdl.org.
[16] Zaagsma 2013 (21).

The numerous methodological problems involved in this kind of 'pattern images' are widely recognized and discussed.[17] All sort of questions remain unanswered: What was written about the *Illustrated London News*? Who wrote about it? And, maybe most importantly, what was the impact of all these texts? To solve these problems, digital humanities scholars propose to combine quantitative and qualitative methods. New software, like Texcavator introduced in the last *TS Tools*, is designed to make a switch between the two methods easier.[18] It should enable what the literary scholar Franco Moretti has called 'distant reading'.[19]

This is a step in the right direction, but it also leads to new questions. When should a researcher start (and stop) to apply qualitative methods? Can certain parts of quantitative results be explained with qualitative research, while others parts remain unsubstantiated in this way? Researchers are prone to qualitatively explain bursts; the exceptional parts of the quantitative results that often do not fit their expectations. However, because of the unstable corpus of most digital newspaper archives, there is no clear methodological reason why perceived unexceptional results can be taken for granted.

The digital mediation of historical newspapers and periodicals results in a problematic loss of context. Pre-digital press historical research, which entailed wading through countless issues of certain clearly demarcated publications, placed a strong emphasis on context.[20] Digital newspaper archives disconnect texts, mostly in the form of an article, from their original context – the entire publication – and tend to present 'hits' as pieces of neutral information.[21] As media historian James Mussel puts it: 'The genre of the digital newspaper archive attempts to structure an encounter in which users stop perceiving the resource and start perceiving its content. In other words, (…), the resource disappears leaving only apparently unmediated information.'[22]

An example will illustrate the problems that a lack of context can cause. Let's say that I am interested in how people judged the quality of illustrations in the *Illustrated London News*. I search Trove with the terms 'Illustrated London News' and 'drawing'. An article from the *Freeman's Journal* of 1851 pops up with the title: 'The English Press and the Catholic Hierarchy (*From the Dublin Review)*'.[23] In this article the *Illustrated London News* is mentioned in the following sentence that is directly pointed out to me by the Trove interface: 'The unscrupulous compound of bad drawing and bad writing, the *Illustrated London News*, which hitherto had been innocently amusing, must needs have its fling at Catholics.' If a quantitative method is used the article will be categorized as

[17] See G. Janssen & K. Wils, 'The End of Humanities 1.0.' *BMGN-Low Countries Historical Review* 128.4, 2013 (145-146) and the other contributions in the *Forum* section of this special issue on digital history.
[18] J. van Eijnatten, T. Pieters & J. Verheul, 'TS Tools: Using Texcavator to Map Public Discourse.' *TS. Tijdschrift voor Tijdschriftstudies* 35, 2014, 59-65 (64).
[19] F. Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary Theory*. London: Verso 2005. Cited in Nicholson 2013 (68).
[20] See for a discussion of digital newspaper archives and press historical research: M. Broersma, 'Nooit meer bladeren? Digitale kranten archieven als bron.' *Tijdschrift voor Mediageschiedenis* 14.2, 2012, 29-55.
[21] *Ibid.* (40-41).
[22] James Mussel, 'Elemental Forms.' *Media History* 20.1, 2014, 4-20 (16).
[23] *Freeman's Journal*, 12 June 1851.

'negative of the quality of illustration'. However, when the article is placed in its proper context a different picture emerges. The *Freeman's Journal* is a catholic publication and the *Illustrated London News* can indeed be qualified as (mildly) anti-Catholic. The qualification of the illustrations by the *Freeman's Journal* has almost nothing to do with an aesthetic appreciation, but mainly with the context of publication of both newspapers.

Technology can, and probably will, solve some of the abovementioned problems. In the foreseeable future software may be able to bridge the gap between quantitative and qualitative methods in a more satisfying way. The problem of lack of context can also partly be solved by another technological intervention. Detailed information about a publication – its owners, major editors, circulation, political outlook etc. – could be directly linked to articles that are found in digital newspaper archives. In this way, the political stance of the *Freeman's Journal* in 1851 would immediately become apparent. This, of course, means that digital newspaper archives have to invest in old-fashioned monographs of the publications in their corpus. Linking them to other open access databases like *Wikipedia* would also constitute a major improvement.

A POSITIVE CONCLUSION

Can the methodological difficulties of research based on digital newspaper archives be overcome? A positive answer to this question depends on the willingness of scholars to recognize and assert their agency over the archives and in the debates surrounding them. The basic components of digital archives are unlikely to change in the near future. It is high time that researchers treat digital archives as any other archive: a set of sources with countless possibilities as well as specific limitations. In relation to this, it can be seen as comforting that most of the methodological problems of digital archives and tools are not new at all. Researchers in the humanities have been interested in 'big data' ever since punched cards were first used to record censuses in the late nineteenth century. Answers to questions of previous methodological debates could be used to tackle the problems that surround the digital newspaper archive today.

The question whether new digital possibilities are transformative for the humanities, a proposition that is fervently defended and attacked, can only be answered by the quality of research that is done in the years to come. This quality can only be guaranteed if researchers are willing to leave the theoretical trenches of the digital humanities debate and have a pragmatic discussion about a new and digital methodology. This position sounds very familiar because it is repeated over and over again in the digital humanities debate. The intention to come to a new methodology seems to be omnipresent but consensus is still lacking. Students and researchers are in dire need of a (text)book that prescribes (enough) methodological rules for sound digital research. It would be a major step forward if a small part of the funds available for digital humanities research would be allocated to such a project.

•> THOMAS SMITS *is writing a PhD on the production of (trans)national identity by European illustrated newspapers in the nineteenth century.*