Interdisciplinary Collaboration in Studying Newspaper Materiality

Eetu Mäkelä[1][0000−0002−8366−8414], Mikko Tolonen[1][0000−0003−2892−8911], Jani Marjanen[1][0000−0002−3085−4862], Antti Kanner[1][0000−0002−0782−1923], Ville Vaara[1][0000−0001−7924−4355], and Leo Lahti[2][0000−0001−5537−637X]

[1] Department of Digital Humanities
University of Helsinki, Finland
first.last@helsinki.fi
[2] Department of Mathematics and Statistics
University of Turku, Finland
first.last@utu.fi

Abstract. This paper presents a collaboration between computer scientists, linguists and historians studying the material aspects of newspapers and developing a tool for that purpose. The paper describes how the back-and-forth collaboration in terms of research questions and technical challenges yielded insights both for solving computational problems as well as refining historical analysis. In the project, existing metadata was amended by reconstructing new materiality data from the Finnish digitised newspaper corpora. The analysis of such data is crucial for studying the development of newspapers, but can also inform other computational studies on the same data. The use of enriched materiality data allows for better understanding subdivisions in large corpora such as digitised newspapers, but also highlight that content and form interact. Content analysis of newspapers should therefore always take into account material properties of the studied material to properly grasp the cultural, social and political meanings embedded in the sources.

Keywords: Materiality of newspapers · Collaboration · Digital humanities.

1 Introduction

This paper offers a view to the collaboration undertaken at the Helsinki Computational History Group (COMHIS)[3] between computer scientists, historians and linguists on a project that studies the material dimensions of newspapers and their development [3].

The present day transformation from print to digital is not the first time newspapers have evolved drastically. Instead, this change of format reminds of similar transformations when the newspaper first appeared as a distinct material genre. One influential definition separating a newspaper from a newsbook or

[3] http://helsinki.fi/computational-history

pamphlet in its early days was that a newspaper was a "sheet of two or four pages, made up in two or more columns" [10]. The Dutch had two-column news at the time, while civil war in Britain saw both the rebels and the crown printing their propaganda. It took, nevertheless, centuries before journalism became a profession of its own and newspapers took their particular shape in the mid-nineteenth century [20,1,2,11,13,23].

In the context of digital humanities, newspapers have become an iconic example of "big data" research (cf. [5,15,7], https://numapresse.org/). While in localised research [8,28] the material can be thought uniform, in the big data approaches it is striking how little attention is paid to what the data consists of. A telling example of waking up to this is the Oceanic Exchanges project (https://osf.io/wa94s/) where M.H. Beals and Ryan Cordell quickly concluded that mapping metadata across its many datasets is to be one of its most important contributions (https://twitter.com/ryancordell/status/1001845719341285377).

Framed against this background, the idea of this paper is to outline how we developed a tool to uncover and explore the varied materiality of newspapers. As part of the large-scale digitisation, the accessibility of historical newspapers has improved drastically, but at the same time much of the information about the size, shape and feel of the newspapers, that was so central to past readers in understanding what kind of documents they were perusing, has to a large extent been hidden from view. Interestingly, the digitised versions of the newspapers also allow for large-scale study of their material dimensions – an opportunity that has so far been paid very little attention to. In our case, our focus on materiality is also just one aspect of the group's larger interest in studying the nature of early modern public discourse through the analysis of structured and unstructured data relating to newspapers and other printed materials.

In what follows, we will first briefly explain the background for this study and how it fits the group's publication history. Then, we'll shortly discuss the type of data we started our work from, before going into detail on how the research process that led to the materiality explorer tool actually happened. Finally, we will describe the tool itself and the tentative results we've obtained using it, before concluding by outlining directions for future work.

2 Studying the Materiality of Newspapers

The first time that data on the materiality of newspapers was extracted and studied by us at the COMHIS group was as part of the Helsinki Digital Humanities Hackathon of 2015[4]. After that, intermittent analyses on both the content as well as metadata such as language, location and form of the newspapers was done as part of the internal dialogue of the research group, in part in the context of the Academy of Finland funded project on "Computational History and the Transformation of Public Discourse in Finland, 1640-1910"[5].

[4] http://heldig.fi/dhh15

[5] http://www.aka.fi/globalassets/32akatemiaohjelmat/digihum/hanke-esitteet/salmi-digihum.pdf

Slowly, these explorations coalesced into multiple conference presentations on the subject. Mostly, the actual work happened in sporadic bursts, often with one of the more computationally oriented researchers in the group being inspired to run a particular analysis, which then led to back-and-forth exchange between the historians and the experts in quantitative methods to better interpret and fine-tune the analysis. In this process analyses were also designed to be more aligned with research questions pertinent to newspaper history, and new analyses were requested by the historians.

In time, these explorations led to more focused research questions, dealing with the modernisation of newspapers in Finland in two main languages. As newspapers became more frequent, more topical and gained a larger format, they started resembling the modern newspapers that we encounter today (or perhaps those of our childhood). In particular, we wanted to trace the asynchronicity that was present between Finnish-language and Swedish-language papers. Editors and other intellectuals in Finland operated mostly in both languages, and thus the newspapers were developed in constant cross-fertilisation across the language border, but still the different language spheres developed at different paces. While Swedish-language papers were generally more advanced up to the 1860s and 1870s, Finnish-language papers became leading by the turn of the century 1900 due to growth both in terms of readership and places of publication.

A problem with our early explorations was that they had been done in a haphazard, off-the-cuff manner by different people using different versions of the data, so they were not mutually consistent and reliable. An impetus to change this came when one of the conference presentations led to an invitation to write up the work more formally for the Journal of European Periodical Studies (JEPS). At this point, it was decided to take one single version of the data as the source, and calculate all material and linguistic indicators from that. A more thorough analysis of the trustworthiness of the pipeline and the dataset itself was also undertaken.

For the JEPS article, the figures and analyses used to inform the content started as those that had arisen organically as part of the internal dialogue within the group. However, when polishing the art, a dialogue was held between the historians and the statistical visualisation experts on what the core message was. This led to replacing earlier more explorative versions of the visualisations with ones designed specifically to convey particular arguments. At the same time, the visual outlook of all graphs was unified.

After working on the JEPS article, the group had a relatively good notion on what the important aspects of materiality in the data were, and how they could best be visualised and explored in a unified manner. This led way to the development of an interactive materiality explorer. Through this, there was more freedom for the content experts to explore the phenomenon, with much less frequent need for the computer scientists to run customised analyses or change the parameters of the exploration.

2.1 Extracting and Deriving Material Aspects from ALTO XML

In order to understand what the group was working with, it is relevant to understand the usefulness of ALTO (Analyzed Layout and Text Object, https://www.loc.gov/standards/alto/) files that were luckily available for the project. ALTO files contain a description of the visual organisation of content on a page, at the core of which are the individual words and their page coordinates. At the same time, the words are also grouped into blocks, often corresponding to paragraphs or columns. The format also contains general layout information, such as the sizes of margins and main printed area.

The usefulness of ALTO for analysing materiality crucially depends on the choice of the measurement unit in which all coordinate and size information is given. Here, the format gives a choice from three options: mm10 (tenth of a millimeter, the default value), inch1200 (1200th of an inch) or pixel. Of these, the first two directly relate all measurements to actual physical dimensions, while the pixel coordinates do not. However, even then, the information on original physical dimensions can be recovered if the DPI value of the image is known, information given in the METS metadata files originally often accompanying the ALTOs. Unfortunately, many collections such as the Dutch Delpher (https://delpher.nl/) and French Gallica (https://gallica.bnf.fr/) provide their ALTO data specifically using pixel coordinates, while not giving out the METS files (which would also contain logical segmentation information, separating the text into articles and adverts). Similarly, the National Library of Finland (http://digi.kansalliskirjasto.fi/), while providing the METS files, explicitly removed scanning information from them until requested otherwise.

These examples highlight how little thought is given to the material dimension of the newspapers in most digital processing pipelines even before the user interface layer. Luckily, the ALTO files of the National Library of Finland had a MeasurementUnit of mm10. Given this, we could easily extract page size, printed area and character and words counts for each page. Besides these, the ALTO file also contains some style information that can be extracted. Currently, we disregard the information on left/center/right alignment, but do extract font information. Directly given are the size, face, style (bold/italic/underline) of each font used, to which we add the calculated number of characters and words written using that font, as well as the overall page area covered.

For each page, we also extract all text box coordinates (visualised in Figure 1). While these are primarily meant to locate text visually on the page in reader interfaces, they can be processed to yield layout information. First, we extract column counts using a lighter-weight process than the computer vision approach used in [6]. We scan the page from top to bottom, for each Y coordinate counting the number of text boxes present there. This yields a distribution associating all column counts with the area they control on the page. Mapping shifts in the amount of columns seems to be one of the clearer indicators of changes in layout. This is useful both for assessing the general development of newspaper layout, but also for identifying particular instances in which editors felt they needed to introduce changes to the layout. Columns obviously roughly correspond to page

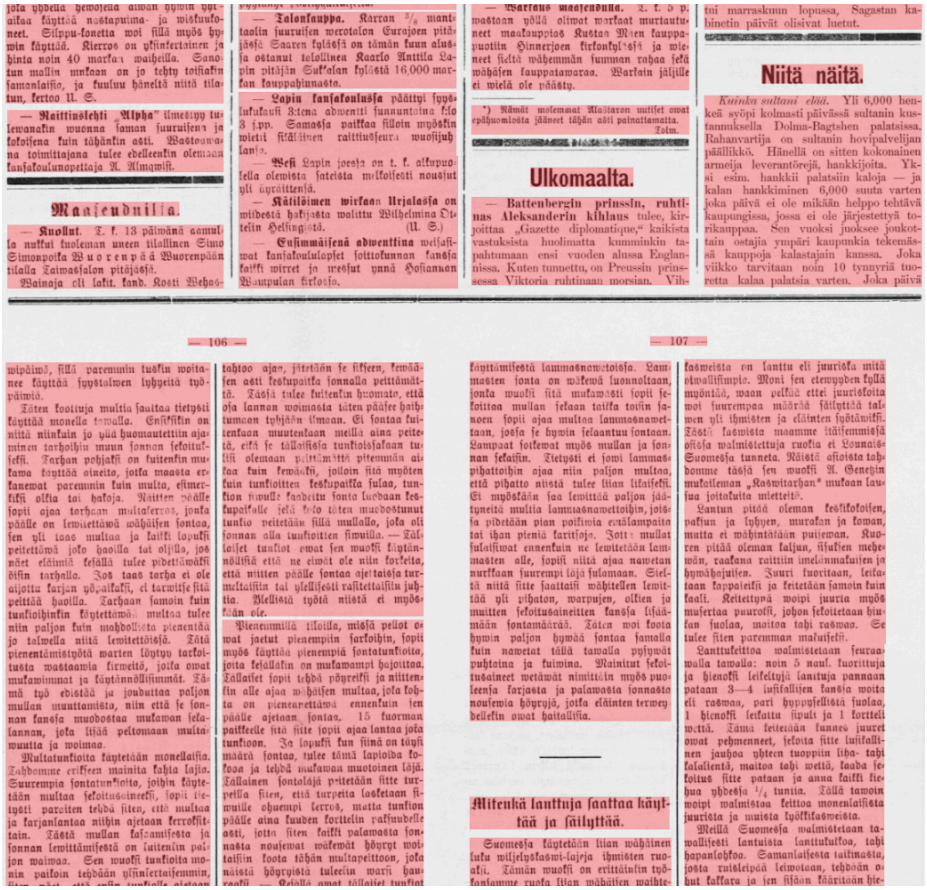

Fig. 1. ALTO text blocks overlaid on a newspaper page.

size, but changes in the width of columns are also indicative of how newspapers explored issues of readability.

2.2 Developing the Materiality Explorer

The Helsinki Computational History Group sits along the same corridor at the University of Helsinki. This physical presence is an important part of the group's work, but so is Slack. As a tool, Slack is an effective way of communicating while sharing research ideas and findings, but it also has the benefit of functioning as a means of documenting much of the group's efforts. To provide an example of this, we will present shortly below an analysis of our Slack communication relating to developing the materiality explorer.

On this particular project, the intensive work started – according to the comments on Slack – on 30 October 2018. It began when Eetu Mäkelä posted first images of a general visualisation unifying multiple aspects of materiality data. From the beginning, it was clear that the point of the materiality explorer was to experiment with different ways to define gross materiality categories in newspapers. It took however few days before the work on the development got going seriously.

Nevertheless, by 12 December 2018, there were altogether 355 different messages (8-9 on average / day) on the group's slack channel dedicated to newspapers about this work. Altogether 9 people participated in this online discussion with different kinds of input. While some people just posted one or a few notes, two group members had more than 100 messages each devoted to this project. There was also, of course, actual human interaction in real life, which is unfortunately not recorded. What drove the work was a looming deadline for the DH2019 conference at the end of November.

Analyses undertaken on development versions of the materiality explorer soon led us to realise that some of our data was problematic. Here, an important point to notice is that computational processing of the data did not start with us, but included also the scanning and OCR of the pages, as well as the metadata work done on the collection at the National Library of Finland. What we found out was that the National Library of Finland had used altogether 22(!) versions of scanning software. A key problem for us was that some of these did not differentiate between Fraktur and Antiqua fonts. By using metadata to analyse which newspapers were scanned with which version, we determined that reliable font identification could only be had up to the year 1910. We also employed some spot checking to compare algorithmic results to the manually keyed metadata, and for example decided to use the raw data directly for page size and date range estimation instead of the same information as keyed.

After a few days of pondering about the effects of these technical problems for analysis, we started focusing more on the question of cramming information on one sheet of newspaper – thinking also about the readability of the text on the page. At the same time, a more extensive reading of relevant secondary sources begun to figure out the technological development (especially with the DH2019 conference submission in mind). The reason for doing this was to find possible

identifiable markers to flag differences and effects caused by changes in printing techniques. For example, the emergence of lithography offset printing was one such technique whose effects we could clearly identify also in the data.

We also soon advanced to thinking about layout and the relevance of the front-page. The idea was to figure out ways of detecting typographic changes on the front page within the context of a single newspaper to understand its development. At this time, it came as an idea to try to identify an instance of a (statistically) typical front page for each decade over time for both Finnish and Swedish language newspapers. Once we knew that this is possible based on the tools at hand, several different kinds of experiments to find "typical" newspaper proportions using the materiality explorer were made. Our deliberations particularly echoed those by Myllyntaus [21], who has done a huge amount of work on these issues without the statistical apparatus that we have on hand today. What was visible in our data was that importing the rotary press and offsetting technology to Finland changed the newspaper layout in the papers that could afford this technology in a very short period of time. We were able also to see that the linguistic and geographic diversity in Finland led to a situation where print runs were smaller and there was more type-setting ongoing than in some larger European countries.

We realised also that we could group different language newspaper published by the same publisher in the same year at the same location together in order to study their layout and content. This would help us to understand how news possibly circulated from one language to another and how different advertisements for example are presented in different languages in Finland. Many previous scholars have been interested about different language profiles of newspapers in different Finnish towns. What these scholars haven't realised is that the question of type, layout etc. can also have intellectual relevance. So, to ask if parallel newspapers are coming from the same publishing house (as they at times do) is a relevant question to ask.

On Sunday 25th of November, Eetu Mäkelä posted an image of the mean front page of Helsingin Sanomat in 1907. This also marked the saturation point of the development phase of this part of the work. There were still new ideas coming in, for example, about terseness of language in newspapers in order to allow cramming, but the main thing for us at this point was to prepare for the DH2019 deadline that was on 27th of November. Perhaps we need to wait for the next deadline to get back seriously to this project.

2.3 The Materiality Explorer Interface

As it currently stands, the materiality explorer has three main functionalities, each aimed at a different use cases. Common to all views are a set of selectors, allowing to limit the set of newspapers under study. Currently, these hold facilities for limiting study by 1) time, 2) newspaper language, 3) newspaper lifetime, 4) printing location and 5) individually by title.

In the overview view shown in Figure 2, first presented is the absolute amount of data. This is important, as all the other graphs display their information as

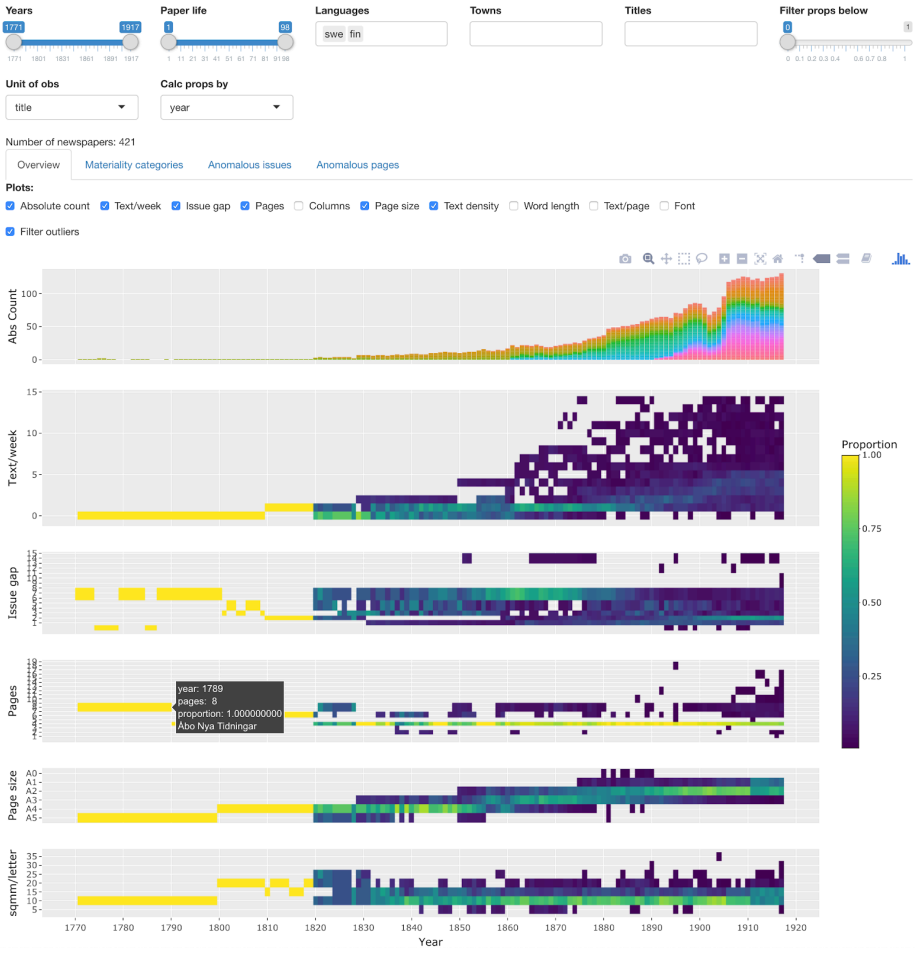

Fig. 2. The materiality explorer.

proportions of the whole. Depending on a user selected option, this proportion may be calculated by year, by month or by week. In addition, the user can select whether they want an observation to be titles, issues or pages. Here, the choice depends on what one is interested in. Counting by titles treats each newspaper as a single unit, allowing exploration of the breadth of newspapers without regard to how often they appeared or how large they were. On the other hand, if one is more interested in the amount of information consumed by an end reader, then possibly counting by issue or even by page is appropriate. Another use case where observing by page or issue may be more interesting is when studying the development of a single newspaper, where the differing publication frequencies and page sizes no longer matter, but instead even singular aberrant pages are interesting.

After this absolute count, a baseline measure of text per month is given, against which all the materiality information can be contrasted. This baseline was developed in consultation between the computer scientists, historians and linguists to provide a language-neutral measure for throughput. By counting the number of characters each newspaper produces in a month without regard to how they are divided between issues or pages, this measure shows how much content needs to be transmitted. As this quantity rises, newspapers must respond with material innovations, whether by increasing page count, page size or publication frequency, or by cramming more material into available space by decreasing font size, line breaks or margins.

A second view allows grouping the data by a combination of material dimensions, thereby allowing exploration of archetypal materiality categories. Finally, two distinct views allow the user to explore respectively page and issue-level material anomalies in the data: for example pages which have much more text than others or pages with abnormal layout, or issues with appendixes or which appear on the same day as another issue. These both lead the way for interesting qualitative analyses, but can also be used to remove abnormal data from further quantitative computational analyses of either form or content. In our project, the anomaly detection served as a central method for exploring the data as well as identifying errors in the code or metadata. Here historians had a rich source for detecting counter-intuitive findings, and often those findings led to feedback that could further improve coding efforts.

At present, we are using the interface to exploratively develop hypotheses on common development patterns as well as archetypal materiality categories. Both of these are interesting in themselves as objects of study, but can also be used later to partition datasets for other computational processing such as OCR retraining or content analysis. While this current stage of explorative hypothesis development is interactive, visual and qualitative, our plan for the next stage is to explore statistical validation of such hypotheses using for example Granger causality and archetypal coverage measures. Once developed and tested, these again will be added to the interface to enable further self-sufficient analysis in a more trustworthy manner.

3 Discussion

At the outset of this project, we asked in particular how the modernisation of newspapers published in Finland could be better understood by looking at the form, shape, location and publication frequency in newspapers published in different languages (Finnish and Swedish being the main publication languages). The project produced one article that pays particular attention to the different speed in development with regard to Swedish-language and Finnish-language newspapers in Finland. Further, we produced an interactive materiality explorer that helps researchers understand the development of material aspects of newspapers. We also developed preliminary hypotheses that will be shortly discussed below with regard to different categories of materiality.

For the Finnish newspapers, the data shows a general order in how they expanded: first, layout was changed to include more words per page; second, page size was increased; third, publication frequency was increased and only after that was the amount of pages increased. This last step often coincides with the introduction of rotary presses, which allowed newspapers to more easily be composed of more than four pages, and also allowed them to move back from large page sizes to more easily handled formats. Simultaneously, the data shows also high variability, where papers not only frequently printed supplements, but could switch back and forth between formats inside a single week, or cram text into a special issue through diminished line breaks. Similar shifts took place also with regard to fonts. Newspapers explored different Fraktur and Antiqua fonts to try out readability, but also because fonts were oftentimes used to signal that the contents was aimed for a particular audience. While there are plenty of exceptions to this, it seems that Fraktur was more often used when dealing with economy and religion, whereas Antiqua was reserved to politics, philosophy and the high arts. To test such hypotheses about different uses for fonts and relating that to the overall development of newspapers, we still need more robust statistical information. We also aim to compare used fonts and with other factors, such as language frequency and size of newspapers. (For the history of newspaper layout and design, see [4,17,22,24,26,27,12].) Compared to earlier studies, our data driven approach gives us a great opportunity to evaluate the main findings of earlier historical studies of newspaper materiality [18,30,21,9,16,32].

What we also aim to do with these patterns is to develop evidence-based archetypal categories of newspapers across history. We are then able to trace and compare these through time and place, but also use them to study the evolution of individual newspapers. These categories will also help us understand the newspapers as objects of intellectual activity, creating a theory of different historical maturity levels of newspapers. This in turn will help us chart the development of public discourse over time.

Besides presenting the research process regarding the material development of newspapers as a genre in itself, we argue that content and form interact, and thus big data approaches to newspaper analyses also need to pay attention to material differences in order to accurately understand the subdivisions in large corpora. Here, this paper continues on a path previously charted by for example

[19,29,14], while providing an orthogonal axis to those expanding study from text to visual elements [25,31]. For example, using the metadata we can create meaningful subsets of the data that are balanced by paper type for for example topic modelling or teaching automated transcription algorithms.

Here, Finland makes an intriguing case for digital history because its public sphere is bilingual, with newspapers in both Swedish and Finnish. One interesting phenomena that arises from this are publishers publishing newspapers in both languages. For example, in Kotka there are both Finnish and Swedish newspapers by the same publisher with identical layouts and advertisements. Such could be used to create parallel corpora, interesting for the study of commonalities and differences between the different language public spheres, but also perhaps as material for machine translation.

References

1. Allen, J.E.: The Modern Newspaper: its typography and methods of news presentation. [With illustrations. New York & London, Pp. ix. 234. Harper & Bros. (1940)
2. Baldasty, G.J.: Commercialization of News in the Nineteenth Century. University of Wisconsin Press, Madison (2014)
3. Bode, K.: The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History. Modern Language Quarterly **78**(1), 77–106 (Mar 2017). https://doi.org/10.1215/00267929-3699787, https://read.dukeupress.edu/modern-language-quarterly/article/78/1/77-106/19924
4. Broersma, M. (ed.): Form and style in journalism: European newspapers and the representation of news 1880-2005. Peeters, Leuven, Dudley, MA (2007)
5. Buntinx, V., Bornet, C., Kaplan, F.: Studying Linguistic Changes over 200 Years of Newspapers through Resilient Words Analysis. Frontiers in Digital Humanities **4** (2017). https://doi.org/10.3389/fdigh.2017.00002
6. Buntinx, V., Kaplan, F., Xanthos, A.: Layout analysis on newspaper archives. In: DH2017 abstracts (2017), https://dh2017.adho.org/abstracts/193/193.pdf
7. Cordell, R., Smith, D.: What News is New?: Ads, Extras, and Viral Texts on the Nineteenth-Century Newspaper Page. In: DH2017 abstracts (2017)
8. Cristianini, N., Lansdall-Welfare, T., Dato, G.: Large-scale content analysis of historical newspapers in the town of Gorizia 1873–1914. Historical Methods: A Journal of Quantitative and Interdisciplinary History **51**(3), 139–164 (Jul 2018). https://doi.org/10.1080/01615440.2018.1443862
9. Gustafsson, K.E., Rydén, P.: A History of the Press in Sweden. NORDICOM, Göteborg (2010)
10. Hutt, A.: The changing newspaper; typographic trends in Britain and America 1622-1972. Gordon Fraser, London (1973)
11. Høyer, S., Pöttker, H.: Diffusion of the news paradigm 1850-2000 (2014)
12. Kapr, A., Forssman, F., Willberg, H.P.: Fraktur: Form und Geschichte der gebrochenen Schriften. Hermann Schmidt, Mainz (1993)
13. Kutsch, A.: Journalismus als Profession: Überlegungen zum Beginn des journalistischen Professionalisierungsprozesses in Deutschland am Anfang des 20. Jahrhunderts. In: Blume, A., Böning, H. (eds.) Presse und Geschichte: Leistungen und Perspektiven der historischen Presseforschung, pp. 289–325 (2008)

14. Lahti, L., Marjanen, J., Roivainen, H., Tolonen, M.: Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly **0**(0), 1–19 (Jan 2019). https://doi.org/10.1080/01639374.2018.1543747
15. Lansdall-Welfare, T., Sudhahar, S., Thompson, J., Lewis, J., FindMyPast Newspaper Team, Cristianini, N.: Content analysis of 150 years of British periodicals. Proceedings of the National Academy of Sciences **114**(4), E457–E465 (Jan 2017). https://doi.org/10.1073/pnas.1606380114
16. McReynolds, L.: The News under Russia's Old Regime: The Development of a Mass-Circulation Press. Princeton University Press (1991). https://doi.org/10.2307/j.ctt7zth51
17. Moen, D.R.: Newspaper layout and design. Iowa State University Press, Ames (1989)
18. Moran, J.: Printing Presses: History and development from the Fifteenth Century to Modern Times. University of California Press (1973)
19. Moreux, J.P.: Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment (Aug 2016), https://hal-bnf.archives-ouvertes.fr/hal-01389455/document
20. Morison, S.: The English Newspaper, 1622-1932: An Account of the Physical Development of Journals Printed in London. Cambridge University Press, 1 edition edn. (Oct 2009)
21. Myllyntaus, T.: Suomen graafisen teollisuuden kasvu 1860-1905. University of Helsinki, Helsinki (1981)
22. Olson, K.E.: Typography and Mechanics of the Newspaper. D. Appleton and Company (1940)
23. Pettegree, A.: The Invention of News: How the World Came to Know About Itself. Yale University Press (2014)
24. Presbrey, F.: The history and development of advertising. Doubleday, Garden City, N.Y. (1929)
25. Smits, T.: Illustrations to Photographs: Using computer vision to analyse news pictures in Dutch newspapers, 1860-1940. In: DH2017 abstracts (2017)
26. Sutton, A.A.: Design and makeup of the newspaper. Prentice-Hall (1948)
27. Swanson, G.: Graphic Design & Reading: Explorations of an Uneasy Relationship. Allworth Press (2000)
28. Tilles, D.: The Use of Quantitative Analysis of Digitised Newspapers to Challenge Established Historical Narratives. Roczniki Kulturoznawcze **7**(1), 83–97 (2016). https://doi.org/10.18290/rkult.2016.7.1-4
29. Tolonen, M., Lahti, L., Roivainen, H., Marjanen, J.: A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History **0**(0), 1–22 (Dec 2018). https://doi.org/10.1080/01615440.2018.1526657
30. Tommila, P., Salokangas, R.: Sanomia kaikille: Suomen lehdistön historia. Kleio ja nykypäivä, Edita, Helsinki (1998)
31. Wevers, M., Smits, T., Impett, L.: Modeling the Genealogy of Imagetexts: Studying Images and Texts in Conjunction using Computational Methods – DH2018. In: DH2018 abstracts (2018), https://dh2018.adho.org/en/modeling-the-genealogy-of-imagetexts-studying-images-and-texts-in-conjunction-using-computational-methods/
32. Wilke, J.: Belated modernization: form and style in German journalism 1880-1980. In: Broersma, M. (ed.) Form and style in journalism : European newspapers and the presentation of news, 1880-2005, Groningen studies in cultural change, vol. 26. Peeters, Leuven, Dudley, MA (2007)