Textual Migration Across the Baltic Sea: Creating a Database of Text Reuse Between Finland and Sweden

Petri Paju[1], Hannu Salmi[1], Heli Rantala[1], Patrik Lundell[2], Jani Marjanen[3] and Aleksi Vesanto[1]

[1] *University of Turku, Department of Cultural History, Turku, FI-20014, Finland*
[2] *Örebro University, School of Humanities, Örebro, SE-70182, Sweden*
[3] *University of Helsinki, Department of Digital Humanities, Helsinki, FI-00014, Finland*

Abstract

In this paper, we present a database and an interface on text reuse between newspapers and journals published in the Swedish language in Sweden and Finland during the 1645–1918 time frame. Using two national, digital newspaper collections, we detected their textual similarities with a computational method to study the textual migration, i.e., information flows, between the two countries. For purposes of this project, we developed a database of detected clusters of text reuse and an online interface to search, examine and analyse the transnational movement of information. The database, *Text Reuse in the Swedish-language Press, 1645–1918*, is accessible online and includes texts from over 1,100 newspapers and journals published at approximately 150 locations at various times during the 274-year time frame.

Keywords

computational history, text reuse, historical newspaper, digital collections, database construction, Finland, Sweden, transnational history, information flow

1. Introduction

This short paper presents a database and an interface on text reuse among Swedish-language newspapers and journals during the 1645–1918 time frame. The database and interface were built as part of the project, *Information flows across the Baltic Sea: Swedish-language press as a cultural mediator, 1771–1918*. The database and the accompanying project aim to study information flows, particularly between Sweden and Finland from the period when present-day Finland was part of the Swedish kingdom to the establishment of Finland as a Grand Duchy in the Russian Empire after 1809 and until the Independence of Finland in 1917 and Civil War in 1918. Even after the 1809 separation, news and other texts circulated because of the common cultural heritage and shared language, i.e., Swedish. The border was relatively easy to cross, and newspapers circulated between Sweden and Finland regularly. However, because their national histories eventually diverged, these press materials have been preserved, processed, and siloed in two national libraries. Still, print media digitisation makes it possible to study overlaps in large collections of texts and see how information was spread across the Baltic Sea.

In our project, textual migration was traced using a method based on the software BLAST, which can be applied to text-reuse detection. With the method, we detected every text passage with 300 or more characters of similarity and combined these passages into reuse clusters. We included Swedish-language papers published in Sweden and Finland, but excluded the Finnish-language press in Finland, as textual migration within Finland has been studied in previous publications [1, 2]. To strengthen the

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.
EMAIL: petpaju@utu.fi (A. 1); hansalmi@utu.fi (A. 2); hemara@utu.fi (A. 3); Patrik.Lundell@oru.se (A. 4); Jani.Marjanen@helsinki.fi (A. 5); avjves@gmail.com (A. 6)
ORCID: 0000-0002-2486-2364 (A. 1); 0000-0001-8607-6126 (A. 2); 0000-0002-4108-4904 (A. 3); 0000-0002-6221-9089 (A. 4); 0000-0002-3085-4862 (A. 5)

project's transnational dimension, we also included Swedish-American newspapers available via the National Library of Sweden, originally from the Minnesota Historical Society's collection.

In this paper, besides introducing the new database and its interface, we particularly examine the question of how a database can serve as a tool and a novel way to study transnational information flows. The paper also aims to inform database users about what they need to know and consider when conducting searches of text-reuse material and analysing results drawn from it.

The results of text-reuse detection are presented in the form of a database and its interface, which include all detected text-reuse passages and provide clusters of reused passages. The database bears the name *Text Reuse in the Swedish-language Press, 1645–1918,* and it has been accessible online since October 2021.

The interface is an easy-to-use tool for examining passages, as well as sometimes-viral clusters to which they belong. It allows for studying both long-term and short-term text reuse; text reuse between different countries, towns and newspaper titles; and reuse within countries. A map function illustrates potential viral chains as well. Our tool also can be used to examine the circulation of particular newspaper items, although it is not so such something that BLAST picks up, but rather what the most widespread texts include, e.g., advertisements, announcements and telegrams. Overall, the database and interface provide ample opportunities for studying concrete cases of virality and cultural mediation between Sweden and Finland, but they also provide a way of quantitatively assessing cultural asymmetries present between the two countries during the long nineteenth century.

The paper also discusses the critical issues involved when combining historical newspaper collections from countries with different size and historical trajectories. It further raises the issue of how the different choices made in the two newspaper collections' digitisation process influenced how these data sets can be connected and processed further. The paper concludes with an assessment of both the benefits and pitfalls of a database as a historical representation.

2. Textual Migration and Text Reuse

Despite Harold Innis' modern classic *Empire and Communications* [3] and regardless of a more recent so-called spatial turn in media research [4], the existing literature on the characteristics and contours of information geography in the 1800s is relatively scarce. Scholars have investigated economic, journalistic, political, rhetorical and technological frameworks and implications of new communications, globally or nationally [5–7]. Other studies have approached nineteenth century communication technologies from the perspective of 'great power rivalries' [8] and how they were used as 'tools of empire' [9]. And yet some have pointed to this perspective's limitations, as it fails to acknowledge implementations in specific contexts. This type of research usually offers case studies [10, 11].

Of course, a wide array of earlier scholarship on textual migration exists, both qualitative and quantitative. For example, text movements have been traced in the study of medieval manuscripts and their itineraries, in the study of book trade and book publishing, and in the study of quotation and paraphrasing practices. However, the present project's interest in text reuse emanates from a desire to nuance present-day assumptions of information virality related to digital media [12, 13], and from the previously somewhat-neglected fact that nineteenth century press items were 'shared' and continuously republished to a considerable extent [14–16], eliciting a wide range of historical questions. The *Viral Text Project* (VTP)[2] and the *Oceanic Exchanges* [17], both published in the United States, and the *Computational History and the Transformation of Public Discourse in Finland* (COMHIS),[3] as well as the newly launched *Information Highways of the 19th Century* in Sweden, all work with large data sets to trace textual migration in time and space. VTP investigates 'the great unread', the plentiful poems and short stories that circulated via newspapers, and the recontextualisation of texts and how authorships were transformed as texts were copied and reprinted [16, 18]. The *Oceanic Exchanges* examines information flow patterns across national and language borders. COMHIS builds on a more refined algorithm (BLAST), detects a larger number of reuses and identifies long chains of fast-moving text-

[2] Viral Text Project, https://viraltexts.org.

[3] See the COMHIS database at http://comhis.fi/clusters.

reuse cases, as well as the distribution of ads and slow-moving text items [1, 2, 19]. The present project, *Information Flows,* uses the same algorithm as COMHIS and studies the Swedish-language press as a cultural mediator between Finland and Sweden.

3. Database Construction

The basic idea behind constructing the database was to create a tool for analysing text reuse in the Swedish-language press on a transnational scale. In addition to Sweden, Finland also has had, and still has, a vivid Swedish-language publishing culture. In the construction of the database, it was possible to draw on the comprehensive digitisation of Finnish Swedish-language papers from the first published issue in 1771 onwards, realised by the National Library of Finland [20, 21]. The copyright-free collection reaches up to 1920, and includes practically all published issues. We restricted our database up to the year 1918 for historical reasons, i.e., to include the Independence of Finland in 1917 and the Civil War in 1918. The OCR'd XML files of published issues are downloadable from the Language Bank of Finland.[4]

In Sweden, a large digitisation project with historical newspapers is in an active phase, with the aim of covering newspapers from the inauguration of the press in 1645 up to 1906, which is viewed as the last year in the open collection. The OCR'd content was made available to us with an API via the Betalab of the National Library of Sweden in fall 2020.[5] At the time of construction of the current database, approximately half the Swedish newspaper collection had been digitised. This material comprised the bulk of our corpus, which also was supplemented with digitised magazines from the Swedish Language Bank's collection, including material up to the 1910s [22]. By these means, it was possible to extend the time frame up to 1918.

Finally, for text reuse detection, we had more than 5 million pages of digitised content: 1.79 million pages from Finland and 3.24 million from Sweden. The database includes texts from over 1,100 titles of newspapers and journals published at approximately 150 locations. However, three aspects must be noted about the material. First, although we were able to add material from the collections of the Swedish Language Bank, the content, published in Sweden after 1906, is very thin, i.e., text-reuse cases from 1907 to 1918 provide only a fragmentary view of what really was published. Second, the publishing business already had started in Sweden in 1645. In turn, the first newspaper in present-day Finland was published in 1771 as part of the expansion of the press in the Kingdom of Sweden. Therefore, it is obvious that most of the cross-border reuse cases are from the nineteenth century. However, by including earlier papers from the seventeenth century, we wanted to emphasise that long reuse chains also were possible, i.e., the texts travel not only in space, but in time. Content from the very early papers could have been printed later, which is of historically significant interest and can be brought to light through a reuse database. Third, it is important to note that the material, downloaded via the API of the National Library of Sweden, also included newspapers published outside of Sweden and Finland, mostly in the US. Therefore, the database also covers several Swedish-American papers from the Minnesota Historical Society's collection. In constructing the database, we decided to include these issues for the benefit of users in general, who then could get the maximum amount of results and have a wider understanding of the public sphere formed by the Swedish-language press.

The text-reuse detection method applied in this project was based on the National Centre for Biotechnology Information Basic Local Alignment Search Tool (NCBI BLAST). This software initially was developed to match biological sequences, but it also can be used to trace duplicated text passages from a corpus of scanned and OCR-recognised newspapers and journals. Researchers in the University of Turku's Department of Computing developed this BLAST application, i.e., text-reuse-BLAST (for more information on the technical details of BLAST and the processing of data, see [2, 23]). To avoid boilerplate results in the reuse chains, the minimum length of passages was set at 300 characters. The original OCR data were not segmented perfectly into articles, so elements such as page breaks or

[4] The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771–1874), published 2011, http://urn.fi/urn:nbn:fi:lb-201505112; Newspaper and Periodical OCR Corpus of the National Library of Finland (1875–1920), published 2017, http://urn.fi/urn:nbn:fi:lb-201405275.

[5] For further information, see KB Data Lab, https://github.com/Kungbib/kblab, accessed 15 January 2022.

pictures in the original images can cause multiple clusters of similar passages. Therefore, the absolute number of discovered passages does not necessarily reflect the actual level of reuse.

First, similar passages were recognised with BLAST by using CSC Finland's supercomputers. After this, instances of similarity were clustered so that text-reuse cases could be presented through the database's interface. In the end, we found 17.8 million clusters, out of which 2.4 million were shared between countries.

The database and its web interface, at https://textreuse.sls.fi, draw on many existing open source libraries and software: Solr is a database software; pysolr is a Python client that communicates with Solr; and nginx, in turn, takes care of the traffic to Django, which is a web framework needed to construct the database's actual web pages. Other services also are used, e.g., the clusters' map function draws on flowmap.blue. Construction of the database required intensive collaboration among the whole research team, e.g., to verify metadata. We manually and qualitatively checked all printing sites, which was necessary because multiple newspapers had identical titles, e.g., *Aftonbladet* (the *Evening Paper*), and certain papers shifted publishing sites over time. Careful identification of printing sites per year was needed for the map function to be accurate and for the visualisation of reuse clusters.

4. Database Functions

The database and interface's functions and parameters have been co-designed in particular to meet the project's objectives, e.g., to help answer questions on transnational information flows. Simultaneously, the project has aimed to provide a new tool for others interested in historical newspapers more broadly.

To begin with, the database allows for effective searches of the detected text-reuse cases. When choosing advanced searches, it is possible to search either individual hits or clusters. A *hit* is a single instance of a passage being repeated in the data set. The hit will be a passage from a page of an issue in the data set. Every newspaper page usually contains multiple hits, although they are generally parts of different clusters. A *cluster*, in turn, refers to a group of hits that all share the same (or similar enough) text passage.

A hit search is useful when the user searches for a specific detail, e.g., a name, event or term. The power of a cluster search becomes evident when the user is interested in text circulation in its many varieties and scales. By using the expressions of cluster search, a user can examine text circulation, e.g., from a specific time period and/or across (present) national borders.

Hits and clusters have been given different attributes in their metadata, and they can be searched and sorted with these parameters. Both hit and cluster searches have different available search fields: Clicking the i-button beside the search box opens detailed instructions.[6]

4.1. Sorting options for hits and clusters

The interface offers several features for filtering and organising search results. In the interface (see Figure 1), search limitations are on the left, and a variety of sorting options are in the upper-right corner. These differ from each other based on whether one is searching for hits (individual hits) or clusters. Here, the focus is on expressions regarding clusters because they are probably new to most users. These expressions are also more project-specific.

On the left, the clusters can be limited through eight parameters: starting country; starting location; starting year of appearance; span across multiple countries; port city; port country; incoming city; and incoming country. From the perspective of information flows, 'span across multiple countries' is important. If the user clicks 'Yes', the search will be limited to clusters in which the text has been published in two of the geographical regions (Finland, Sweden, and the US) or in all of them. *Port city* means the last city of the reprint cluster in its country of first printing, i.e., the city that presumably 'sends' the text overseas. *Incoming city* refers to the first-printing location of a text in another *incoming country*.

[6] See Guidelines for the database, at https://textreuse.sls.fi/guidelines.

In the upper-right corner, the results can be organised by average length, starting and ending dates, starting country or location, number of unique locations and starting year. It is also possible to sort by count, timespan (in days), gap (in years) and virality score. *Count* refers to the number of hits in the cluster. *Timespan* refers to the length of the cluster in days. *Gap* allows the user to find clusters with significant breaks in the chain of texts. If a text was printed for the first time in 1850 and several times from 1900 onwards, there is a maximum gap of 50 years.

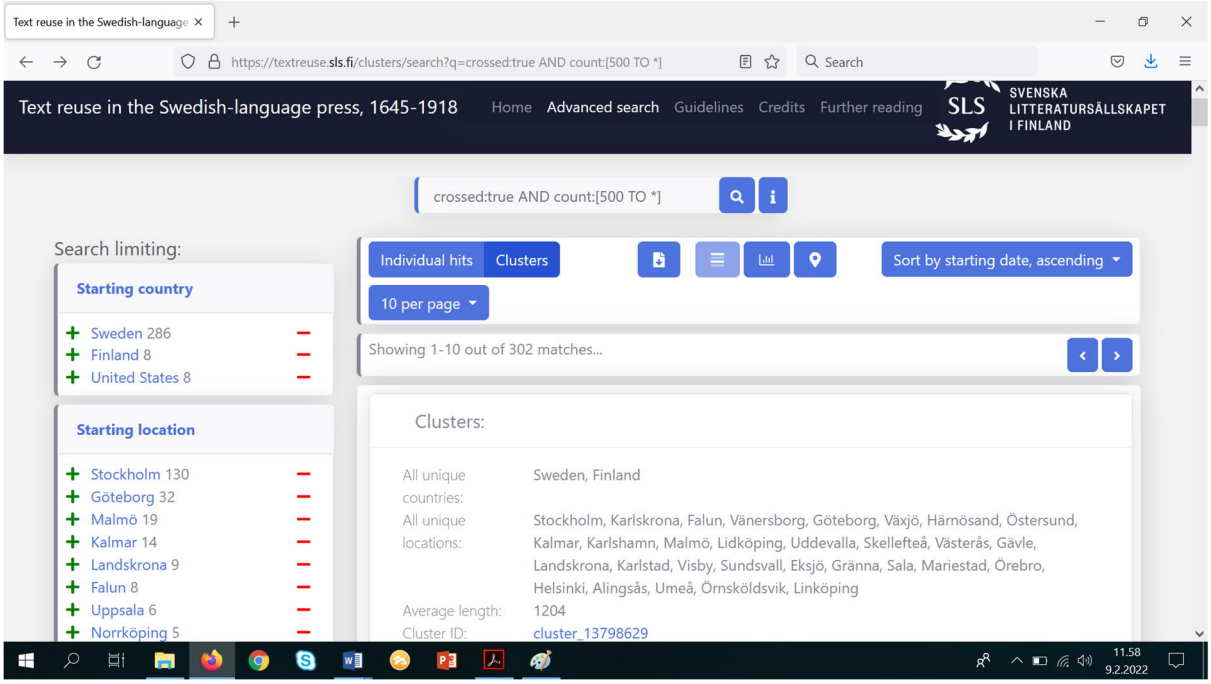

Figure 1: The database interface. Here, the user has searched clusters with 500 or more texts within two or more countries (Source: https://textreuse.sls.fi/).

Finally, the user can sort the results of a cluster search by *virality score*, which is a way of approaching how efficiently a particular text has circulated through the media network, its value calculated from the metadata [2]. All the clusters' virality score values were normalised between 0–100 for clarity. We view virality score as a tool among many others: The values are not decisive in themselves, but they offer the user a tool for filtering out material and perhaps finding interesting cases. This filtering can be combined with other features, e.g., by checking which were the most efficiently spread cross-border texts annually.

4.2. Charts and maps

The search results can be viewed as charts or on a map. After pushing the chart button, the user can view the search results, hits or clusters as a chart with either absolute or normalised values and choose to organise them by year or by month. By choosing the map function, the user can examine, e.g., a cross-border text-reuse cluster geographically.

Mapping can be conducted only with a single cluster, but in two ways: *Origin to rest* shows the text reuse as if radiating from the location of the first hit in the cluster. *Chain hits together* assumes that the cluster flows linearly in time from its starting location to the rest of the hits' locations chronologically. Often, the text-reuse chains do not conform to these patterns, but rather comprise a combination of the two. Still, the two options help gauge possible chains of influence in text reuse.

4.3. Downloading search results

One important function in the interface is the option to download search results as .tsv files (tab-separated values), which can be imported into most spreadsheet applications. Downloaded results can

be processed and scrutinised further with other tools, e.g., network analysis software, or they can be used to create graphical presentations to visualise trends in the search results.

5. Critical Issues

Digital sources need source-critical attention as much as any other historical material. The digitisation process itself raises several critical issues, e.g., errors created by OCR processing (see, e.g., [24, 25]). Moreover, considering that digital newspaper corpora, as a rule, have been built up and designed as national collections, the process of combining them into one corpus raises issues of which database users should be aware. Several asymmetries are evident between the Swedish and Finnish digital newspaper collections, some deriving from historical differences between the countries, while others are a product of different solutions made in the digitisation process.

For historical reasons, the first 'Finnish' publications could be counted as Swedish ones because Finland belonged to the Swedish realm until the early nineteenth century. However, in our database, all reuse cases between Sweden and Finland have been classified as 'cross-border' movement. Moreover, the former Finnish towns of Vyborg and Sortavala (both founded during the Swedish reign) are included as historical Finland, although they are presently part of Russia.

Compared with the press in Sweden, the volume of the Finnish press was still quite modest before 1850; thus, few publications could have circulated information. Furthermore, the database provides information only on reuse cases within the Swedish-language press, i.e., all Finnish-language publications printed in Finland fall outside this database's scope. In Finland, the material includes journals, but in Sweden, journals are not included. It is essential to realise that newspapers and journals historically sometimes have been classified quite arbitrarily. Thus, any close readings and case studies can benefit from an extra search in the national collections, *Svenska dagstidningar* and Finnish newspapers and journals.[7]

The digitisation of newspapers has been a long ongoing process during which the technical aspects of scanning have changed. For scanning purposes, old newspapers are noisy material. All sorts of faults are evident, e.g., scratches and ink spillage that have resulted in varying quality in scanned newspaper images [26]. Furthermore, they have changed the OCR software several times during the long scanning process [27]. Thus, OCR quality varies inside one national collection, as well as between the Swedish and Finnish collections. However, text-reuse-BLAST is tolerant to noise and, therefore, also can be used in cases in which the quality of material is poor or variable.

Apart from these critical points deriving from the national collections' features, database users should be aware of some other shortcomings and limitations. In our database, *text reuse* refers to various forms of textual overlap, including direct quotations and intentional or unintentional borrowing. However, the nature of repetition is case-specific. The database displays this overlap in text-reuse chains (clusters), but it is important to be aware that errors can appear in these chains. The aforementioned problem of multiple clusters means that BLAST sometimes splits reuse chains into several clusters with identical or very similar content. This produces additional clusters and results in higher cluster numbers than actual cases of repetition. Therefore, one must treat the precise number of clusters as approximate.

Furthermore, although the database offers a tool for studying text movement between different localities and regions, it does not reveal the chronology of this movement *per se*, but simply lists a specific reuse chain's publishing order. This does not necessarily mean that the papers actually quoted each other in that order.

To illustrate the difficulty of drawing straightforward conclusions about a single cluster case, we can imagine a text passage printed first in Helsinki on July 30, 1850, then in Stockholm on August 10 and in Gothenburg on August 20. This hypothetical, yet realistic, example might mean that 1) they all reprinted a letter or an ad sent to them, 2) all three newspapers independently quoted a fourth newspaper (not included in the data set), 3) the Stockholm and Gothenburg editors were both subscribing to the Helsinki paper and quoted directly from it, 4) the Gothenburg paper quoted the Stockholm paper, which, in turn, quoted the Helsinki paper or 5) some combination of these scenarios. Thus, establishing the actual physical (or electrical) transportation of content requires more information. Sometimes this can

[7] See https://tidningar.kb.se/ (Sweden) and https://digi.kansalliskirjasto.fi/etusivu (Finland).

be found in the texts that these passages are part of, while in other cases, different sources need to be consulted, and often, it is still not possible to determine. However, if the case in question is viewed in terms of an immaterial dissipation of content, then something did, indeed, spread in time and space, from Helsinki via Stockholm to Gothenburg. Therefore, terms like *port city, incoming city* and *virality score*, which are used in the database, must be employed with both imagination and care.

6. Impact

Although newspapers typically are not studied through individual authors, studying text reuse in newspapers further highlights how texts not only were written by individual authors, but also appropriated and republished for different purposes in other newspapers. Ryan Cordell analyses this through the lens of the 'network author' [16], but more than a reconceptualisation of authorship, text reuse highlights how reception and republishing entail historical agency in a way that historical scholarship has not been accustomed to acknowledging. A text may have been very important – an instant classic – from the very beginning, but sometimes its importance has become cemented only after active reception.

The database makes it possible to study practices of text reuse both qualitatively and quantitatively. Qualitative cases may include studying a particular text and its history of reuse in different newspapers in Sweden and Finland, but also may depart from a particular theme or topic defined by a keyword search or a focus on particular towns or newspapers defined by using the metadata available. Any search results then can be downloaded and analysed further using other available tools. For instance, a researcher interested in text reuse between Gothenburg and Vyborg can download search results and use a corpus linguistics tool for further examination.

The searches also provide quantitative data that can help with quantitative assessment of text-reuse patterns. As mentioned above, the clusters identified do not always offer one-to-one matches with historical republications of texts due to complexities in layout changes, but the figures nonetheless provide a good indicator of how many texts were circulated, from where they departed and which newspapers republished them. For instance, the database provides solid figures on how common it was for texts to be reused across the Gulf of Bothnia, compared with internal Swedish text reuse. The figures can be narrowed further geographically and based on publication date, making it possible to study regional influence patterns over time. Any thematic interest, defined through a keyword search, also may be quantified through figures provided on the interface.

Both the qualitative and quantitative cases attest to the cultural asymmetries present between Sweden and Finland. On one hand, the Swedish press dominates text-reuse clusters in sheer amounts. On the other hand, individual cases complicate the situation by providing examples in which the influence has been reversed or is not that straightforward. A crucial amount of agency is present in the reception and reuse of texts, and the reuse of Swedish texts in Finland should be viewed as both a testament to Swedish influence, as well as a Finnish choice to uphold a cultural connection. This connection seems to have become more important towards the end of the nineteenth century, when language relations in Finland became more strained [28]. At that time, Swedish-language papers in Finland yielded significant symbolic power by drawing from newspapers in Sweden.

7. Conclusion

This paper has presented a database as a tool for studying transnational information flows. It can benefit researchers from various disciplinary backgrounds – from literary studies to historical research. Its use benefits from previous knowledge of the Swedish language and the history of newspaper publishing within the region. The database of text reuse between Sweden and Finland was sparked by the project's aim to understand information flows across national and geographical borders. For the database, we designed specific parameters and functions to best answer our interests. We expect to develop these features further in the future. In the previous discussion, we emphasised the benefits of the database's functionalities and also discussed its critical issues. Database users should keep its limitations in mind, and one way of doing this is to form a practice of cross-checking findings in the original, national digital newspaper repositories.

According to our test phase so far, the database is well-suited for detecting transnational text flows, providing ample possibilities to filter results and visualise reuse clusters. For instance, large-scale transnational marketing campaigns could be made visible through the database. Likewise, the database works well for examining viral news that spread rapidly in the newspaper network. Because of its long time span, ranging from the seventeenth century to the early twentieth century, the database also is useful for examining how texts travel over time and how they are reprinted later in history.

The database *Text Reuse in the Swedish-language Press, 1645-1918* includes several levels of historical representation. First, the database's interface serves as a representation of the clusters found during the reuse-detection process. The clusters themselves are only chains of text passages, which is why they must be enriched through available metadata so that their dimensions can be examined. Second, the interface aims to represent the text flows of the past. What is at stake are not only the clusters, detected via a computational process, but also the migration of texts in the historical world from the seventeenth century to the twentieth. In this sense, the database's interface works as a historical interpretation, although it must be kept in mind that clusters, without reservation, cannot be viewed as actual text-reuse chains. One chain in the historical world might be split into several clusters due to data quality and the nature of the process. In addition to these two layers of representation, a third also exists. In the world of the past, historical actors themselves understood that continuous text-reuse processes existed, but their abilities to grasp the extent of the newspaper network were restricted. The database allows the user to investigate the actors' positions as well, particularly in situations when they took advantage of the network in the form of marketing and information campaigns.

In conclusion, we aimed to find a novel way to grasp cross-border information flows. Useful databases for historical research are available, but what is new in our case is that, drawing on computational methods, we constructed a database that connects and combines digitised content from multiple countries and allows for the study of text flows. The database was developed under the guidance of the researchers themselves and contains functions tailored to examine text reuse and information flows. This is important also from a more general perspective. Digitised newspaper repositories have been built predominantly on a national basis through libraries' initiatives worldwide. This has been valuable work, but simultaneously, the digitised collections have been confined to national domains. Our database and its interface are one solution for going beyond the siloed histories of newspaper publishing and contributing more broadly to an increasingly transnational understanding of history.

8. Acknowledgements

We would like to thank Erik Edoff, Johan Jarlbrink, Pelle Snickars, the National Library of Finland and the National Library of Sweden. The research consortium *Information flows across the Baltic Sea: Swedish-language press as a cultural mediator, 1771–191*8 (2020–2023) is funded by Svenska litteratursällskapet i Finland (Society of Swedish Literature in Finland). The computational resources have been provided by the CSC-IT Centre for Science; Espoo, Finland.

9. References

[1] H. Salmi, A. Nivala, H. Rantala, R. Sippola, A. Vesanto, F. Ginter, Återanvändningen av text i den finska tidningspressen 1771–1853. Historisk Tidskrift för Finland. 103, 46–76 (2018).

[2] H. Salmi, P. Paju, H. Rantala, A. Nivala, A. Vesanto, F. Ginter, 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, 14–28 (2021). https://doi.org/10.1080/01615440.2020.1803166.

[3] H. A. Innis, Empire and communications. Rowman & Littlefield, Lanham, Md (2007).

[4] J. Falkheimer, A. Jansson, Geographies of communication: The spatial turn in media studies. Nordicom, Göteborg (2006).

[5] T. Rantanen, When news was new. Wiley-Blackwell, Chichester, U.K. (2009).

[6] P. Flichy, Dynamics of modern communication: The shaping and impact of new communication technologies. Sage Publications, London; Thousand Oaks (1995).

[7] J. W. Carey, Communication as culture: essays on media and society. Routledge, New York (2009).
[8] J. Hills, The struggle for control of global communication: the formative century. University of Illinois Press, Urbana (2002).
[9] D. R. Headrick, The tools of empire: technology and European imperialism in the nineteenth century. Oxford University Press, New York (1981).
[10] M. Hampf, S. Müller-Pohl (Eds.), Global communication electric: business, news and politics in the world of telegraphy. Campus-Verl, Frankfurt am Main (2013).
[11] J. Harvard, P. Stadius, Conclusion: Mediating the Nordic brand - history recycled. In: P. Stadius, J. Harvard (Eds.), Communicating the North: Media Structures and Images in the Making of the Nordic Region. pp. 319–332. Ashgate, Burlington (2013).
[12] K. Nahon, J. Hemsley, Going viral. Polity Press, Cambridge, England (2013).
[13] H. Jenkins, S. Ford, J. Green, Spreadable media: creating value and meaning in a networked culture. New York University Press, New York; London (2013).
[14] C. Blevins, Space, Nation, and the Triumph of Region: A View of the World from Houston. Journal of American History. 101, 122–147 (2014). https://doi.org/10.1093/jahist/jau184.
[15] J. Jarlbrink, Mobile/sedentary: News work behind and beyond the desk. Media History. 21, 280–293 (2015). https://doi.org/10.1080/13688804.2015.1007858.
[16] R. Cordell, Reprinting, Circulation, and the Network Author in Antebellum Newspapers. Am Lit Hist. 27, 417–445 (2015). https://doi.org/10.1093/alh/ajv028.
[17] R. Cordell, M. Beals, I. Galina, M. Priewe, E. Priani, H. Salmi, J. Verheul, R. Alegre, S. Koch, T. Hauswedell, P. Fyfe, J. Hetherington, E. Lorang, A. Nivala, S. Pado, L.-K. Soh, M. Terras, Oceanic Exchanges. (2017). https://doi.org/10.17605/OSF.IO/WA94S.
[18] R. Cordell, A. Mullen, "Fugitive Verses": The Circulation of Poems in Nineteenth-Century American Newspapers. American Periodicals: A Journal of History & Criticism. 27, 29–52 (2017).
[19] A. Nivala, H. Salmi, J. Sarjala, History and Virtual Topology: The Nineteenth-Century Press as Material Flow. Historein. 17, (2018). https://doi.org/10.12681/historein.14612.
[20] T. Pääkkönen, J. Kervinen, A. Nivala, K. Kettunen, E. Mäkelä, Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. D-Lib Magazine. 22, (2016). https://doi.org/10.1045/july2016-paakkonen.
[21] K. Kettunen, T. Pääkkönen, Kansalliskirjaston historialliset sanoma- ja aikakauslehdet avoimena digitaalisena datana: datapaketteja, rajapintoja, käyttäjiä ja tutkimusongelmia. Informaatiotutkimus 37, (2018). https://doi.org/10.23978/inf.77412.
[22] D. Dannélls, The Kubhist corpus of Swedish newspapers, https://spraakbanken.gu.se/blogg/index.php/2019/09/15/the-kubhist-corpus-of-swedish-newspapers/.
[23] A. Vesanto, A. Nivala, H. Rantala, T. Salakoski, H. Salmi, F. Ginter, Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910. In: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. pp. 54–58. Linköping University Electronic Press, Gothenburg (2017).
[24] J. Jarlbrink, P. Snickars, Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. JD. 73, 1228–1243 (2017). https://doi.org/10.1108/JD-09-2016-0106.
[25] M.J. Hill, S. Hengchen, Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. Digital Scholarship in the Humanities. 34, 825–843 (2019). https://doi.org/10.1093/llc/fqz024.
[26] M. Koistinen, K. Kettunen, T. Pääkkönen, Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In: Proceedings of the 21st Nordic Conference on Computational Linguistics. pp. 277–283. Association for Computational Linguistics, Gothenburg, Sweden (2017).
[27] E. Mäkelä, K. Lagus, L. Lahti, T. Säily, M. Tolonen, M. Hämäläinen, S. Kaislaniemi, T. Nevalainen, Wrangling with non-standard data. In: S. Reinsone, I. Skadiņa, A. Baklāne, J. Daugavietis (Eds.), Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 81–96. CEUR-WS.org, Germany (2020).
[28] M. Engman, Språkfrågan: Finlandssvenskhetens uppkomst 1812–1922. Svenska litteratursällskapet i Finland, Helsingfors (2016).