Petri Paju, Heli Rantala, Hannu Salmi

Towards an Ontology and Epistemology of Text Reuse

Cycles of Information Flows in Finnish Newspapers and Journals, 1771–1920

Abstract: The article explores the ontological and epistemological ramifications of text reuse, drawing on a digitized corpus of newspapers and journals from the National Library of Finland and covering the time span of 149 years from 1771 to 1920. The article examines three types of reuse cycles, rapid, slow and mid-range repetition. The argument is that text reuse has ontological ramifications on how the processes of a media network are conceived. With ontology we mean that the study of history always includes conceptualizations, either implicit or explicit, of which kinds of entities and things, as well as forms of being, there were in the past. Text reuse offers a perspective for the analysis of these "forms of being." In the epistemological part of the study, the article studies the aspects that influence and may bias the results, focusing on the material conditions of digitization process, the problems of metadata, and the possible methodological nationalism of drawing on nationally siloed corpora.

Keywords: text reuse detection, newspaper history, media history, ontology, epistemology

1 Introduction

Text reuse detection is an expanding field of research. It offers possibilities for exploring large text corpora and the circulation of texts and text passages. It has been successfully employed, for example, in the study of how authors have quoted previous literary works, such as ancient classics.[1] It has also proved to be an efficient tool in understanding how newspapers share each other's contents.[2]

1 M. Büchler, G. Crane, M. Moritz, and A. Babeu. 2012. "Increasing recall for text re-use in historical documents to support research in the humanities." In: (Proceedings) Second International Conference on Theory and Practice of Digital Libraries, vol 7489, pp. 95–100. DOI: 10.1007/978-3-642-33290-6_11.

2 D. A. Smith, R. Cordell, E. Maddock Dillon. 2013. "Infectious texts: Modeling text reuse in nineteenth-century newspapers." In: Proceedings of the Workshop on Big Humanities. IEEE Computer Society Press 2013, 86–94, http://www.ccs.neu.edu/home/dasmith/infect-bighum-

https://doi.org/10.1515/9783110729214-012

In this article, we explore the ontological and epistemological ramifications of text reuse. The article draws on the research project *Computational History and the Transformation of Public Discourse in Finland* (Academy of Finland, 2016–2019), in which we studied text reuse in the Finnish press.[3] Our work is based on a digitized corpus of newspapers and journals from the collection of the National Library of Finland, covering a time span of 149 years, from 1771 to 1920. The advantage of the Finnish corpus is that it is complete, including in principle all published titles and their issues up to the year 1920. The corpus consists of all kinds of serial and periodical publications which in principle could reprint texts from each other. By drawing on both newspapers and journals we wanted to emphasize a comprehensive view on periodical press. From the perspective of Finnish history, the period is also essential: until 1809, Finland was part of Sweden, and after the War of 1808–1809, it came under Russian rule. Finland became an independent state in 1917. Throughout the timespan covered in this study, periodicals in Finland were published both in Swedish and Finnish.

Although the reprinting of particular texts in a range of different locations can be regarded as an old and well-acknowledged practice, a systematic examination of this phenomenon has not been possible until the digitization of the press. Our primary research material derives from the digitized and OCR'd corpus, including newspapers and journals, published by the National Library of Finland, in sum 5 million pages. For this project, we developed our own solution for text reuse detection, which we called *text-reuse-BLAST*. It was based on NCBI BLAST (National Center for Biotechnology Information Basic Local Alignment Search Tool), originally developed for detecting similarities in biomedical sequences. As a software originally created for comparing and aligning DNA and protein sequences, text

2013.pdf. H. Rantala, A. Nivala, H. Salmi, P. Paju, R. Sippola, A. Vesanto, and F. Ginter. 2019. "Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920. Digitaalisten ihmistieteiden näkökulma." Historiallinen Aikakauskirja 1: 53–67. H. Salmi, A. Nivala, H. Rantala, R. Sippola, A. Vesanto, and F. Ginter. 2018. "Återanvändningen av text i den finska tidningspressen 1771–1853." Historisk tidskrift för Finland 1: 46–76. H. Salmi., P. Paju, H. Rantala, A. Nivala, A. Vesanto, and F. Ginter. 2020. "The Reuse of Texts in Finnish Newspapers and Journals, 1771–1920: A Digital Humanities Perspective". Historical Methods: A Journal of Quantitative and Interdisciplinary History, https://doi.org/10.1080/01615440.2020.1803166.

3 The project was a large consortium, including researchers from the Universities of Turku and Helsinki and the National Library of Finland. The text reuse study was realized as a cooperation between historians (PI Hannu Salmi) and data scientists (PI Tapio Salakoski) at the University of Turku. The group included Filip Ginter, Asko Nivala, Petri Paju, Heli Rantala, Hannu Salmi, Reetta Sippola, Tapio Salakoski and Aleksi Vesanto. The software development was done by Aleksi Vesanto under the supervision of Filip Ginter.

reuse detection with BLAST is character-based – not word-based as in many previous text reuse detection solutions – and it requires preprocessing of the material so that the letters are encoded into the alphabet of 23 amino acids. BLAST was originally designed to be highly applicable to material that includes a substantial amount of noise. The tolerance of noise was significant in processing the Finnish corpus of newspapers and journals, which includes many OCR errors.[4] In the end, text-reuse-BLAST proved to be effective in recognizing textual similarity. From the whole corpus, we found 61 million hits or occurrences of similarity, which formed 13.8 million clusters of text reuse. These clusters are simply chains of similar passages of text that were initiated between 1771 and 1920 in Finland. At the outset, we set the minimal threshold for similarity to 300 characters.

This article analyzes the different cycles of text reuse within this material, especially their ontological and epistemological aspects. In an earlier work, we identified and studied long-term text reuse[5] and presented the main findings of the project in two publications.[6] Cases of long-term reuse were those in which texts were reprinted after a gap of 50 years or more. Some repetition chains were very long and slow, lasting over 140 years. In contrast to this, there was also very short-term, rapid, viral circulation of texts. Virality was manifested when the same text spread across the network within a few days or weeks. The duality between slow and quick repetition is challenged by the significant amount of what we call mid-range repetition, characterized by texts that were copied infrequently within 5, to 10, or even 20, years. Our argument is that text reuse has ontological ramifications on how we conceive the past processes of a

4 On this methodology, see A. Vesanto, A. Nivala, H. Rantala, T. Salakoski, H. Salmi, and F. Ginter. 2017. "Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910." In: (Proceedings) 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017 (Linköping 2017), http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf. An updated discussion has been published in Salmi *et al.* 2021. For source code, see https://github.com/avjves/textreuse-blast.

5 H., Salmi, H. Rantala, A. Vesanto, and F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf.

6 Rantala H., H. Salmi, A. Vesanto, F. Ginter. 2019. "Tekstien pitkä elämä: Ajassa liikkuvat tekstit suomalaisessa sanomalehdistössä 1771–1920." Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41858179. Salmi, H., P. Paju, H. Rantala, A. Nivala, A. Vesanto, F. Ginter. 2021. "The Reuse of Texts in Finnish Newspapers and Journals, 1771–1920: A Digital Humanities Perspective". Historical Methods: A Journal of Quantitative and Interdisciplinary History 54 (1): 14–28, https://doi.org/10.1080/01615440.2020.1803166.

media network. In information science, "ontology" refers to a model and an abstraction based on certain concepts, and often to a data structure. Our point of departure emphasizes a complementary, philosophical approach. With ontology we mean that the study of history always includes conceptualizations, either implicit or explicit, of which kinds of entities and things, as well as forms of being, there were in the past. Text reuse offers a perspective for the analysis of these "forms of being." Simultaneously, it is important to concentrate on the epistemological aspects of text reuse: our conditions of knowing about those "forms of being."

This article examines the three cycles of text reuse, rapid, slow and mid-range repetition, as entities that shed light on newspaper activity in the analyzed time frame. Thereafter, we focus on the epistemological aspects of text reuse and concentrate on three perspectives, on the material conditions of both newspaper publishing and digitization process, the problems of metadata, and the possible methodological nationalism of drawing on nationally siloed corpora.

2 Towards an Ontology of Reuse

In our analysis, we were able to detect 13.8 million text reuse clusters.[7] Since the Finnish OCR corpus is unsegmented, the process of text reuse detection was made on all newspaper content types, including news but also advertisements, anecdotes, obituaries, timetables, announcements, and any other items. These reuse clusters were not necessarily cases of conscious reuse from paper to paper. In this kind of a computational analysis, it was not possible to identify external sources, like letters sent to several editorial offices at the same time. For us, clusters were simply chains of similar passages. These chains were different in nature: in the present corpus, the shortest clusters included only two or three hits, but there were also long chains of repetition. These clusters represent flows of information from the late eighteenth century to the early twentieth century, and thus offer the possibility to explore the ontology of newspaper printing.

During the project we gradually separated the mass of reuse clusters into three main categories which broadly correspond and describe different cycles of reprinting. These categories help us understand and analyze the vast number of clusters. We first concentrated on rapid, short-term circulation and slow, long-term circulation. Within the rapid circulation, special attention was paid to exploring and measuring viral repetition. Finally, we included mid-range circulation to cover the

7 Text reuse data can be explored via a search interface at http://comhis.fi/clusters.

wide variety of reprinting between rapid and slow circulation. Table 1 presents key features of these three cycle types, forming our ontological considerations.

Importantly, these cycles of reprinting are not all mutually exclusive. A reprinted text could first spread very rapidly, "go viral", and then get reprinted again after a considerable time although this was untypical.

Table. 1: Distinctive features of the three reprinting cycle types.

Reprinting cycles	Rapid repetition (including viral circulation)	Mid-range repetition	Long-term repetition
Timespan	within a year; in viral cases: days or weeks	several years, ranging from 1 to 49 years	50 years or more
Share of all clusters	85,29 % (11,768,371 clusters)	14,67 % (2,023,615 clusters)	0,04 % (5,888 clusters)
Movement	synchronic, geographic (fast, incl. viral cases)	diachronic (both regularly and sporadically repeated texts)	diachronic (slow)
Contents, typically	wide range, from ads to news (incl. boilerplate)	wide range, from announcements to religious stories (incl. boilerplate)	news, literary works

2.1 Rapid Circulation of Information

A distinct category of text reuse is the rapid circulation of information. In the time period studied, this was mainly horizontal: news traveled in the geographical space within a short period of time. In the Viral Texts project, Smith, Cordell, and Maddock Dillon analyzed nineteenth-century American newspapers and divided the reprinted texts into fast and slow ones. In their fast set, the median lag time was under one year.[8] In our findings, most of the chains were also realized within twelve months, which comprised 85 percent of all reuse clusters, and these can be seen as fast repetition.[9]

8 Smith, D.A., R. Cordell, E. Maddock Dillon. 2013. "Infectious texts: Modeling text reuse in nineteenth-century newspapers." In: (Proceedings) Workshop on Big Humanities. IEEE Computer Society Press 2013, 86–94, http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf, 93.

9 Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019.

We can define texts that spread within twelve months as being "fast," but not all of these will be viral texts. In present-day discourse, the word "viral" is regularly used in connection to social media, in describing the rapid sharing of media content: rapidity is its essential feature. The roots of "virality" are obviously in microbiology and medicine, and the concept refers to the capacity of viruses to replicate in host cells and cause diseases in an epidemic curve. In public discourse, "viral" suggests that something is "like a virus" or "spreads like a virus."[10] Although the concept has a contemporary undertone, it can also be applied to historical material: in the nineteenth century, media capacity grew exponentially. In Finland, there were so many publications all around the country towards the end of the century that it created a resonance base for viral information flows. We decided to measure this capacity by paying attention to printing locations around the country and the number of newspaper titles.

To be able to filter the results, we defined a *virality score*, reflecting the diffusion of the clusters and measuring both rapidity and capacity (in relation to the volume of the press and to the geographical coverage of the cluster). This score is calculated by multiplying the number of different publication titles within a cluster and the number of unique printing locations by the inverse of the elapsed time. Here, "elapsed time" means the length of the cluster in days. The virality score helped us to capture how many titles the information spread into and how many places it geographically reached, and then penalized the value the slower the process was.[11] After this, the values of clusters were normalized into the range 0–100. The most viral text proved to be a paid advertisement by the Finnish tobacco industry against American cigarettes, published between 1 and 31 March 1916 in 45 different newspapers or journals in 26 different locations, 75 times altogether.[12]

The use of a virality score in our study was an experiment in measuring virality in a historical setting. It can be further refined into a tool that helps in understanding the rhythms of information flows and how they saturated the prevailing media system. Of course, this can be done only within the limits of

---

Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf, 394–404.

10 Oxford Dictionary of English. 2015. 3rd online edn. Oxford: Oxford University Press. DOI: 10.1093/acref/9780199571123.001.0001. On virality, see H. Salmi, 2020. *What is Digital History?* Cambridge: Polity, 22–25.

11 The virality score is discussed more in detail in Salmi *et al.* 2021. For the code, see https://github.com/avjves/cluster-viral-score.

12 Cluster no. 11592519, http://comhis.fi/clusters/?f%5Bcluster_id%5D%5B%5D=11592519, accessed June 27, 2020.

the OCR corpus, which means that the possible external sources of information could not be automatically detected. The score, however, helped in the qualitative analysis of results. In the case of the most viral texts, like the tobacco ad, it might be questionable whether paid advertisements, the impetus of which came outside of the media sphere, can be regarded as viral cases at all, but on the other hand this was illuminating from the perspective of an ontology of text reuse. In our material, only 81 clusters had a virality score higher than 50. These cases also include journalistic content, but almost all of them seem to have had sources outside newspaper editorial offices. It is obvious that media capacity enabled the viral spread of news, and this shift was understood by contemporaries who took advantage of the efficiency of the printing press.

An alternative way of approaching virality in the study of nineteenth century press could be to concentrate on reachability, or cluster spread. This would mean emphasizing geographical coverage and scope of the movements of reprinted texts within a chosen window of time (say, one year). In the project we thought about this option too but chose to concentrate on rapidity of reprinting. It might however be that reachability could better reflect the conditions of the printed press before the late nineteenth century, and in further studies it might be especially promising when studying border-crossing flows of news. Exploring the geographical spread of clusters as complementary to rapidity could improve our understanding of virality as a historically changing phenomenon.

2.2 Long-Term Cycles of Reuse

One of the valuable characteristics of the Finnish digital newspaper corpus is its long timespan. The corpus offers a view into Finnish newspaper publishing in its full scale from the very first publications of the 1770s to the era of wide national coverage of the press in the 1920s. In the beginning of our project, we did not really have any hypotheses on the temporal scale of text reuse. In this respect, it was a surprise that the longest chains of reuse covered over 140 years, almost the whole timespan of the corpus. We have analyzed this feature of the press, the so-called long-term reuse, in detail elsewhere.[13]

13 Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf. For an article in Finnish on the results, see H. Rantala, H. Salmi, A. Vesanto, F. Ginter. 2019. "Tekstien pitkä elämä: Ajassa liikkuvat tekstit suomalaisessa sanomalehdistössä 1771–1920." Ennen

The discovery of long-term reuse – 50 years or more – reminds us of different temporalities that have been present in the press. In addition to news material and editorial texts, the papers have comprised a great variety of repetitive content, such as advertisements and announcements. This is not a Finnish phenomenon as such, but we are not aware of other studies that have focused on the practices of slow publishing or on the reuse of texts within as long a time span as a century.

In the case of long cycles of reuse, the time lag between the first appearance and republication can be counted in several decades. Sometimes, although rarely, the span of reuse is even one hundred years or more. This means that the papers have benefited from material from the earlier publications and republished both extracts and full stories or news from the old newspapers. Thus, past publications have been used as an archive of possible material for contemporary newspapers. With its historical depth, the press participated in constructing a sense of community. Referring to Pierre Nora's notion *lieux de mémoire*, we have earlier suggested that the press could be understood as a site of memory.[14] On the whole, this type of text recycling is not a dominant form of reuse in the Finnish press but it is worth attention. If rapid reuse was often horizontal movements in geographical space, long-term reuse was vertical or diachronic: information that traveled in time. This does not mean that long-term reuse would realize without geographical spread, but time seems to be its dominant property: this includes cases where the same item has been republished by the same paper later on, again and again, and cases where the geographical spread is not characterized by the saturation of the capacity of the press but evolved rather randomly in time.

We discovered that, within these long-lasting cases of text reuse, there is variation in the actual cycles of republishing. While some texts had been reused only once or twice during a very long period of time, there have also been cases in which the old text was activated several times during different decades. One

ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41858179.

14 For further details, Rantala, H., A. Nivala, H. Salmi, P. Paju, R. Sippola, A. Vesanto, F. Ginter. 2019. "Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920. Digitaalisten ihmistieteiden näkökulma." Historiallinen Aikakauskirja 1: 53–67; Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf. See also P. Nora. 1997. "General introduction: Between memory and history." In (P. Nora, ed.) The Realms of Memory: Rethinking the French part, Vol. 1: Conflicts and divisions. Columbia University Press: New York, 1–21.

clear reason for this kind of cyclic reuse is anniversaries of certain culturally and/or politically loaded events. For example, in 1863 the gathering of the Finnish Estates was a major event since the Emperor of Russia had not summoned the Diet in over 50 years. Many Finnish newspapers published the opening speech of the Tsar at the Diet meeting. During the following decades, the same speech was republished every now and then by several papers. Then, in 1913, 50 years after the historical Diet meeting, the words of the Emperor were republished by over 40 Finnish newspapers to honor the anniversary.[15]

Apart from the above-mentioned remembrance of different national anniversaries or other important dates, newspapers and journals recycled a variety of old material. Among the clusters of long-term reuse, there are several anonymous stories or anecdotes, as well as old news clippings, which have probably been reprinted for the sake of amusement and curiosity. Furthermore, we have found examples in which the reprinting of old material had a clear connection to the contemporary culture, for example, of the political life of the country; for censorship reasons, it was sometimes impossible to describe the current state of affairs, but contemporary concerns could be implied by reprinting old content.[16]

On the whole, the slow cycles of reuse represent only a very small amount of texts of the corpora in question (see Table 1). This feature of the press is nevertheless interesting and offers the possibility to rethink those temporal scales in which the newspapers have operated. The existence of publishing cycles that have covered a century or more demonstrate the manifold functions of the press. Along with the topical functions, the press has many other scopes, including those that strengthen our ability to remember the past.

2.3 Mid-Range Text Circulation

In addition to rapid circulation and long-term use of texts, text reuse detection also revealed clusters that escaped both categories, or did not completely fit into either of them. We call this gray area *mid-range repetition*. The mid-range cycles comprise a significant share of total reuse cases. Their amount can be measured in different ways: it can range from anything over a year to those just under 50 years, thus covering a variety of reuse cycles.

15 Cluster no 6216311, http://comhis.fi/clusters/?f%5Bcluster_id%5D%5B%5D=6216311, accessed June 27, 2020.

16 Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf.

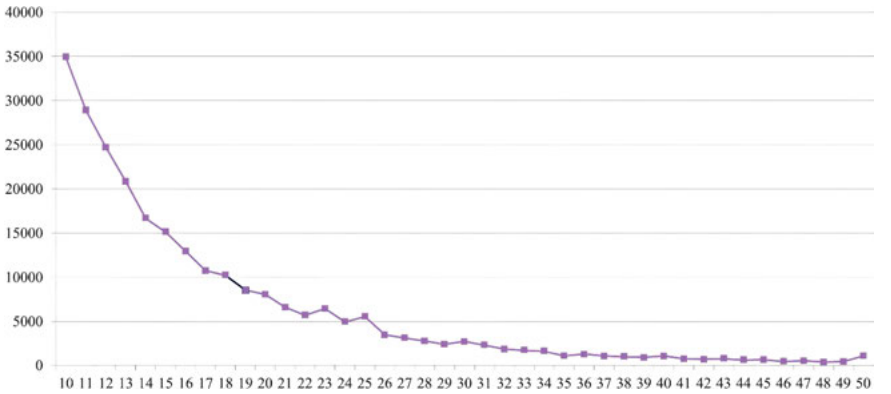

Fig. 1: The number of clusters in relation to their span (10 to 50 years). Source: comhis.fi, accessed September 8, 2022.

Instead of reuse cycles lasting a span of 1–49 years, we began to examine these cycles with simpler time frames, basically asking which kinds of texts were copied infrequently within 5 to 10, or even 20 years. To illustrate the distribution of clusters of different timespans, Fig. 1 shows clusters whose spans range from 10 to 50 years.

In our earlier publications, we only paid attention to fast and slow repetition, in a similar way to Smith, Cordell and Maddock Dillon.[17] However, after considering the results further, it seems clear that the dichotomy between fast and slow gives too limited a view on the types of text reuse. What we call 'mid-range text reuse' represents a significant, albeit often overlooked, practice of newspaper production. Mid-range repetition is mainly about undramatic content. It is easily forgotten especially compared to the texts encountered in long-lasting circulation chains. Studying mid-range circulation may help us to highlight the characteristics of other text circulation cycles. It may also be a fruitful entry point to further nuance the understanding of the forms of text reuse in the nineteenth century.

Typical cases of mid-range text reuse were official announcements using standard sentences annually and even across decades, and other announcements, such as for a research scholarship, published regularly. There were also

17 Smith, D.A., R. Cordell, E. Maddock Dillon. 2013. "Infectious texts: Modeling text reuse in nineteenth-century newspapers." In: (Proceedings) Workshop on Big Humanities. IEEE Computer Society Press 2013, 86–94, http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf, 93.

commercial, textual advertisements published repeatedly over several years either in one location or in several regions. This may also include texts that were reprinted by the same newspaper again and again. The virality score penalizes texts that were not printed by several papers, but these texts still have a significant role in newspaper publishing.

Religious texts and biblical stories were typically circulated in newspapers, and their long-lasting messages made them highly reusable. Their movements from paper to paper took years and even decades. Other short stories and educational texts are also among those that newspapers occasionally republished or borrowed from each other. For instance, one article encouraging farmers to use reindeer lichen and certain natural plants for fodder, or to add lichen to fodder when their crop of hay had failed, was circulated in newspapers at least 83 times around Finland during eight years from 1894 onwards.[18]

Further, the detection process has helped to recognize certain types of texts, or printing conventions, in which exact phrases were repeated. These were, for instance, strings of obituaries of different persons, spanning many years and repeating phrases (such as "died believing in the Savior"). These were then identified as similar and collected in clusters of text reuse. Since the detection process was set to find similarities that are over 300 characters, this means that the phrases were accompanied by other phrases. These can be regarded as boilerplate text, which we aimed to exclude by ignoring similarities under 300 characters. This did not completely succeed, however, since template phrases formed a sequel which was interpreted by BLAST as a larger textual whole. On the other hand, these texts were closely related to the development of newspapers as a media platform since, through these text templates, media visibility was sold and offered for various purposes. The templates were often not reused long-term, meaning decades or fifty years; they were shorter, but in many cases, long-lasting chains of repetition that tell of the changes in public writing patterns and also of how Finnish organizations learned to interact with and utilize the press effectively.

18 P. Paju. 2019. "Jäkälän paluu: Jäkälävalistus ja tekstien uudelleenkäyttö historiallisen tutkimusteeman jäsentäjänä. (Return of the Lichen. Lichen education and outlining a historical research topic by studying text reuse.)" Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41942380.

3 Epistemological Ramifications

Ontology and epistemology are always intertwined in research. Our ability to know about the past, about past entities and "forms of being," is conditioned and framed by our retrospective position, by the available sources, and by the methods selected for research. These conditions, we believe, do not prevent us from making conclusions on the ontological premises, but it is necessary to ponder how these conditions direct our perspective and influence our findings. The following discussion on the epistemological ramifications of text reuse detection is based on three aspects that need to be articulated. These are the materiality of the digital, the problem of metadata, and the question of methodological nationalism that must be considered when national and regional corpora are used.

3.1 Materiality of the Digital

When studying digitized periodicals as sources, the material conditions that framed the development of the press have to be kept in mind. This involves the materiality of newspaper publishing itself, and its many changes, in the nineteenth century.[19] Previous research has explored the development of text layout, including pagination, font size and number of columns as well as on material proportions, such as paper size and quality over this period.[20] We also studied the relationship of text reuse patterns with the growing number of characters per page and larger page sizes in the late nineteenth century, which shows that text reuse grew more moderately than would be indicated by absolute numbers of clusters in our results.[21] The absolute number of reuse clusters grew rapidly towards the end of the research period, which reflects the fact that the volume of the press increased accordingly. Paper size enlarged, especially

19 On previous discussion on source criticism, see for example, R. Abel. 2013. "The Pleasures and Perils of Big Data in Digitized Newspapers.". Film History, 25.1–2: 1–10. Milligan I. 2013. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." The Canadian Historical Review, 94.4: 540–69. M. Koolen, J. van Gorp, J. van Ossenbruggen. 2019. "Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice." Digital Scholarship in the Humanities, 34.2: 368–85.

20 J. Marjanen, V. Vaara, A. Kanner, H. Roivainen, E. Mäkelä, L. Lahti, and M. Tolonen. 2019. "A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917." Journal of European Periodical Studies, 4.1: 54–77.

21 See the figure in Rantala H., H. Salmi, A. Vesanto, F. Ginter. 2019. "Tekstien pitkä elämä: Ajassa liikkuvat tekstit suomalaisessa sanomalehdistössä 1771–1920." Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41858179, 60.

in the 1880s and 1890s, while at the same time font size became smaller. This allowed even more information to be printed by the press. It must be noted, however, that the smaller the font size became, the more difficult it has been for the OCR to recognize the text. Therefore, the OCR accuracy is lower in the 1880s and 1890s than in the 1850s and 1860s, and likewise, there is more OCR noise towards the end of the period. From an epistemological perspective this is fascinating, since – although text reuse detection finds exponentially more results from the period compared to previous decades – it may well be that the real amount of actual reprinting cases is even higher than what could be retrieved with this selected method.

Further to this, the digitization process of newspapers is also impacted by material aspects that influence the results. In many countries, the preservation of old and fragile newspapers started with microfilming projects, since microfilm surrogates were regarded as a stable means for archival preservation.[22] The use of microfilm would also reduce the use of actual newspapers, which again would help the originals to be preserved. When digitization of Finnish newspapers started in the 1990s, it was done mostly on the basis of these microfilms, which was a technically inexpensive and practical solution. This allowed the National Library of Finland to proceed efficiently on the project. The first online collection was opened in 2001, and today all issues published prior to 1920 have been digitized and opened for researchers. The essential epistemological feature lies in the fact that optical character recognition was done on the images taken from the microfilm, not from the original newspaper.[23] If one compares the Finnish development to Sweden, for instance, many historical newspapers there have only been digitized for the first time in 2020. Because the National Library of Finland acted early on in large-scale OCR processing, the present collection contains a lot of noise and random, erroneous characters created by the OCR software in the

22 See, for example, A. Prescott. 2018. "Searching for Dr Johnson: The digitisation of the Burney newspaper collection." In (S. Gøril Brandtzæg, P. Goring and C. Watson, eds.) Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, Brill: Leiden, 49–71, 57.

23 On the history of the collection, see Bremer-Laamanen, M. 2006. "Connecting to the past – newspaper digitisation in the Nordic countries." Journal of Digital Asset Management, 2(3–4): 168–171. See also M. H. Beals, and E. Bell, with contributions by R. Cordell, P. Fyfe, I.G. Russell, T. Hauswedell, C. Neudecker, J. Nyhan, M. Oiva, S. Padó, M. Peña Pimentel, L. Rose, H. Salmi, M. Terras, and L. Viola. 2020. The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges. Loughborough. DOI: 10.6084/m9.figshare.11560059. https://www.digitisednewspapers.net/.

1990s and the early 2000s.[24] Overcoming this problem of OCR inaccuracy, in fact, motivated our project to study text reuse in the first place, and to choose an approach that is tolerant to noise. According to Aleksi Vesanto, text-reuse-BLAST can recognize similarity even if 40 percent of the characters are wrong.[25] Our chosen method thus enabled the project to circumvent the problem and also had the advantage of analyzing a corpus that is complete in relation to the real volume of newspaper publishing.

In these ways, the quality of OCR in the Finnish digital corpus still carries features of its production history, which again conditions the ways in which we can know about text reuse, and the past in general. It has to be added that this production history is an open process, and there are ongoing development actions to improve OCR accuracy; re-digitization is even being considered with selected newspapers.

3.2 The Problem of Metadata

Another aspect of our epistemological assessment deals with the metadata that is available for the OCR corpus of Finnish newspapers. The quality of the metadata used is in general very good, with exact timestamps and the names of publications. There are rare cases of metadata mistakes, but they do not impact the results in this large-scale exploration. There are other features, however, that do influence the investigation.

Most importantly, the Finnish OCR corpus is not segmented, which means that the metadata does not include information on how the texts are divided into different forms of content, like editorials, news items, advertisements, obituaries, and so on. The OCR corpus includes newspaper and journal issues as separate folders that include pages as XML files. This arrangement has two sides: on the one hand, all forms of newspaper publishing are mixed together and cannot be automatically separated from each other. On the other hand, page breaks form ruptures within the material in a way that influences the analysis. If there were segmentation, it would be easier to connect, for example, broken-up news stories,

24 On the problem of OCR noise, see J. Jarlbrink, and P. Snickars. 2017. "Cultural heritage as digital noise: nineteenth century newspapers in the digital archive." *Journal of Documentation*, 73(6), 1228–43.

25 Vesanto, A., A. Nivala, H. Rantala, T. Salakoski, H. Salmi, F. Ginter. 2017. "Applying BLAST to Text Reuse Detection in Finnish News-papers and Journals, 1771–1910." In: (Proceedings) 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017 (Linköping 2017), http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.

the beginning of which is on a different page from the end. In newspapers, the continuation is not necessarily on the next page. Our reuse clusters could perhaps be employed in developing categorization of the contents and in enhancing the metadata, but that is a future prospect.

The fact that the OCR corpus consists of pages influences the estimation of the extent of text reuse. Our text-reuse-BLAST identifies similar passages of text, but page breaks confuse the results. If, for example, an article has been printed twice, and in the first instance it is divided into two pages but in the second instance it is on one page, the algorithm finds two reuse clusters instead of one, since it treats the beginning and the end as different passages. Otherwise, text reuse detection has no upper limit: it identifies similarity as long as it remains similar, but page breaks interrupt the process of recognition. In our epistemological consideration, the previous point on the substantial OCR noise prevents the algorithm of finding all reuse cases, but here the trend is the opposite: page breaks have produced more text reuse clusters than would have been the case if the problem of page breaks had been overcome. This is a problem of course, but it does not undermine the usefulness of our reuse cluster database. It does mean, however, that the absolute numbers of clusters have to be discussed with reservations.

Let us return to the question of the boilerplate, which was mentioned in the earlier discussion on mid-range reuse. Text-reuse-BLAST does not require lower or upper limits for the recognition of similarity, but we set a lower limit of 300 characters to avoid too much boilerplate content. OCR noise also makes it problematic to quantify very short reuse cases. This also relates to the question of metadata. The metadata of the corpus includes information on images on the pages, but many graphic aspects of the paper have been ignored. This leads to the situation that in the OCR corpus smaller text passages might be connected together, although in the actual newspaper they were separated with a visual element. Thus, separate boilerplate texts might aggregate so that their sum exceeds 300 characters, and in the end mix the results. There are advantages too, since this makes boilerplate elements more perceptible in the overall interpretation of text reuse, which is historically fair, considering how much repetitive content there actually was in the newspapers of the past.

In addition to these points, another aspect of metadata must be mentioned. The OCR corpus of the National Library of Finland includes, as already noted, both newspapers and journals. We included all periodical press in the analysis to make it more comprehensive. There lies a problem, however, since some of the journals did not have an exact date of publication, at least to the day. In these cases, the timestamp is not as accurate as in the case of newspaper issues. This means that the time spans of the clusters might be

inaccurate in those cases where one of the publications of the reuse chain was a journal without an exact date. While this creates a problem, it seems that the participation of journals into reuse chains was limited, so this issue does not harm the results of the project much.

3.3 Methodological Nationalism and Digital History

In the last perspective in our epistemological assessment, we would like to draw attention to a more general question, the tendency of researchers to see nation-states as units of analysis, which has been called methodological nationalism.[26] Although the digitization of newspapers, and the creation of digital newspaper corpora have been a welcome and much-appreciated development for historians, digital newspaper history includes challenges that are embedded in the very processes of large digitization projects. As a rule, newspaper corpora are nationally, sometimes regionally, siloed collections provided by an operator/actor responsible for the collection and preservation of the national heritage.[27] As such, digital collections do not differ from other national collections available in the museums and libraries. The easy accessibility of digital corpora, however, might obscure the fact that from the perspective of transnational history, digital newspaper corpora are often biased by design.

For example, in the case of the Finnish history of print culture, one cannot understand the developments in the field without a wider context: Finland only became an independent state in 1917, after having been a part of the Swedish Kingdom until 1809, and a Grand Duchy within the Russian Empire between 1809 and 1917. The early decades of what is regularly defined as the Finnish newspaper history actually also belong to Swedish press history, since Finland was a province of the realm of Sweden. The Finnish National Bibliography regards the *Tidningar Utgifne Af et Sällskap i Åbo*, published in Turku, present-day Finland, in 1771 as the first Finnish newspaper. It was published in Swedish, as most Finnish newspapers were until the mid-nineteenth century. The present Finnish territory was detached from Sweden in 1809, but the

26 Wimmer, A., Schiller, N. G. 2003. "Methodological Nationalism, the Social Sciences, and the Study of Migration: An Essay in Historical Epistemology." The International Migration Review. 37 (2): 576–610. https://www.jstor.org/stable/30037750.

27 On the siloed nature of newspaper collections, see for example M. Ehrmann, E. Bunout, and M. Düring, 2017. Historical Newspaper User Interfaces: A Review. http://library.ifla.org/2578/1/085-ehrmann-en.pdf.

Swedish press stayed as an important reference point and source of information for Finnish newspaper editors. One clear reason for this was the common language: until the second half of the nineteenth century, Swedish was also a dominant language in Finnish newspaper publishing. When literacy among the Finnish-speaking population increased towards the end of the century, the volume of the Finnish-language press also grew in a rising curve.

In Finland, and certainly everywhere, newspapers had an active role in sharing information across borders. An epistemological problem lies in the issue that it is difficult to examine the role of transnational news flows with nationally siloed corpora, which tend to undermine these cross-border news flows. News traveled across state borders, and before the establishment of news agencies or electronic telegraph lines, other papers were often important sources of information, particularly for foreign news. For the historian interested in information flows, it is therefore useful to draw from several newspaper corpora and to try to find ways of combining them. In our project, we analyzed several cases of text reuse, where the chain of circulation actually started outside the Finnish corpus.[28] This can be done by combining qualitative close reading methods with computational tools. If one only draws on a nationally limited or restricted collection, there is a danger of strengthening the methodological nationalism that is inherent in the building of national cultural heritage collections. This bears not only on the issue of sources but also methods. Our text reuse detection method could not identify similarity across language borders, meaning that, for example, a news item published first in Swedish and then in Finnish could not be automatically clustered together. We identified many of these through close reading of clusters, but this method of detection could not be done on a corpus level. Machine translations are not a solution either, since OCR noise makes this more than challenging. This is an avenue for future research in newspaper corpora, as there is an urgent need to combine national and regional collections and find ways of identifying information flows across linguistic borders.

28 See, for example, Salmi, H., A. Nivala, H. Rantala, R. Sippola, A. Vesanto, F. Ginter. 2018. "Återanvändningen av text i den finska tidningspressen 1771–1853." Historisk tidskrift för Finland 1: 46–76.

4 Conclusion

This article has been an effort to articulate ontological and epistemological issues in text reuse detection, and contribute to the wider heuristic discussion on digitized newspaper collections. In the research project *Computational History and the Transformation of Public Discourse in Finland*, we did not start from specific research topics or preconditions, but from the newspaper corpus itself. As a collaboration between historians and data scientists, we developed a special method, text-reuse-BLAST, which proved to be highly productive in aligning similar passages in the OCR corpus. We produced a new dataset that included all reuse clusters and provided for the possibility of further exploring the movements and routes of information. Thus, the project has enriched the original digital collection by providing new entry points to the history of Finnish media. Internationally, the Finnish case is illuminating since newspaper publishing was in its state of emergence throughout the nineteenth century. Up to the 1850s, there were only very few printing locations, but thereafter there was rapid development. By 1920, the periodical press was already a network that covered the whole country, including a distinct division of labor between newspapers on a local level and nationally. Our method can also be applied elsewhere in the effort to understand how newspaper publishing developed over a long period of time. The ontological approach introduced in this article could similarly be useful in other national and international settings. We are continuing this research in the project *Information Flows over the Baltic Sea*, where we combine the Swedish-language newspaper collections from Finland and Sweden to understand transnational news flows.[29]

In our project we found out that text reuse characterized the whole period under study and was not only a phenomenon of the rise of the press in the late nineteenth century. From an ontological perspective, it is essential that news travelled both synchronically and diachronically, in geographical space but also in time. The ontology of reuse can be further developed in the future, for example, by concentrating on the "gray area" we described as mid-range repetition. These reuse cases can shed more light on the changes in the publishing practices and thus help to understand how newspapers and journals were culturally positioned and re-positioned in Finland from the late eighteenth century to the early twentieth century. By revealing how much the content of the newspapers

29 *Information Flows across the Baltic Sea: Swedish-language press as a cultural mediator, 1771–1918* (The Society of Swedish Literature in Finland, 2020–2023), https://blogit.utu.fi/informationsfloden/.

was shared, text reuse detection can, for instance, be used in highlighting what, if anything, made the individual papers stand out among the press. Reprinting patterns can further inform us what was *not* reprinted, and prompt us to ask why this was the case. These omissions of reprinting could reflect emerging political or other divisions in the press.

In the epistemological part of the study, we aimed to track down aspects that influence, and may bias, the results found with computational tools, drawing on existing corpora. The more general aim of these considerations has been to try to shift the discussion on a broader level to issues that are relevant for any project on historical newspapers in the effort to try to understand how they circulated information and, especially, how we, and under which conditions, can know about the past.

To conclude, there is an existential question embedded in our study and its results: How do databases produced in fixed-term research projects survive in the long run? For such datasets as the clusters of text reuse to remain usable, they need maintenance and preferably some improvements over time, all of which might be difficult to sustain in a research environment based mostly on external funding. In similar undertakings, this is a vital perspective to conceive already in the outset. Projects in computational history need long-term research infrastructures both nationally and internationally, in order to secure the sustainability of the field in the future.

Database

Vesanto, A., F. Ginter, H. Salmi, A. Nivala, R. Sippola, H. Rantala, and P. Paju. 2018. Text Reuse in Finnish Newspapers and Journals, 1771–1920, http://comhis.fi/clusters.

Bibliography

Abel, R. 2013. "The Pleasures and Perils of Big Data in Digitized Newspapers.". Film History, 25.1–2:1–10.

Beals, M. H. and E. Bell, with contributions by R. Cordell, P. Fyfe, I.G. Russell, T. Hauswedell, C. Neudecker, J. Nyhan, M. Oiva, S. Padó, M. Peña Pimentel, L. Rose, H. Salmi, M. Terras, L. Viola. 2020. The Atlas of Digitized Newspapers and Metadata: Reports from Oceanic Exchanges. Loughborough. DOI: 10.6084/m9.figshare.11560059. https://www.digitisednewspapers.net/.

Bremer-Laamanen, M. 2006. "Connecting to the past – newspaper digitization in the Nordic countries." Journal of Digital Asset Management, 2(3–4): 168–171.

Büchler, M., G. Crane, M. Moritz, A. Babeu. 2012. “Increasing recall for text re-use in historical documents to support research in the humanities.” In: (Proceedings) Second International Conference on Theory and Practice of Digital Libraries, vol 7489, pp. 95–100. Doi: 10.1007/978-3-642-33290-6_11.

Ehrmann, Maud, Estelle Bunout, and Marten Düring. “Historical Newspaper User Interfaces: A Review.” In *Proceedings of the 85th IFLA General Conference and Assembly*, 1–26. Athens, Greece: IFLA Library, 2019. https://doi.org/10.5281/zenodo.3404155.

Jarlbrink, J. and P. Snickars. 2017. “Cultural heritage as digital noise: nineteenth century newspapers in the digital archive.” Journal of Documentation, 73(6),1228–43.

Koolen, M., van Gorp, J., van Ossenbruggen, J. 2019. “Toward a Model for Digital Tool Criticism: Reflection as Integrative Prac-tice.” Digital Scholarship in the Humanities, 34.2: 368–85.

Marjanen, J., V. Vaara, A. Kanner, H. Roivainen, E. Mäkelä, L. Lahti, and M. Tolonen. 2019. “A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917.” Journal of European Periodical Studies, 4.1: 54–77.

Milligan, I. 2013. “Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010.” The Canadian Historical Review, 94:4: 540–69.

Nora, P. 1997. “General introduction: Between memory and history.” In (P. Nora, ed) The Realms of Memory: Rethinking the French part, Vol. 1: Conflicts and divisions. Columbia University Press: New York,1–21.

Oxford Dictionary of English. 2015. 3rd online edn. Oxford: Oxford University Press. DOI: 10.1093/acref/9780199571123.001.0001.

Paju, P. 2019. ”Jäkälän paluu: Jäkälävalistus ja tekstien uudelleenkäyttö historiallisen tutkimusteeman jäsentäjänä. (Return of the Lichen. Lichen education and outlining a historical research topic by studying text reuse.)” Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41942380.

Prescott, A. 2018. “Searching for Dr Johnson: The digitisation of the Burney newspaper collection.” In (S. Gøril Brandtzæg, P. Goring and C. Watson, eds.) Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, edited by. Brill: Leiden, 49–71.

Rantala H., H. Salmi, A. Vesanto, F. Ginter. 2019. ”Tekstien pitkä elämä: Ajassa liikkuvat tekstit suomalaisessa sanomalehdistössä 1771–1920.” Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41858179.

Rantala, H., A. Nivala, H. Salmi, P. Paju, R. Sippola, A. Vesanto, F. Ginter. 2019. “Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920. Digitaalisten ihmistieteiden näkökulma.” Historiallinen Aikakauskirja 1: 53–67.

Salmi, H. 2020. What is Digital History? Cambridge: Polity.

Salmi, H., A. Nivala, H. Rantala, R. Sippola, A. Vesanto, F. Ginter. 2018. “Återanvändningen av text i den finska tidningspressen 1771–1853.” Historisk tidskrift för Finland 1: 46–76.

Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. “The Long-Term Reuse of Text in the Finnish Press, 1771–1920.” In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf.

Salmi, H., P. Paju, H. Rantala, A. Nivala, A. Vesanto, F. Ginter. 2021. “The Reuse of Texts in Finnish Newspapers and Journals, 1771–1920: A Digital Humanities Perspective”. Historical Methods: A Journal of Quantitative and Interdisciplinary History 54 (1): 14–28, https://doi.org/10.1080/01615440.2020.1803166.

Smith, D.A., R. Cordell, E. Maddock Dillon. 2013. “Infectious texts: Modeling text reuse in nineteenth-century newspapers.” In: (Proceedings) Workshop on Big Humanities. IEEE Computer Society Press 2013, 86–94, http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf.

Vesanto, A., A. Nivala, H. Rantala, T. Salakoski, H. Salmi, F. Ginter. 2017. “Applying BLAST to Text Reuse Detection in Finnish News-papers and Journals, 1771–1910.” In: (Proceedings) 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017 (Linköping 2017), http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.

Wimmer, A., Schiller, N. G. 2003. “Methodological Nationalism, the Social Sciences, and the Study of Migration: An Essay in Historical Epistemology.” The International Migration Review. 37 (2): 576–610. https://www.jstor.org/stable/30037750.