DHQ: Digital Humanities Quarterly
2023
Volume 17 Number 1

“The Page Is an Image Again:” Bleedmapping as an Analysis Technique for Historical Newspapers

Quintus van Galen <quintusvangalen_at_gmail_dot_com>, University of Amsterdam https://orcid.org/0000-0003-0776-3871

Abstract

Analysis of historical periodicals through digital tools is still a predominantly text-based field. This paper seeks to expand the toolbox of researchers of these sources, and addresses the observed capability gap, by proposing a new technique for the discovery of appearance patterns in historical newspaper collections. It calls this technique Bleedmapping after the effect it produces, making it seem as if a multitude of pages have “bled through” each other and showing the greatest concentration of articles matching a chosen search parameter visually in an intuitive and easy to interpret manner. This approach greatly extends the toolbox of scholars by providing them a method to explore the metatextual and spatial context of the texts they study. Using a case study on Reynolds' Newspaper this article shows the value of Bleedmapping when used alongside a text-based LDA analysis, showing the prevalence of the British Empire in the paper's “Notices to Correspondents”.

“The Page Is an Image Again:” Bleedmapping as an Analysis Technique for Historical Newspapers

When I was younger, I used to know exactly which page of the paper the sports column was on, just as my father could find the stock updates without looking. The TV-guide had a rear cover dedicated to children's letters and could always be found upside-down on the couch. Books were filled with dogears and would fall open on favourite pages. That we could navigate through these complex textual compositions by referring to their spatial nature – what page and column a certain kind of content lives in are just some intimately everyday examples of the relationship between a text and the spatial context it occupies. 1

When considering the meanings behind any word, our first instinct as researchers, and textually-minded humans, is to look at the words in the spaces around it. Are the signs that point to a more complete understanding of its meaning in the surrounding sentence, paragraph or page? There exist a multitude of techniques that provide researchers with these insights: Latent Dirichlet Analysis, Latent Semantic Indexing, Hierarchical Latent Tree Analysis, Concordance analysis, and many more. Yet all of these ignore the non-textual aspects of a text: the choice of font, the addition of images or embellishments, or the placement on the page. Already in 1927 the French philosopher and poet Paul Valérie wrote in an essay on the importance of the materiality of text. He stated: “Une page est une image. Elle donne une impression totale, présente un bloc ou un système de blocs et de strates, de noirs et de blancs, une tâche de figure et d'intensité plus ou moins heureuse.”[1] However, digitised historical collections are often accessible only through a text-based interface [Mussell 2012], which has the effect of masking its non-textual aspects. 2

Other methods have been developed to analyse periodicals in a manner that rescues these sources from being reduced to just text. Some early work on the form of the newspaper stems from [Barnhurst and Nerone 1991], [Barnhurst and Nerone 2002], who provided the theoretical foundations and proved its relevance to studying the cultural connotations of the patterns of placement. Their work relied on a visual close reading of their selection of newspapers, and sought to reach a general context true for all papers, rather than a specific context of one article. It also meant their methods do not scale well to encompass a digitised archive of thousands of newspapers. Additionally, the result of their visual close reading is itself textual, which has been described as problematic by some scholars, as it occludes the original nature of the material [Haskell 1993]. Some effort to address these problems has been made recently by Beals [Beals 2018], however her proposed visualisation of word counts doesn't show the space itself, only an estimate based on the length of text. Meanwhile Moreaux [Moreaux 2016] has shown the value of metadata analysis and visualisation for nineteenth-century newspapers. 3

This article presents a solution to this problem in the form of a tool for visualising the spatial pattern of article placement based on the positional data encoded in its metadata. The visualisation proposed by this paper will allow the heatmapping of common places where a given subset of articles in a periodical appear, allowing for the identification of the context in which a text needs to be seen. 4

Bleedmapping was developed as a tool to answer a simple research question: where do articles appear within a paper's pages, in single issues and over a longer print run of a title? It does so by extracting the positional metadata for each article in a generated subset, and visualising these as grids on a density map (heatmap), creating the effect of thousands of pages being put on top of each other and bleeding through to reveal the meta-textual structure that frames its contents. Answering this question is relevant for all research searching 5

for non-textual contexts of appearance, but it is of particular interest for studies in historical identities. It has long been acknowledged that newspapers signpost the reader whether a story should be read in an national or international context by grouping stories together. The banality of this act serves as a marker of identity for the reader, who sees their own perspective on the world validated by the paper's distinction between “homeland” and “foreign”, between “ingroup” and “outgroup” [Billig 1995]. Beyond studying the identity of any prospective readers of a publication, bleedmapping also allows for an insight in the periodical itself. As the distribution of articles through a print run of issues is not random, but instead a deliberate act on behalf of the editor, we can look at the pattern of article appearance as part of construction of the identity of the publication itself [Mussell 2012].

There are several advantages to the approach Bleedmapping takes. First, it is methodologicaly and epistemologically transparent. This means it avoids being a black box by virtue of not just being transparent to tool critics wishing to investigate the source code, but also to any scholar unfamiliar with digital approaches who may want to use it. The basic function of bleedmapping can be understood as a digital version of a manual operation: drawing and counting rectangles. In this way, it stands vis-à-vis methods such as topic modelling, which to fully understand require a degree of mathematical aptitude that is seldom found in the humanities. The second advantage is from the data it uses. Instead of being forced to rely on approximations of place and size Bleedmapping can rely on (relatively) accurate positional data. 6

It pays to take a moment to reflect on that positional data itself. Most digitised newspaper archives generate article blocks as part of the image processing undertaken when a page is transformed from a “dumb” image to an archival page [Gatos et al. 2000]. Such a step is required to have the OCR software “read” the text accurately, and to allow text to be read across a page boundary. This article segmentation step is the field where advances in Document Image Recognition and Analysis have made major contributions to accuracy [Meier et al. 2017], [Oyallon and Rabin 2015]. The coordinates of these article blocks are then retained in the metadata of a page, and are key part of keyword search implementation, allowing the user to be dropped exactly on the article they requested. Crucially, the way search indexes are set up suggests that these coordinates are never used outside of this purpose: the coordinate field cannot be searched, nor easily requested. Bleedmapping therefore relies on creative reuse of this data, and has to take at times roundabout ways to extract it. 7

When it comes to discussing the process of Bleedmapping, it is important to first discuss the source of the data used. Not each digital newspaper archive is created equally, and some were produced in a more structured manner than others. For example, the Delpher-archive of the Dutch Royal Library is very consistent in the quality of its segmentation and has a uniform resolution for all page images, while the British Library 19th Century Newspapers archive was digitised by two different external partners, producing different image resolutions [Fyfe 2016], [Beals and Bell 2020]. This has substantial implications for the Bleedmapping process, as it means in one case resolving scaling issues is vital, while in the other it can be approximated in favour of processing speed. The implementation below is for the British Library 19th Century Newspapers archive on the “legacy” Gale Text Mining Drives produced prior to 2018 [Beals and Bell 2020]. Yet for all implementations, as alluded to above, the existence of the positional metadata is crucial. Without this, bleedmapping is not possible. Thus, if this data is not present in the archive, the researcher has to either generate zoning data themselves from page images, or opt for a text-length approximation approach, such as used by Beals [Beals 2018]. 8

Overall, the process of generating the visualisation is divisible into four main steps: 9

1. The *(Meta)data lookup* collects the data that the visualisation corpus needs from the different parts of the archive with a keyword search.
2. The *Scaling Step* prepares the data for visualisation by harmonising the coordinate systems that each article is in, and resolving, as far as possible, ambiguities in the assignment of blocks of text to multiple pages. The harmonised coordinate systems are then handed off to the *Table-of-Occurrence Generator*. It should be noted that this step is only necessary in the case an archive does not consist of images of a uniform size and resolution.
3. The *Table-of-Occurrence Generation* tallies the number of articles that occupy a certain space on a certain page into a tabular format.
4. The *Heatmap Visualisation* generates the final image.

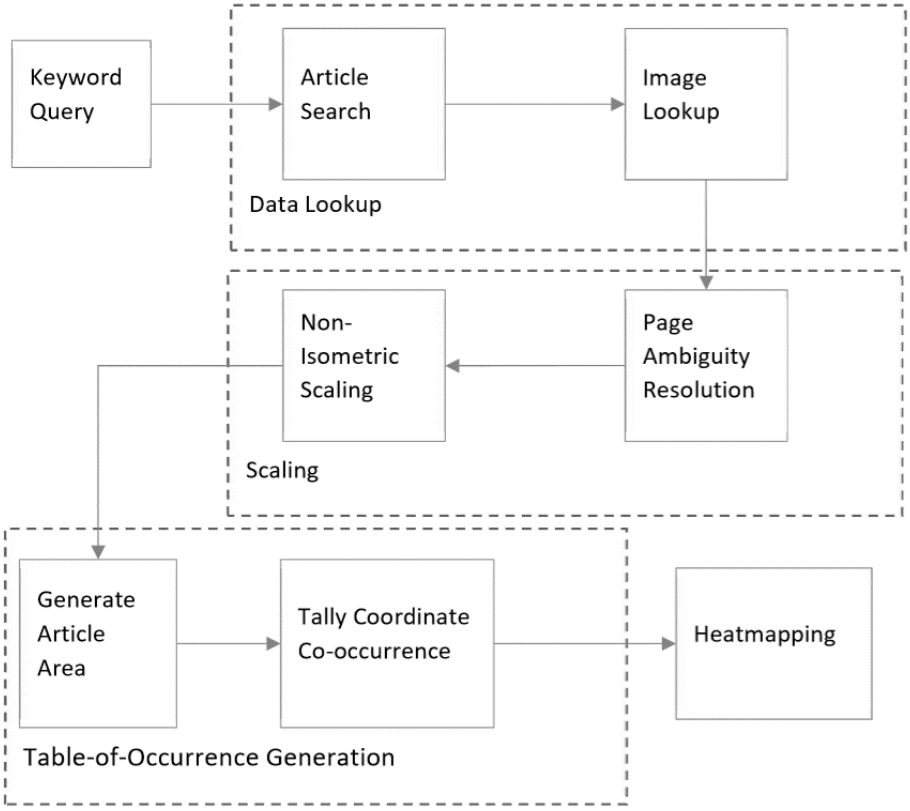


Figure 1. Graphic representation of the steps involved in generating article placement heatmaps. Each dashed rectangle corresponds to one of the sections below.

Metadata Lookup

The first step to visualising the placement of newspaper articles is to create a subset out of the corpus. This subsetting serves an epistemological goal. After all, if we were to visualise all the articles in the corpus, we would produce a representation of all articles in the newspaper, which means there would be no areas of greater intensity from which to draw any conclusions. Thus, the creation of a subset of articles whose positions might produce insight into the subject of choice, or which might answer a research question is a necessity. For example, if we wanted to visualise patterns in the placement of poetry columns across four decades of a specific newspaper, we would first need to identify and create a subset of all the poetry columns that we wish to measure. Theoretically, this subset could be assembled by manually identifying each column, but in order to visualise long-term trends and deal with large bodies of journalism, we need to identify material for our subset using digital search methods. 10

The visualisation corpus is first generated by the keyword search, which extracts the article ID's, article coordinates, and the article text. Next, the list of article ID's is used to perform a reverse lookup of the source images, using the page ID's derived from the articles. This produces a link to the page's raw image file on disk, which is then accessed and the image size in pixels and resolution in dpi pulled from its metadata. The results are used to build a dictionary of page ID's, sizes, and resolutions for the nest steps, alongside a regular list of articles from the keyword search. This process is represented in Figure 2. 11

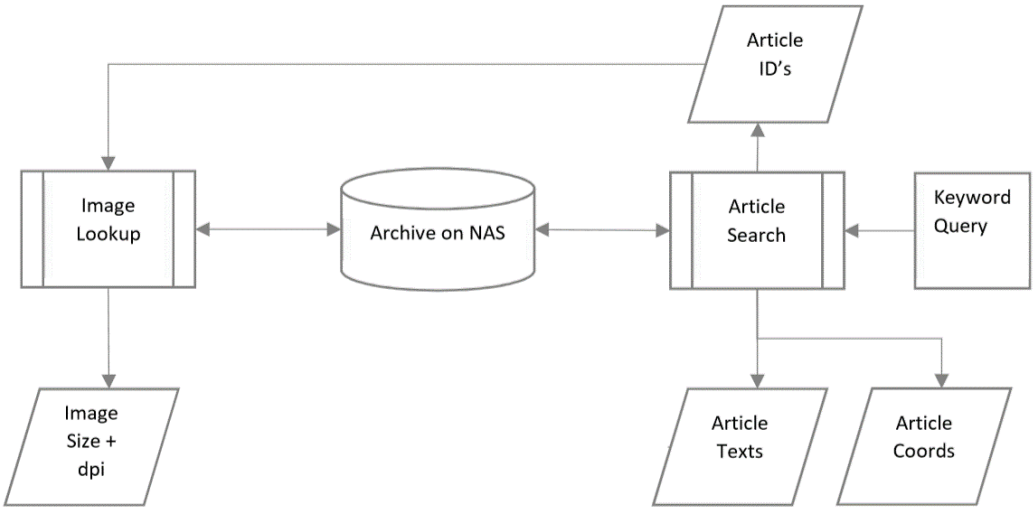


Figure 2. Diagram of the retrieval of article- and page data. The keyword query is used to subset articles, for which Text, Coocrdiante and ID are used further.

Scaling and Harmonisation

The newspapers contained in this archive are far from uniform in size, ranging from the regular-sized dailies, which were typically 12 ¼ by 18 ¾ inches, to the much smaller *Pall Mall Gazette*, which was only 6 by 9 inches [Beals 2018]. Additionally, even if the pages were the same, the idiosyncrasy of the digitisation process mean that a page might have gained or lost an inch or two to the binding or simply to the way it was placed on the scanner. Compounding this issue is the fact that the coordinates of an article are in pixels, not in inches, and there is no universal translation between these two units. The reason for this is simple: the coordinates were never intended by the archive creators to be used this way. All it was designed to do was to provide a visual indication to the user where on the digitised page the article they were looking at was located, either in a preview thumbnail or in an image viewer. This means these coordinates were always intended to be image-specific, and there was no need for uniformity of any kind between the images. In a similar vein, the choice halfway through the project to use higher resolutions took place for the sole reason of providing better segmentation and OCR. However these choices were made, they shape the archival reality researchers have to deal with [Fyfe 2016]. 12

Both of these issues need to be addressed in order to compare different articles on different pages with one another. First, the frames of reference in which the coordinates of the article exist need to be normalised; that is, both coordinate systems need to be given the same meaning. In the case of the *British Library Newspapers* parts I and II, there is a disparity in resolution, with part I images being 300 dpi and part II at 400 dpi. This provides us with an unworkable situation for visualisation if the goal is to compare an entire print run of a paper or even multiple different newspapers for a year or decade, as these may be from different parts of the archive. Using the image size and dpi information of the page from which an article comes, we transpose the coordinates of the article onto an “ideal” newspaper page at a resolution of 400 dpi that is the same for all articles, by multiplying the 300 dpi coordinates with 1.33. Next, the result is scaled non-isometrically to fit the ideal page's width and height, as shown in Figure 3. A separate X and Y scale are used for this, to assist in columns over different pages aligning with each other. The size of the ideal page was chosen so it has an aspect ratio that is equal to modern A4 for easier printing. 13

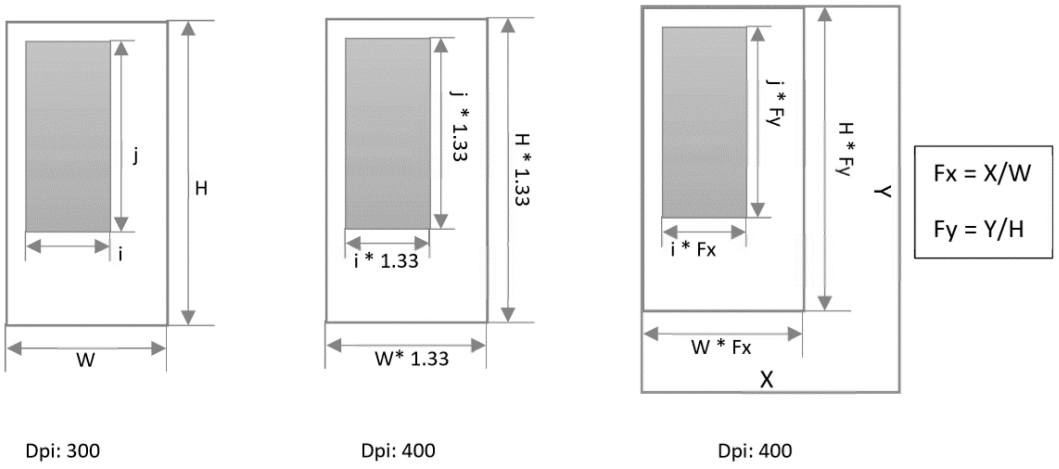
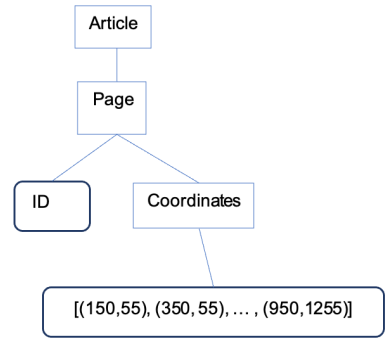


Figure 3. Stages of Normalisation and Scaling. Left: Original Page. Centre: Normalised for resolution. Right Non-Isometric Scaling for size.

Table-of-Occurrence Generation

Once the pages have been normalised and scaled, and the coordinate systems harmonised, it remains to generate the table of occurrence from which the heatmaps are drawn. These are the values that inform the intensity of the heatmap, and represent the amount of articles that occupy any given space. For this we need to determine the coordinates of each pixel within the article's area. It was found that using every pixel was impractical and unnecessary: it increases computation time and memory used substantially, but when analysing it offers no additional benefit. The tool therefore uses 200 by 200 pixel blocks in its calculations, as it saves significant resources. These represent an area of 0.5 inch square in physical terms, which is small enough to show the detail we need, but big enough to not squander computer time. 14

At this point in the process, we have generated a collection of articles, with each article containing all the necessary information needed to generate the occurrence tables: a harmonised and unified set of coordinates covering each point within the bounds of an article, with each rectangle assigned to their correct page, for each article in the visualisation corpus. An example instantiation of this form is illustrated in Figure 4, realised for an article consisting of a single column on page 3, covering the area from point (150,55) to point (950,1255). This process is repeated for every article. 15



X	Y	N	Page
150	55	1	3
350	55	1	3
…	…	…	…
950	1255	1	3

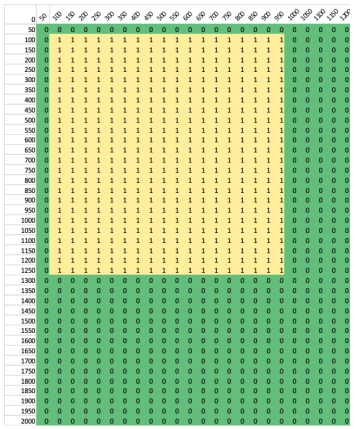


Figure 4. Example instance of a single page article stretching from 150,55 to 950,1255 during different stages of the visualisation process. Top left: schematic representation of the data structure. Top right: tabular representation of position data with number of observed articles. Bottom left: heatmap showing underlying number of observances and x- and y-coordinates.

Thus, this table has the X and Y coordinate of the 200-pixel square, the amount of times an article occurs in that square, and the page number on which the article sits as their own columns. The program iterates over the articles it needs to visualise, and looks up the space occupied by the article in the table. A model of the table is shown as Table 1 below. If an article has already been observed in that place, it increments the number of observations by one; if not, it adds a new row to the table with the article's position. 16

Row	Contents
X	X-coordinate of the 200 pixel square
Y	Y-coordinate of the 200 pixels square
No. Articles	Number of articles that occur in the square corresponding to these X and Y coordinates
Page	The page from which the article originated, as given in its archival metadata
Topic	Which topic in the topic model an article was assigned to. If assigned to multiple topics, the one with the highest certainty is chosen. Optional.

Table 1. Model of the table underlying article placement visualisation. Following the tidy data convention

Image Generation

With the data scaled and presented in a tidy format, it only remains to generate the images. A key problem is visualising multiple pages at the same time, while maintaining a uniform density scale between the graphs. If all pages were simply visualised on their own without such a precaution, each would default to a local scale, and the same hue on the heatmap could then indicate widely varying numbers of articles observed. Neither seaborn nor matplotlib were designed to support generating multiple heatmaps with a common scale, so a workaround using the manual scale settings had to be found. This necessitated discovery of the maximum number of articles observed before generating the heatmaps, in order to place a set value as the maximum, by generating the table of occurrence for all pages in the paper, and using the maximum value from the incidences column. This results in a sequence of heatmaps, which each represent a page, but which share a common colour scale. An example of this is included as Figure 5. 17

Various forms of colouration were experimented with. The initial visualisations used a schema of increasingly dark shades of a single colour. This was highly effective at showing the areas with the strongest presence of articles, but it was found that in places where there were only slight variations in the number of article occurrences, such as on page 2 or page 5 of the visualisation in Figure 5, the slight variation in shade was not easily spotted. For these visualisations a three-tone colour scheme is advisable, as it combines ease of interpretation with an aesthetically pleasing form. The bleedmaps generated for this paper all use a yellow-green-blue colourmap. 18

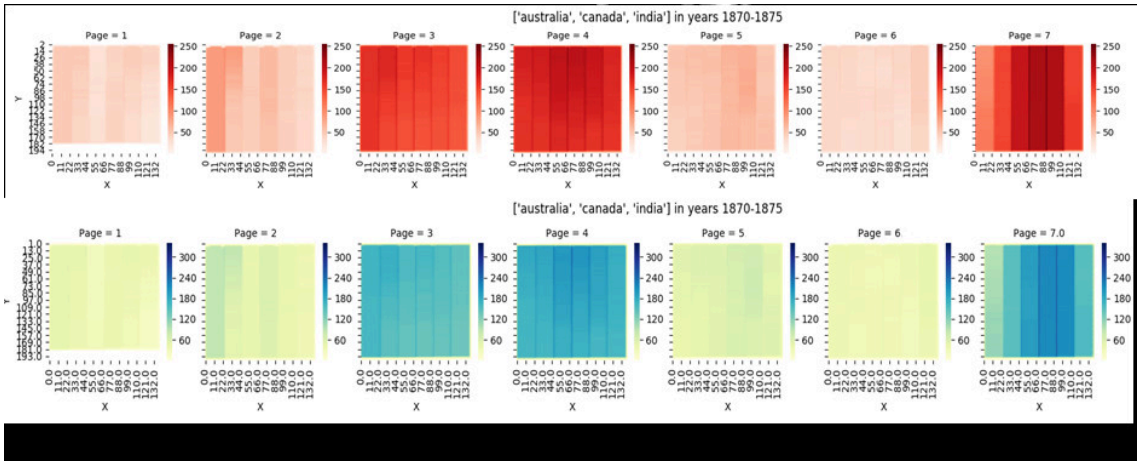


Figure 5. Article placement for three “Imperial” keywords in Reynold's Newspaper for 1870-1875 shown to illustrate interpretative difference in colour. The variance in colour scheme is particularly noticeable for pages where the number of observed articles are close to each other.

Reading these heatmaps is simple and intuitive. Each square heatmap represents a page, which is compressed into a unified scale. This is more apparent in the vertical than in the horizontal, as this makes the columns more pronounced. The columns on the page form by themselves, as the article placement data of each overlaps slightly with the adjacent column. This has the effect of creating a darker area delineating the column. The origins of this overlap lie in the segmentation undertaken during the digitisation process, which appears to have defined the article zoning with relatively wide margins. This is a direct result of the image being slightly curved or misaligned on the scanner, while the zoning is limited to drawing perfect rectangles between its topmost left and bottommost right points. The relative inaccuracy of such an approach would not impact the goal of using the zoning to highlight areas in a preview window. Occasionally, this can lead to artefacts forming when the newspaper changes its layout within a subset; during tests with *Reynolds' Newspaper* it was found that at one point between 1885 and 1889 that paper changed over from a six-column to a seven-column layout, which distorted the image. This in itself was a surprise, as the literature has described *Reynolds'* as an eight-column newspaper for its entire existence [Shirley 2009]. 19

In these bleedmaps, a darker hue of blue represents a stronger presence of the subset in that space. In the example above, the strongest concentration is in the bottom-right corner of page 8, with a medium concentration on pages 3 and 4. All pages have some level of article occurrence; if there are no articles (the observance count is zero), the space on the image would be white (for example at the bottom edge of page 1). In practice, these snapshots, be they per year, per decade, or per month, can be stacked on a page, showing the way focal points of keywords move through the title in their repetition. 20

A case study: Renolds' Newspaper

Having laid out the process of generating and interpreting Bleedmaps, it now remains to make a case for their usefulness as a means for research in digitised historical newspapers. To do this, a compact historical case study is in order, showing the difference in the findings attainable by commonly-used textual analysis methods such as LDA Topic Modelling and Corpus Analysis using Antconc, and the more metatextual approach offered by Bleedmapping. This difference in analytical level means this is not an either-or comparison, but rather a way in which Bleedmapping may supplement other techniques. The case study will center around the hypothetical research question: how did *Reynold's Newspaper* report on imperial news compared to foreign news between 1850 and 1900? 21

The data underpinning this comparison will be drawn from Gale's Legacy text Mining Drives [Beals and Bell 2020]. For this comparison, we need two datasets generated by keyword search, and one composed of random articles, for a total of three. India, Canada and Australia are used as markers of empire. These were chosen because they were three of the major possessions of the British Empire, spread around the three major continents on which it held territory, and they related to colonies in different stages of development and of different types: Canada as a developed settlement colony, Australia as a pioneer settlement colony, and India as the “crown jewel” of the empire. Earlier tests to find suitable search terms for Britain's African holdings proved unsuccessful, as there the geographic descriptors were too fluid. For the comparison with foreign news, three imperial competitors were chosen. France, Britain's oldest enemy and competitor in Africa, Russia, which competed with Britain for influence in Central Asia, and Germany, which began to challenge British naval power towards the end of this period, but with whom Britain had strong dynastic ties in the earlier part of the century. These are respectively the imperial and foreign 22

datasets.

Starting on a metatextual level with the bleedmaps, both datasets were divided into five-year slices to ease computational load and to improve the level of detail with which they reveal change over time. This will provide the spatial context in which the semantic elements that emerge through corpus analysis and LDA exist. Of course, understanding this spatial context comes with a major caveat: there has been very little research into the layout and visual language used in nineteenth-century newspapers. As modern scholars used to certain stylistic tropes we need to be cognisant of the fact that we may not understand the importance of certain cues embedded in the paper's layout, or interpret them incorrectly. For example, we may consider, in our twentieth- or twenty-first century mindset, the front page the most important part of a paper - but this need not have been the case for a Victorian reader who would first see the colourful advertisement wrapper. 23

This being said, there is still great value in exploring the pages of *Reynold's* as a space in a general sense, before involving semantics. Figure 6 shows two selected pages; from these we can see *Reynolds's* went through a redesign in 1885/86, changing from a six-column layout to a seven-column design. This was completely unexpected, as the literature describes the paper as being laid out on eight larger pages and eight columns [Shirley 2009]. A verification using the images themselves confirms the presence of six and later seven columns. Having more columns allowed the editor more options when composing the various articles into a coherent page and made more room for adverts, though at the cost of column width. 24

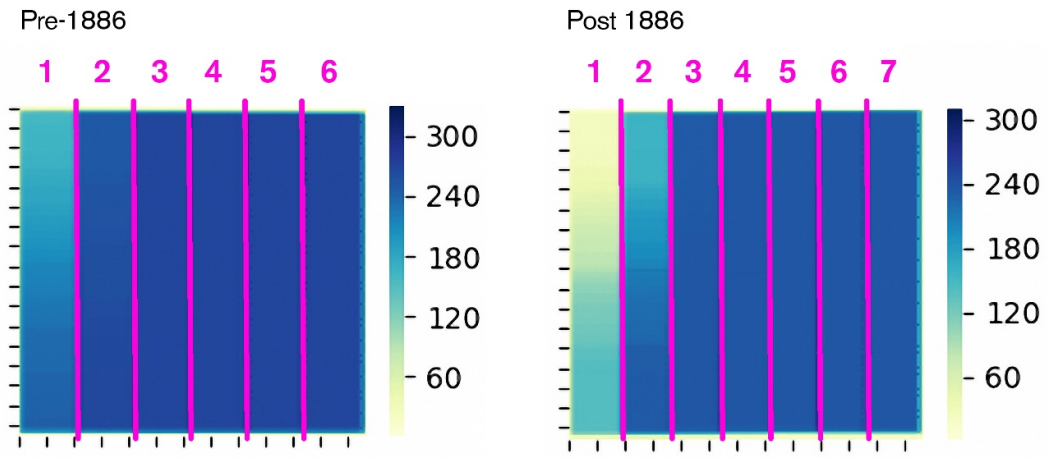


Figure 6. Columns in Reynolds Newspaper before and after the 1886 redesign. Only page 7 is shown. This illustrates the difference in number of columns observed.

During second half of the nineteenth century covered here, *Reynolds's* never switched to a columnless or modern layout. However, this does not mean it was without an underlying design philosophy: despite the appearance of disorder and immutability of newspaper design in the second half of the nineteenth century, rational and bureaucratic design elements did emerge. The centre and the periphery were made visual on the page by dynamic whitespacing, which had the most central and important content in airy, double-spaced lines on the top left and centre columns, with the density of the text increasing towards the bottom-right, where more peripheral content was placed [Barnhurst and Nerone 2002]. However, more recent newspaper scholars such as Liddle [Liddle 2012] have argued that Victorian newspapers were very much in flux, with the information density, and by necessity, organisation, of their pages changing throughout the century. He states that the pages and genres within them only stabilised during the latter decades of the nineteenth century. If this is the case, analysing article placement would be of little value, as it would show no underlying structure. Additionally, whatever the editor intended the placement of content to be, the final say on the allocation of space in a newspaper was reserved for the foreman at the printers. Usually, until the 1870s his task required cutting back material or cramming in content wherever it would fit. 25

However, based on the Bleedmaps produced here (IMG), we may conclude that *Reynolds Newspaper* did possess a consistent layout for the entirety of the period investigated. This conclusion derives from the clearly present clustering of articles for both keyword-selected subsets: if articles that contain the same keywords, and thus cover the same or similar topics, appear in similar places over a long period of time, we can safely consider there to be evidence for a “rational and bureaucratic” design being imposed on the paper. The way these clusters shift shows that after 1886 the paper was redesigned in such a way that the political content, both foreign and imperial, occupied a very different space, barring one major exception. The most obvious section where the imperial and the foreign subsets overlap is in the first two columns of page 4 after 1886, which contains the densest concentration for both these families of article by far. With between 250 and 300 articles, these rows are hotbeds of occurrence. 26

Their appearance becomes even more intriguing once the space is given proper context: these columns housed the very popular “Notices to Correspondents” section of the paper. These kinds of sections have been theorised as “the principal forum for reader opinion”, and the 27

space for items that have “earned their legitimate place in the public debate” [Richardson 2008]. Started by *Reynolds's* as a way to connect with his audience, readers could send questions, both on the mundane and the political to the newspaper's offices, which the editor (G.W.M. Reynolds himself until his death in 1879, subsequently his brother Edward and son William) would respond to. “Notices to Correspondents” helped create a community of readers and was an integral piece of the paper's formula. The change we observe in the placement of political content in general, but of foreign political content in particular, from adverts, economics and news sections to these discussion pages is suggestive of a deeper change in the way newspapers were read in the nineteenth century.

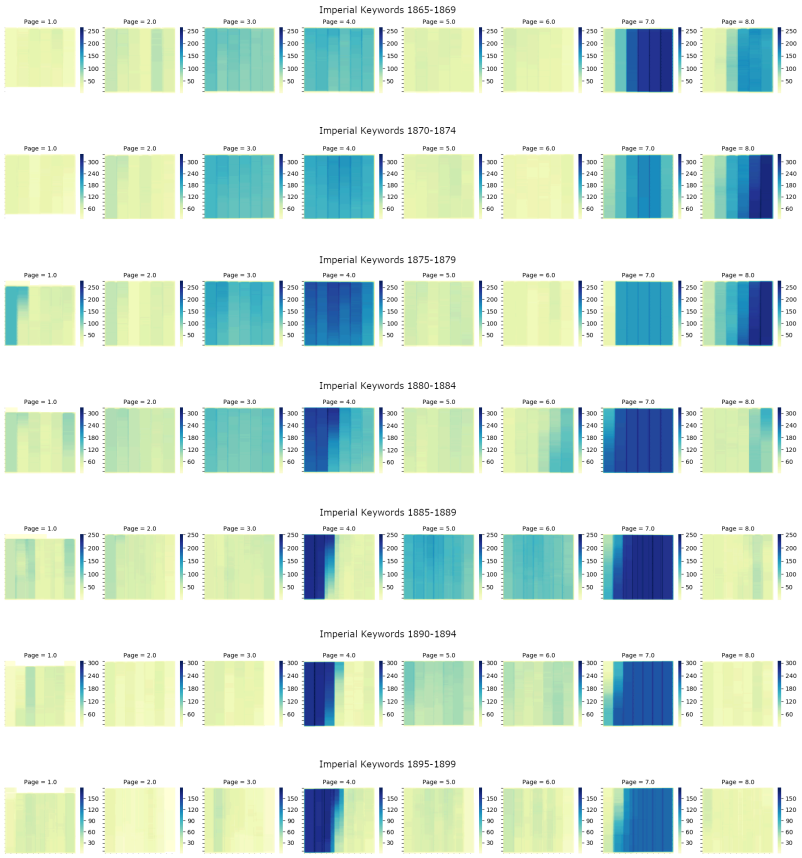


Figure 7. Imperial article placement in Reynolds Newspaper visualised in five-year intervals (January 1st of the first year to December 31st of the last). Keywords used: “India”, “Canada”, and “Australia”. Note the prevalence of columns 1 and 2 on page 4 after 1885, representing the imperial debate in the “Answers to Correspondents” columns.

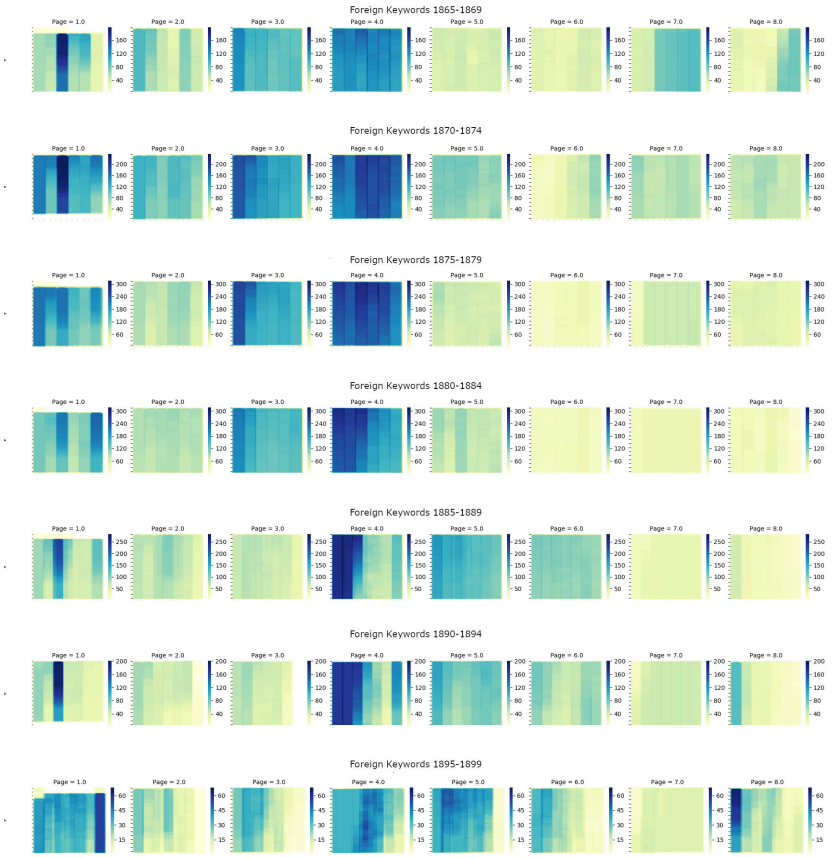


Figure 8. Foreign article placement in Reynolds Newspaper visualised in five-year intervals. Keywords used: “France”, “Germany”, and “Russia”. Note the presence of foreign news on the front page compared to imperial keywords.

The dominance of “Notices to Correspondents” in the discourse of the empire and the foreign in *Reynold's* is remarkable; it outperforms its closest competitor, the advertisements on page seven, by a wide margin. But making note of this is as far as computer-supported research methods can take us: closer reading (or a targeted topic model) is needed to show the ways in which these places were referenced. It showed two contexts in which both the empire and the foreign Other were used. First, they may be invoked in direct response to a question from a reader. For example, in response to a query by “W.G” in 1851: “Theatrical managers in France are made to pay one-tenth of the receipts to the support of the poor.” Or in October 1852, in an answer to A. Milner of Glasgow, when the response reads “If your health, age, character, &c., is such as to meet the view of the Government Emigration Commissioners, a person of your calling might obtain a free or assisted passage to Australia.” In both cases, the original query obviously specified that they were inquiring into something foreign or imperial—in this case the payment into a social security system by French theatre owners or the ways for “a man of certain calling” to make it to Australia. 28

The second invocation of imperial and foreign places in the editor's responses is as a yardstick by which a certain factoid or measurement is to be taken. In these cases, the enquirer is asking after foreign practices to compare them with those within Britain. These are much more difficult to identify, as without the original query, they can look much like the simple queries, and a degree of close reading is required to find them. One example is the response to a question on the use of colonial troops by the French by “M.E.” of Wye: “The Zauaves [sic] are natives of the French provinces of Algiers, disciplined and exercised by French officers, and now forming part of the French contingent employed in the Crimea. They hold exactly the same relation to the French army as the Sepoys in India have to the regular British troops.” In this case, the editor recognised that by making a comparison between an unknown foreign entity and a known imperial one, the reader would better understand the former. 29

While newspapers offered a forum for public discussion since their inception, it has been theorised that the rise of “new journalism” saw these interactive practices reach new heights [Barker and Burrows 2002] [Jackson 2001]. The bleedmaps confirm those readings, as they show the newspaper's shift from being a vessel for news and information about the political developments of the world to being a platform for debating those politics, which suggests a substantial emotional investment by the average Britton into the empire. However, we could also read this as an example of the agenda-setting powers of the press: starting and fostering debate on topics in a way that forces political parties to respond. Said debate has, of course, shades of chickens and eggs, but nonetheless Bleedmaps may contribute to this debate by providing clues to the metatextual background of the words used to debate and discuss. 30

The way in which *Reynold's* places political imperial news suggests that it is placed alongside or amongst articles covering national matters. 31

The articles that are keyword-selected with imperial keywords, and occupy political spaces in the newspaper, do so mainly in the parliamentary columns. Here, the empire is discussed as an integral part of British political life. Additionally, news from parts of the empire, relating to (political) events that take place there are not reported separately. This suggests the empire does hold a special position in British political discourse, and in the political identities it imparts on its citizens. It is not merely an overseas place, but one that, because of the power Britain holds over it, is a space that the British social, economic and political lives intersect with on a regular basis. While the foreign only appears when it is relevant, the empire is on every page, as it is always relevant to Britain's political debates.

32 These insights can be used to better understand the results of more textual analysis methods, such as a corpus analysis by tools such as AntConc or LDA topic modelling. For this case study, LDA topic modelling is an appropriate tool to extend the insights gained into the placement of articles: we have a space and a rough content based on the keyword search, but what more can we learn from them? For this purpose, we use a the Gensim implementation of LDA [Řehůřek and Sojka]. This model was constructed using the same data as was used for the Bleedmaps, divided in the same article-sized chunks. We now turn to the two important aspects that greatly influenced the quality of the resulting model, both of which underline the potential issues inherent in using digitised archival data “as is” and which show the importance of critical data literacy. These two are OCR quality and article segmentation.

33 Article segmentation influences the size (and quality) of the chunks that are used to build the model, and in the case of Gale's Legacy Text mining drives – and the British Library data these drew from – the creators decided to segment articles conservatively, that is, if a page contains six columns of ten adverts each, all these sixty adverts are considered one article. This was not an unreasonable approach to take when the digitisation took place around 2004–2007. Even today, on much more powerful hardware and using much more sophisticated techniques, article segmentation, especially of historic newspapers, remains an unsolved problem [Barman et al. 2021]. Additionally, a greater level of precision was not needed for the intended use as an index of the digitised contents of the newspaper collection [King 2005]. However, the choices made then poses a major problem for a document-level LDA model now, as it means that one of the fundamental assumptions underpinning it no longer holds true: that the writer of each text sought to produce sensible texts where words from the same topic-specific bag-of-words are used together more often than words from other topic-specific bags-of-words. Hence, topic models generated from this dataset are often barely coherent (ie. no meaning can be found in the collection of tokens) or badly fragmented (ie. multiple topics contain almost the same tokens).

34 Compounding the challenges facing the LDA model is the variable quality of OCR in the source material. While the OCR accuracy of this particular archive was excellent when it was digitised between 2004 and 2007 [Tanner et al. 2009], it falls behind when compared to modern standards [Kettunen and Pääkkönen 2016] [Breuel 2017]. The average character transcription error rate in these articles is acceptable for human readers and for search indexing, as this was what it was designed to facilitate, yet it poses significant challenges for LDA. After all, if most tokens are unique, how can a model learn the patterns that underlie their use? With all these caveats out of the way however, the topic model does provide some additional insights in the language that occupies the spaces identified by the Bleedmapping.

35 Topic modelling may, on this particular dataset, raise further research questions on how the various news items of the day interacted and intersected on the pages of Reynolds. It is notable that for the period 1860–69, for example, most topics contain tokens that refer to the American Civil War, such as “Confederate”, “Washington” or “American”. This is not in and of itself unexpected, as the Civil War was followed closely in the British press and loomed large in the public imagination [Grant 2000]. This interest was connected with deep-seated economic concerns about the war's impact on the global cotton supply, and whether supplies from India would be enough to replace unavailable or destroyed stocks. Due to the way articles are segmented in this archive, it is impossible to know if this is a legitimate reading of the topic model, or if it is simply caused by different imperial news articles on “India” being contaminated with larger (yet semantically separate) articles on the Civil War. This problem gets more acute the later in the century we go, as forms of reporting change. While in the middle of the century foreign and imperial reporting tends to be longform, by its end shorter telegraph bulletins are the norm.

36 Questions around this issue of “contamination” of articles by segmentation become even more complex as we consider another popular genre of article in Reynold's: the army and navy list. These articles are by their very design “contaminated” with mentions of places all over the globe, named without much if any contextual information. These lists of regiments of the British army and ships of the royal navy and their postings throughout the Empire form into topics with high probability, because they follow the same pattern for decades on end. These lists were not unique to Reynold's and were often a case of “scissors and paste journalism” [Beals 2018], where newspapers copied whole articles from other -more specialized- periodicals. Interpreting the meaning of these lists is difficult – were they solely intended to inform the loved ones of serving personnel [Jones 2018], or did they also instill a feeling of imperial pride? Whichever is the case, the lack of a clear hotspot on the bleedmap that could correlate with their placement suggests they were regarded as filler, to be placed wherever there was room left.

37 The topic models have given us additional insight into the textual contents of the page, as they are designed to, while they reinforce the findings of the Bleedmaps. While the technical constraints of this project were such that using them truly in unison was not feasible, it is no great stretch to imagine how there two can reinforce each other further. Starting from a topic model, a bleedmap might show the spatial context of each topic. This can ease the topic's interpretation by providing valuable metatextual clues. Alternatively, starting from a bleedmap, a topic model generated for a particular hotspot and modelling only those articles that appear in a (known) space in the newspaper, can be a valuable tool to further explore its textual contents. These improvements, as well as integrating direct web access

through api calls, are being included in the mature version of the Bleedmapper I am currently building for the Delpher newspaper archive.

Conclusion

38 In conclusion, this article has presented a method for visualising the spatial placement of digitised periodical articles, by creating a density map of their positional data. It has shown this method of *bleedmapping* is an improvement over previous (manual) visual analysis methods, as it boasts a high level of scalability and a low reliance on human data tagging, while still retaining human authority in the analysis step. These Bleedmaps are a useful addition to the toolbox of the digital historian when exploring collections of sequential publications, such as newspapers and periodicals. They provide metatextual context for other, text based analysis methods such as topic models, and can in future be made to show the location of individual topics for easier interpretation. The method does not, however, solve issues related to article segmentation and OCR accuracy completely, but it can help mitigate the impact of these known issues when applying methods that are more reliant on accurate segmentation and OCR, such as Topic Modelling.

39 Bleedmapping relies on the creative re-uses of pieces of the archive, which were never intended to be used as such. They show the density of articles that match a set of selection criteria on each page of the newspaper, which can then be linked back to certain categories of article through analysing these specific spaces. They are thus particularly valuable for understanding the patterns of repeated content in a periodical, which are known to possess this spatial identity. They are however limited in their ability to carry the burden of historical evidence alone: until we gain a fuller understanding of the meanings of article placement, they will have to be supplemented by other methods of analysis for the identified areas of interest, be it a closer textual reading or application of computational approaches. Bleedmapping has the major advantage of making the subtext of each article visible by showing the spatial context in which they exist.

40 These bleedmaps fundamentally open up new possibilities for research. Exploring article placement has only been done on a small scale, as it was time-intensive when using the prior existing methods. The development of this tool allows for overviews of article placement to be generated without human involvement. It makes it possible to answer questions about the space occupied by specific genres of content; it makes us ask whether these places were static over time or if articles changed places; it generates questions about the redesigning of the layout by incoming editors. It has the potential to be valuable for both the field of periodicals studies and for historians wishing to gauge the impact or prominence of certain content. However, in its current embryonic form, it suffers from a lack of secondary literature and historical theory to embed itself in. At present, there has been very little work done on the visual language of Victorian newspapers, the way their layout spoke to their readers. Did readers value the front page the same as we do nowadays, or did the presence of the advertising wrapper mean the key content would be most eye-catching on the inner folio? Did readers learn to expect certain patterns to their newspapers that editors themselves were restrained by, such as expecting the Parliamentary debates to be on page three? There are countless questions like this, but most of the answers are yet to be found. Additionally, theories on content placement in a historical context will need to be developed; those that are available for newspapers generally consider only more modern columnless newspapers, which makes them unusable for content such as that in this archive. Bleedmapping allows us to finally explore and theorise the pages of historical newspapers the same way we navigate our everyday reads: spatially.

Acknowledgements

41 I am indebted to the staff at Edge Hill University, specifically Bob Nicholson for his guidance and aid during the research for this article, always being available to bounce ideas off and putting up with a server in his office for two years. Mark Hall was invaluable for sorting out the access to Text Mining Disks Gale-Cengage graciously provided to the university and waded through my amateur attempts at coding to show me how it should be done.

42 The work of Melodee Beals was a great inspiration, and she kindly gave her time to listen to the outline for what would become Bleedmapping and provided feedback and encouragement. Further great feedback came from James Mussell and the attendees at DH2019 in Utrecht.

43 I am also thankful to Melvin Wevers and Mano Delea at Amsterdam University, who helped greatly in ironing out the kinks in the writing process and for providing support when needed. They were great at coming up with workarounds for the loss of access to the source data due to my switching institutions.

44 Lastly, I would like to thank the anonymous reviewers, for being the whetstone to sharpen my arguments, and the editors of this journal, for the polish and shine. Any imperfections or errors that remain are fully my own.

Notes

[1] A page is an image. It gives a total impression, presents a block or a system of blocks and strata, of blacks and whites, to a task of figure and intensity more or less joyous.

Works Cited

Barker and Burrows 2002 Barker, H., and Burrows, S. (2002) *Press, Politics and the Public Sphere in Europe and North America, 1760-1820*,

Cambridge University Press.

Barman et al. 2021 Barman, R., Ehrmann M., Clematide S., Ares Oliveira, S., Kaplan F. (2021) “Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers”, *Journal of Data Mining and Digital Humanities*.

Barnhurst and Nerone 1991 Barnhurst, K.G., Nerone J.C. (1991) “Design Trends in US Front Pages, 1885–1985”, *Journalism Quarterly,* 68(4), pp.796-804.

Barnhurst and Nerone 2002 Barnhurst, K.G., Nerone J.C. (2002) *The Form of News: A History*. New York: Guilford Press.

Beals 2018 Beals, M.H. 2018. “Close Readings of Big Data: Triangulating patterns of textual reappearance and attribution in the Caledonian Mercury, 1820-1840”, *Victorian Periodicals Review* 51(4), pp.616-639.

Beals and Bell 2020 Beals, M.H., Bell, E. (2020) *The Atlas of Digitised Newspapers: Reports of Oceanic Exchanges.* Available at: https://figshare.com/articles/online_resource/The_Atlas_of_Digitised_Newspapers_and_Metadata_Reports_from_Oceanic_Exchanges/11560059/2.

Billig 1995 Billig, M. (1995) *Banal Nationalism*, SAGE, London.

Breuel 2017 Breuel, TM. (2017) “High Performance Text Recognition using a Hybrid Convolutional-LSTM Implementation”, *Proceedings of the Fourteenth IAPR International Conference on Document Analysis and Recognition*, International Association for Pattern Recognition, Kyoto, pp.11-16.

Fyfe 2016 Fyfe, P. (2016) “An Archology of Victorian Newspapers”, *Victorian Periodicals Review*, 49(4), pp.546-577.

Gatos et al. 2000 Gatos, B., Mantzaris S., Perantonis S. and Tsigris A. (2000) “Automatic page analysis for the creation of a digital library from newspaper archives”, *International Journal on Digital Libraries*, 3, pp.77–84.

Grant 2000 Grant, A. (2000) *The American Civil War and the British Press*, Jefferson, NC: McFarland.

Haskell 1993 Haskell, F. (1993) *History and Its Images: Art and the Interpretation of the Past*, New Haven: Yale University Press.

Jackson 2001 Jackson, K. (2001) *George Newnes and the New Journalism in Britain, 1880–1910: Culture and Profit*, Farnham: Ashgate.

Jones 2018 Jones, H. (2018) “She Had Only Navy-Lists and Newspapers for Her Authority”, *Persuasions: The Jane Austen Journal On-Line* 39(1).

Kettunen and Pääkkönen 2016 Kettunen, K. and Pääkkönen T. (2016) “Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, European Languages Recources Association, Portoro, pp.956-961.

King 2005 King, E. (2005) “Digitisation of Newspapers at the British Library”, *The Serials Librarian*, 49(1–2), pp.165–81.

Liddle 2012 Liddle, D. (2012) “Reflections on 20,000 Victorian Newspapers: ‘Distant Reading’ The Times Using The Times Digital Archive”, *Journal of Victorian Culture*, 17(2), pp. 230–237.

Meier et al. 2017 Meier, B., Stadelmann T., Stampfli J., Arnold M., Cieliebak M. (2017) “Fully Convolutional Neural Networks for Newspaper Article Segmentation”, *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp.414-419.

Moreaux 2016 Moreaux, J.P. (2016) “Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment”, *IFLA News Media Satellite Sessions*, 8-11-2016.

Mussell 2012 Mussell, J. (2012) *The Nineteenth-Century Press in the Digital Age*, Basingstoke: Palgrave Mcmillan.

Oyallon and Rabin 2015 Oyallon, E., Rabin, J. (2015) “An Analysis of the SURF Method”, *Image Processing On Line* 5, pp.176–218.

Richardson 2008 Richardson, J. (2008) “Reader's Letters”, in Franklin B. (ed.) *Pulling Newspapers Apart*, London and New York: Routledge, pp.56–66.

Shirley 2009 Shirley, M. (2009) “Renolds' Newspaper”, in Demoor, M. and Brake, L. (eds.) *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland*, London: British Library, pp.539–41.

Tanner et al. 2009 Tanner, S., Muñoz, T., and Ros, P.H. (2009) “Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library’s 19th Century Online Newspaper Archive”, “D-Lib Magazine”, 15(7/8).

Valérie 1927 Valérie, P. (1927) “Les deux vertus d'un livre”, *Arts et Métiers Graphiques,* 1, pp.3-8.

Řehůřek and Sojka Řehůřek, R., Sojka, P. (2010) “Software Framework for Topic Modelling with Large Corpora”, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp.45–50.