DiVA
http://www.diva-portal.org

This is the published version of a paper published in *Historisk Tidskrift för Finland*.

Citation for the original published paper (version of record):

Paju, P. (2022)
Tidningstexter o¨ver O¨stersjo¨n: En databas av delat innehåll mellan Finland och Sverige
Historisk Tidskrift för Finland, 107(2): 192-211

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:oru:diva-101681

PETRI PAJU, ERIK EDOFF, PATRIK LUNDELL, JANI MARJANEN,
HELI RANTALA, HANNU SALMI, ALEKSI VESANTO

Tidningstexter över Östersjön: En databas av delat innehåll mellan Finland och Sverige

Avsikten med denna översikt är att presentera en databas och ett gränssnitt för studier av textåteranvändning mellan svenskspråkiga tidningar och tidskrifter i Finland och Sverige 1645–1918.[1] *Text reuse in the Swedish-language press, 1645–1918* (textreuse.sls.fi), tillgänglig sedan oktober 2021, har skapats inom projektet Informationsflöden över Östersjön: Svenskspråkig press som kulturförmedlare, 1771–1918. Syftet med databasen och det anslutande forskningsprojektet är att studera informationsflöden mellan Finland och Sverige från tiden då dagens Finland var en del av Sverige, över perioden som storfurstendöme i det ryska riket och fram till självständigheten 1917 och inbördeskriget 1918. Även efter separationen 1809 cirkulerade nyheter och andra texter inom språkområdet. Skälen till detta var många, såsom språket och det gemensamma kulturarvet. Gränsen var förhållandevis lätt att passera och tidningar skeppades i båda riktningarna mellan länderna. Eftersom tidningarna har arkiverats och samlats i två olika nationalbibliotek har de dock som regel kommit att behandlas och förstås som två av varandra oberoende samlingar. Digitaliseringen av materialet

1. Texten är en omarbetad och för HTF:s publik anpassad version av Petri Paju, Hannu Salmi, Heli Rantala, Patrik Lundell, Jani Marjanen och Aleksi Vesanto, 'Textual migration across the Baltic Sea. Creating a database of text reuse between Finland and Sweden', Digital Humanities in the Nordic and Baltic Countries (DHNB 2022) 6th conference, *CEUR workshop proceedings*, CEUR-WS.org, Karl Berglund, Matti La Mela & Inge Zwart (eds.), under utgivning.

gör det möjligt att betrakta det som en enhet och studera likheter och överlappningar mellan tidningar som publicerats på olika sidor om Östersjön. På så sätt kan man följa hur information spreds. I projektet spåras denna textmigrering – eller, om man så vill, detta delade innehåll – med en metod som baseras på programvaran BLAST. Med hjälp av metoden har vi identifierat alla textpassager med minst 300 liknande tecken. Dessa passager kombinerades i återanvändningskluster. För att begränsa materialet och databasens storlek omfattar den endast ett språk, i detta fall svenska. Några finskspråkiga tidningar från Finland[2] förekommer således inte i databasen.[3] Däremot, för att stärka databasens transnationella dimension och vidare användbarhet, har vi även inkluderat Kungliga bibliotekets (KB) digitaliserade svenskamerikanska tidningar, vilka ursprungligen härrör från Minnesota Historical Societys samlingar.

Förutom att introducera den nya databasen och dess gränssnitt, fokuserar vi i denna översikt särskilt på databasen som ett verktyg för och ett bidrag till ett nytt sätt att studera internationella informationsflöden. Syftet är också att informera databasens potentiella användare om vad de behöver känna till vid sökningar i kedjorna av textåterbruk för att sedan kunna analysera resultaten på ett adekvat sätt. Programvarans textåteranvändningsdetektering presenteras alltså i form av en databas som är åtkomlig och sökbar via ett webbgränssnitt. Databasen inkluderar samtliga identifierade textåteranvändningspassager och tillhandahåller kluster av återanvända textpassager. Gränssnittet är ett lättanvänt verktyg för att utforska dessa ibland virala kluster och textpassager. Det gör det möjligt att studera kortsiktig textåteranvändning såväl som långsiktig sådan, texttrafik över landsgränser, mellan specifika städer och tidningstitlar, liksom återanvändning inom

2. I den svenska samlingen förekommer ett fåtal finskspråkiga tidningar utgivna i Sverige, som *Haaparannanlehti*, vilka inte aktivt har uteslutits och därför inkluderas i databasen.
3. Textåteranvändning inom Finland studeras i Hannu Salmi, Asko Nivala, Heli Rantala, Reeta Sippola, Aleksi Vesanto & Filip Ginter, 'Återanvändningen av text i den finska tidningspressen 1771–1853', HTF 103 (2018:1), s. 46–76, https://journal.fi/htf/article/download/80405/42644; Hannu Salmi, Petri Paju, Heli Rantala, Asko Nivala, Aleksi Vesanto & Filip 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* 104 (2021:1), s. 14–21, https://doi.org/10.1080/01615440.2020.1803166.

länderna. En kartfunktion illustrerar också potentiellt virala kedjor. Eftersom programvaran BLAST identifierar textpassager som *liknar* varandra utan att för den skull behöva vara identiska (vilket närmast är en förutsättning när man har att göra med ett material som är så fullt av brus som digitaliserade gamla tidningar), kan verktyget även användas för att utforska framväxten (och eventuellt borttynandet) av särskilda format eller stilar, det vill säga över tid mer eller mindre hårt mallade texttyper; bland de största klustren hittas sålunda sådant som annonser, kungörelser och telegram – med varierande innehåll men i en mycket snarlik form. Sammantaget ger databasen och gränssnittet goda möjligheter att studera allt från konkreta fall av viralitet och kulturell förmedling mellan Finland och Sverige till att kvantitativt bedöma de kulturella asymmetrier som fanns mellan de två länderna under det långa 1800-talet.

Utöver denna grundläggande presentation av databasens uppbyggnad och gränssnittets funktioner diskuterar vi kritiska frågor som uppstår när man kombinerar historiska tidningssamlingar från länder med olika storlek och skilda historiska erfarenheter. Det väcker bland annat frågan om hur de olika val som gjorts i digitaliseringsprocessen av de två tidningssamlingarna påverkar hur dessa datamängder kan kopplas samman och ytterligare bearbetas. Vi avslutar med en diskussion om både förtjänster och fallgropar med att betrakta en databas som en historisk representation.

Textmigrering och textåteranvändning

Oaktat Harold Innis moderna klassiker *Empire and Communications* och det färskare talet om en så kallad rumslig vändning inom medievetenskapen, eller för den delen de allra senaste årens intresse för cirkulation inom den så kallade kunskapshistorien, är studier av informationsgeografins egenskaper och konturer under 1800-talet relativt få.[4] Relevant kvalitativ forskning har undersökt ekonomiska, journalistiska, politiska, retoriska och tekniska ramar och konsekvenser av

4. Harold Innis, *Empire and Communications* (Lanham, Md 2007); Jesper Falkheimer & André Jansson, *Geographies of Communication. The Spatial Turn in Media Studies* (Göteborg 2006); Johan Östling, Erling Sandmo, David Larsson Heidenblad, Anna Nilsson Hammar & Karin H. Nordberg (eds.), *Circulation of Knowledge. Explorations in the History of Knowledge* (Lund 2018).

nya kommunikationer, globalt eller nationellt.[5] Somliga har närmat sig 1800-talets kommunikationsteknologier som en aspekt av stormaktsrivalitet och hur de användes som imperiella verktyg.[6] Andra har pekat på detta perspektivs begränsningar, då det inte lyckas att visa dess bäring i mer historiskt specifika sammanhang. Den senare typen av forskning ägnar sig vanligtvis åt fallstudier.[7]

Det finns naturligtvis ett brett utbud av tidigare forskning om textmigration i olika former, både kvalitativ och kvantitativ. Texters förflyttning i rummet har spårats till exempel i studiet av medeltida manuskript och deras färdvägar, i undersökningar av bokhandel och bokutgivning samt i utforskandet av citat- och parafraseringsmetoder. Vårt intresse för återanvändning av text härrör dock ur en strävan efter att nyansera dagens föreställningar om informationsviralitet relaterat till digitala medier och från det tidigare något försummade faktumet att 1800-talets tidningsmaterial ”delades” och återpublicerades kontinuerligt i betydande omfattning.[8] Detta öppnar för en lång rad historiska frågor.

De amerikanska projekten Viral Text Project (VTP)[9] och Oceanic Exchanges[10], det finländska Computational History and the Transformation of Public Discourse in Finland (COMHIS) samt det nystartade svenska Information Highways of the 19th Century arbetar alla med

5. Terhi Rantanen, *When News was New* (Chichester 2009); Patrice Flichy, *Dynamics of Modern Communication. The Shaping and Impact of New Communication Technologies* (London 1995); James W. Carey, *Communication as Culture. Essays on Media and Society* (New York 2009).
6. Jill Hills, *The Struggle for Control of Global Communication. The Formative Century* (Urbana 2002); Daniel R. Headrick, *The Tools of Empire. Technology and European Imperialism in the Nineteenth Century* (New York 1981).
7. M. Michaela Hampf & Simone Müller-Pohl (eds.), *Global Communication Electric. Business, News and Politics in the World of the Telegraph* (Frankfurt am Main 2013); Jonas Harvard & Peter Stadius, ’Mediating the Nordic brand – history recycled’, Peter Stadius & Jonas Harvard (eds.), *Communicating the North. Media Structures and Images in the Making of the Nordic Region* (Burlington 2013), s. 319–332.
8. Karine Nahon & Jeff Hemsley, *Going Viral* (Cambridge 2013); Henry Jenkins, Sam Ford & Joshua Green, *Spreadable Media. Creating Value and Meaning in a Networked Culture* (New York 2014); Cameron Blevins, ’Space, nation, and the triumph of region. A view on the world from Houston’, *Journal of American History* 101 (2014:6), s. 122–147, https://doi.org/10.1093/jahist/jau184; Johan Jarlbrink, ’Mobile/sedentary. News work behind and beyond the desk’, *Media History* 21 (2015:3), s. 280–293, https://doi.org/10.1080/13688804.2015.1007858.
9. Viral Text Project, https://viraltexts.org.
10. Se COMHIS databas på http://comhis.fi/clusters.

stora datamängder för att spåra textmigrering i tid och rum. VTP undersöker ”the great unread”, den rika mängd dikter och noveller som cirkulerade via tidningar, och rekontextualiseringen av texter liksom hur författarskap omvandlades när texter kopierades och trycktes på nytt.[11] Oceanic Exchanges undersöker informationsflödesmönster över nations- och språkgränser. COMHIS bygger, i sin tur, på en mer förfinad algoritm (BLAST), som upptäcker ett större antal återanvända texter och identifierar långa kedjor av snabbrörliga fall av textåterbruk såväl som distributionen av annonser och långsamma textobjekt.[12] Det här aktuella projektet, Informationsflöden över Östersjön, använder samma algoritm som COMHIS och studerar alltså svenskspråkig press som kulturförmedlare mellan Finland och Sverige.

Databasen

Grundidén bakom databasen var att skapa ett verktyg för att analysera textåteranvändning i svenskspråkig press på en transnationell skala. Förutom Sverige har även Finland haft, och har fortfarande, en livaktig svenskspråkig förlagskultur. I uppbyggnaden av databasen har vi kunnat dra nytta av Nationalbibliotekets omfattande digitalisering av den svenskspråkiga pressen i Finland, från det första numret av *Tidningar Utgifne af et Sällskap i Åbo* 1771 framåt.[13] Den upphovsrättsfria samlingen sträcker sig fram till 1920, och innehåller samtliga publicerade nummer. I tid begränsade vi databasen till år 1918 av historiska skäl, för att kunna inkludera självständigheten 1917 och inbördeskriget 1918. OCR-lästa XML-filer av de i Nationalbibliotekets

11. Ryan Cordell, ’Reprinting, circulation, and the network author in antebellum newspapers’, *American Literary History* 27 (2015:3), s. 417–445, https://doi.org/10.1093/alh/ajv028; Ryan Cordell & Abigail Mullen, ”’Fugitive verses”. The circulation of poems in nineteenth-century American newspapers’, *American Periodicals. A Journal of History & Criticism* 27 (2017:1), s. 29–52.
12. Salmi et al., ’Återanvändningen av text i den finska tidningspressen’; Salmi et al., ’The reuse of texts in Finnish newspapers and journals’; Asko Nivala, Hannu Salmi & Jukka Sarjala, ’History and virtual topology. The nineteenth-century press as material flow’ *Historein* 17 (2018:2), https://doi.org/10.12681/historein.14612.
13. Tuula Anneli Pääkkönen, Jukka Kervinen, Asko Nivala, Kimmo Tapio Kettunen & Eetu Mäkelä, ’Exporting Finnish digitized historical newspaper contents for offline use’, *D-Lib Magazine* 22 (2016:7/8), http://dlib.org/dlib/july16/paakkonen/07paakkonen.html; Kimmo Kettunen & Tuula Anneli Pääkkönen, ’Kansalliskirjaston historialliset sanomaja aikakauslehdet avoimena digitaalisena datana. Datapaketteja, rajapintoja, käyttäjiä ja tutkimusongelmia’, INF 37 (2018:4), https://doi.org/10.23978/inf.77412.

digitala samlingar publicerade numren finns att ladda ner från finska Språkbanken.[14]

I Sverige pågår fortfarande det omfattande projektet med att digitalisera tidningssamlingen, från de första numren av *Ordinari Post Tijdender* år 1645 fram till 1906; utifrån en försiktigare bedömning av innebörden av att upphovrätten gäller till och med 70 år efter upphovspersonens död, gör Kungliga biblioteketet successivt tillgängligt material som är minst 115 år. Genom ett API (Application Programming Interface) via KB-labb gjordes det OCR-lästa materialet tillgängligt för oss hösten 2020.[15] Vid insamlingen till databasen var ungefär hälften av den svenska samlingen digitaliserad.[16] Detta material utgör huvuddelen av vår nuvarande korpus som också kompletterats med digitaliserade tidningar ur svenska Språkbankens samling, inklusive material fram till och med 1910-talet.[17] På det viset var det möjligt att förlänga tidsramen fram till 1918.

För textåteranvändningsdetektering hade vi inalles mer än fem miljoner sidor med digitaliserat innehåll: 1,79 miljoner sidor från Finland och 3,24 miljoner sidor från Sverige. Databasen innehåller texter från över 1 100 tidnings- och tidskriftstitlar utgivna på omkring 150 orter. Tre anmärkningar måste dock göras rörande materialet. För det första, trots att vi kunde lägga till material från svenska Språkbankens samlingar är innehållet publicerat i Sverige efter 1906 mycket tunt, vilket gör att textåteranvändningsfall från 1907 till 1918 endast ger en fragmentarisk bild av vad som faktiskt publicerades. För det andra började tidningsutgivningen i Sverige redan 1645, medan den första tidningen i Finland kom ut först 1771, som en del i tidningslitteraturens expansion inom det svenska riket. Därför kommer de flesta fall av gränsöverskridande återanvändning att stamma från 1800-talet. Men genom att ta med även äldre tidningar har vi önskat betona att över tid mycket

14. Nationalbibliotekets OCR-korpus över tidningar och tidskrifter (1771–1874) publicerades 2011, http://urn.fi/urn:nbn:fi:lb-201505112. Nationalbibliotekets OCR-korpus över tidningar och tidskrifter (1875–1920) publicerades 2017, http://urn.fi/urn:nbn:fi:lb-201405275.
15. För vidare information, se KB data lab, https://github.com/Kungbib/kblab (hämtad 20.5.2022).
16. Processen kan följas här: https://feedback.blogg.kb.se/forums/topic/digitaliserade-dagstidningar.
17. Dana Dannélls, 'The Kubhist corpus of Swedish newspapers', https://spraakbanken.gu.se/blogg/index.php/2019/09/15/the-kubhist-corpus-of-swedish-newspapers.

utdragna kedjor av textlån inte bara är möjliga utan verkligen förekommer. Detta understryker det faktum att texterna färdas inte bara rumsligt utan även tidsmässigt. Innehåll från de tidiga tidningarna kan ha tryckts om senare, vilket är av stort historiskt värde och kan synliggöras just genom en återanvändningsdatabas av det här slaget. Som en tredje punkt är det viktigt att notera att materialet, nedladdat via KB:s API, även omfattade tidningar som publicerats utanför Finland och Sverige, mestadels i USA. Därför täcker databasen även en liten del av den svensk-amerikanska pressen, nämligen Minnesota Historical Societys samling. Vi inkluderade dessa titlar för att forskare ska kunna få ut största möjliga antal resultat och, förstås, en bredare och mer realistisk förståelse av den historiska offentligheten, liksom denna formades av den svenskspråkiga periodiska litteraturen.

Den metod vi använde för att upptäcka textåteranvändningen baseras på National Center for Biotechnology Information Basic Local Alignment Search Tool (NCBI BLAST). Från början utvecklades denna programvara för att matcha biologiska sekvenser, men den kan också användas för att spåra duplicerade textpassager från till exempel en samling skannade och OCR-lästa tidningar och tidskrifter. Denna tillämpning av BLAST, textåteranvändnings-BLAST, har tagits fram av forskare vid Institutionen för datavetenskap vid Åbo universitet.[18] För att undvika nonsensresultat (till exempel i form av så kallad boilerplatetext) i återanvändningskedjorna sattes textpassagernas minimilängd till 300 tecken. Den ursprungliga OCR-datan är långt ifrån perfekt segmenterad i artiklar, varför element som sidbrytningar och illustrationer i originalavbildningen kan orsaka flera kluster av liknande passager. Därför speglar det absoluta antalet identifierade passager inte nödvändigtvis den faktiska nivån av återanvändning.

Först identifierades textavsnitt som liknar varandra genom BLAST med hjälp av superdatorerna vid IT Center for Science i Finland. Efter det sammanfördes likartade texter i kluster så att textåteranvändningsfall kunde presenteras via databasens gränssnitt. Totalt innehåller data-

18. För mer information om de tekniska detaljerna för BLAST och databearbetningen, se Salmi et al., 'The reuse of texts in Finnish newspapers and journals'; Aleksi Vesanto, Asko Nivala, Heli Rantala, Tapio Salakoski, Hannu Salmi & Filip Ginter, 'Applying BLAST to text reuse detection in Finnish newspapers and journals, 1771–1910', *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language* (Göteborg 2017), s. 54–58, http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.

basen 17,8 miljoner kluster, av vilka 2,4 miljoner representerar texttrafik över dagens nationsgränser.

Databasen och gränssnittet bygger på många befintliga bibliotek och programvaror med öppen källkod: Solr är en databasprogramvara, pysolr är en Python-klient som kommunicerade med Solr, nginx tar i sin tur hand om trafik till Django som är ett webbramverk som behövs för att konstruera databasens faktiska webbsidor. Även andra tjänster används. Till exempel bygger klustrens kartfunktion på flowmap.blue. Uppbyggnaden av databasen krävde ett intensivt samarbete inom hela forskargruppen, till exempel för att verifiera metadata. Vi kontrollerade bland annat samtliga tryckorter manuellt och kvalitativt. Detta var nödvändigt eftersom det finns flera tidningar med samma namn, till exempel *Aftonbladet*, liksom tidningar som haft skiftande tryckorter. En noggrann identifiering av tryckorter för varje givet år behövdes för kartfunktionen och för visualisering av återanvändningsklustren.

Databasens funktioner

Databasen och gränssnittets funktioner och parametrar har designats för att i första hand uppfylla projektets specifika mål, exempelvis för att besvara dess historiska frågor om transnationella informationsflöden. För att inte utesluta forskare med endast rudimentära kunskaper i svenska valde vi engelska som språk för gränssnittet. Det går att föreställa sig att forskare som är intresserade av regionens textflöden i fågelperspektiv inte nödvändigtvis behöver förstå innehållet för att göra sig en bild av cirkulationen. Samtidigt syftar projektet till att ge ett nytt verktyg åt andra som är intresserade av att använda historiska tidningar för helt andra forskningsfrågor. Som modell för databasens funktioner stod det tidigare projektet COMHIS som behandlade textåteranvändning i tidningar som publicerats i Finland.

Till att börja med tillåter databasen sökningar i de upptäckta fallen av textåteranvändning. Med hjälp av avancerad sökning är det möjligt att söka efter antingen enskilda träffar eller kluster. En träff (*hit*) utgör ett enskilt ställe av en passage som upprepas i datasetet, en textpassage från en sida av ett nummer i datasetet. Varje tidningssida innehåller vanligtvis flera träffar, även om de oftast ingår i olika kluster. Ett kluster (*cluster*) hänvisar i sin tur till en grupp träffar som alla delar samma (eller snarlika) textavsnitt.

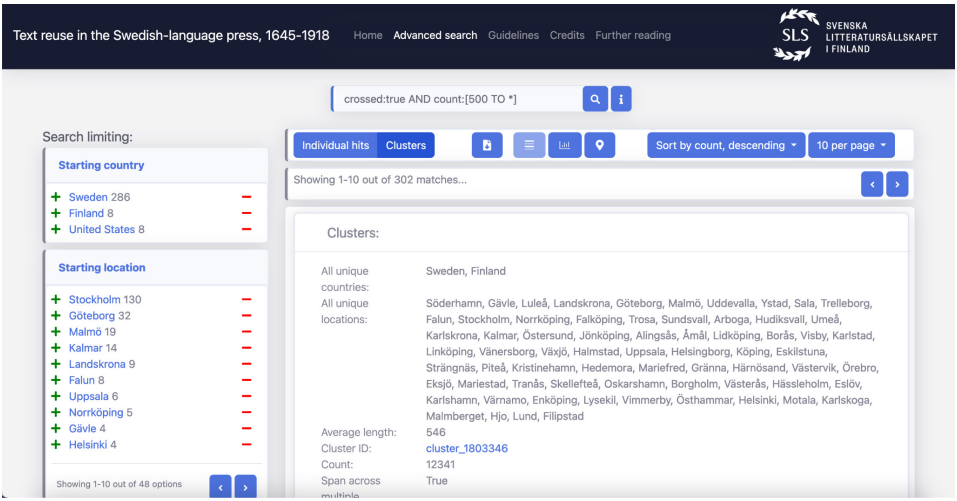
Träffsökning är användbart när användaren söker efter en specifik detalj, till exempel ett namn, en händelse eller en term. Styrkan i klustersökning blir uppenbar när användaren är intresserad av textcirkulationen i sig. Genom att använda klustersökning kan textcirkulationen utforskas till exempel under en viss tidsperiod och/eller över (nuvarande) nationsgränser. Träffar och kluster har fått olika attribut i sina metadata, och de kan sökas och sorteras efteråt utifrån dessa parametrar. Träff- och klustersökningar har i sin tur olika tillgängliga sökfält; genom att klicka på i-knappen bredvid sökrutan visas detaljerade instruktioner.[19]

Gränssnittet erbjuder flera funktioner för att filtrera och organisera sökresultat. I gränssnittet (se figur 1) finns olika filtreringsmöjligheter till vänster och en mängd sorteringsalternativ uppe i högra hörnet. Dessa skiljer sig från varandra beroende på om man söker efter träffar (*individual hits*) eller kluster. I den här texten ligger fokus på kluster eftersom de förmodligen är nya för de flesta användarna.

Till vänster kan klustren begränsas genom åtta parametrar: *starting country*, *starting location*, *starting year of appearance*, *span across multiple countries*, *port city, port country*, *incoming city* och *incoming country*. Ur vårt forskningsprojekts perspektiv är *span across multiple countries* förstås en central funktion. Om användaren väljer ”yes” kommer sökningen att begränsas till kluster där textavsnittet har publicerats i åtminstone två av de geografiska regionerna (Finland, Sverige, USA). Port city betyder den sista staden i klustret i landet för den första publiceringen, det vill säga den stad som kan sägas ”sända” texten utomlands. En *incoming city* syftar följaktligen på den första publiceringsorten för en text i ett annat land, *incoming country*.

Uppe i högra hörnet kan resultaten organiseras efter genomsnittlig längd, start- och slutdatum, startland eller plats, antal unika platser och startår. Det är också möjligt att sortera efter antal, tidsperiod (i dagar), *gap*/lucka (i år) och viralitetspoäng. *Count* betyder antalet träffar som finns i klustret. *Time span* hänvisar till klustrets längd i dagar. *Gap* låter i sin tur användaren hitta kluster där det finns betydande tidsavbrott i textkedjan. Om till exempel en text trycktes första gången 1850 och sedan flera gånger först från 1900 och framåt är det ett maximalt *gap* på 50 år. Till sist kan användaren sortera resultaten av en klustersök-

19. Se databasens användarinstruktioner på https://textreuse.sls.fi/guidelines.

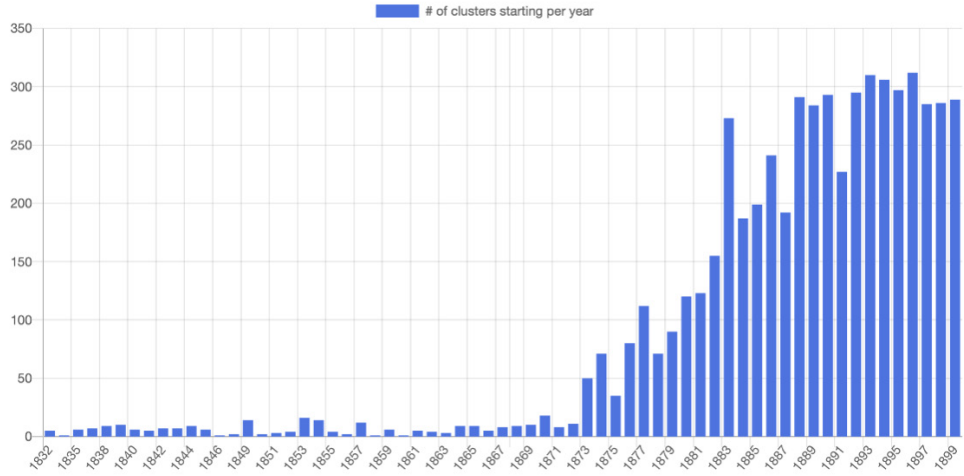


Figur 1: Gränssnittet till databasen. Här har användaren sökt efter kluster med 500 eller fler texter inom två eller flera länder. (Källa: https://textreuse.sls.fi.)

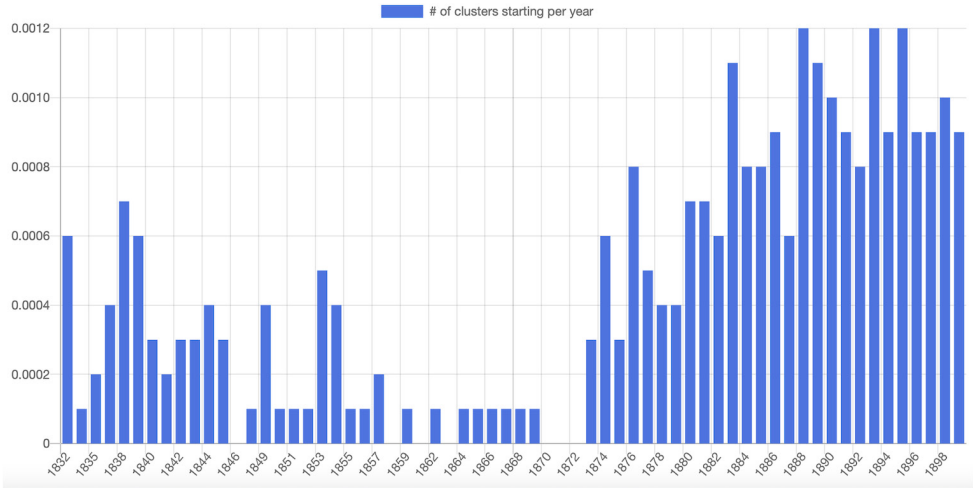
ning efter ett *virality score*. Detta är ett sätt att närma sig hur effektivt en viss text har cirkulerat genom medienätverket, och dess värde beräknas från metadata.[20] Alla klusters viralitetspoängsvärden normaliserades mellan 0 och 100 för tydlighetens skull. Viralitetspoäng bör ses som ett verktyg bland många andra: värdena är inte avgörande i sig, men de ger användaren ett verktyg för att filtrera materialet och upptäcka intressanta fall. Detta filter kan kombineras med de andra funktionerna, till exempel för att få fram vilka som var de mest effektivt spridda gränsöverskridande texterna årligen.

Sökresultaten kan visualiseras som diagram eller på en karta. Efter att ha tryckt på diagramknappen kan användaren se sökresultat, träffar eller kluster som ett diagram med antingen absoluta eller normaliserade värden och välja att organisera dem per år eller per månad. Normaliseringsfunktionen är användbar och viktig eftersom tillväxten av den periodiska litteraturen närmast kan beskrivas som exponentiell; vad som absolut sett kan framstå som små krusningar tidigt under en längre tidsperiod, framträder i relativa termer på ett mer realistiskt sätt.

20. Salmi et al., 'The reuse of texts in Finnish newspapers and journals'.

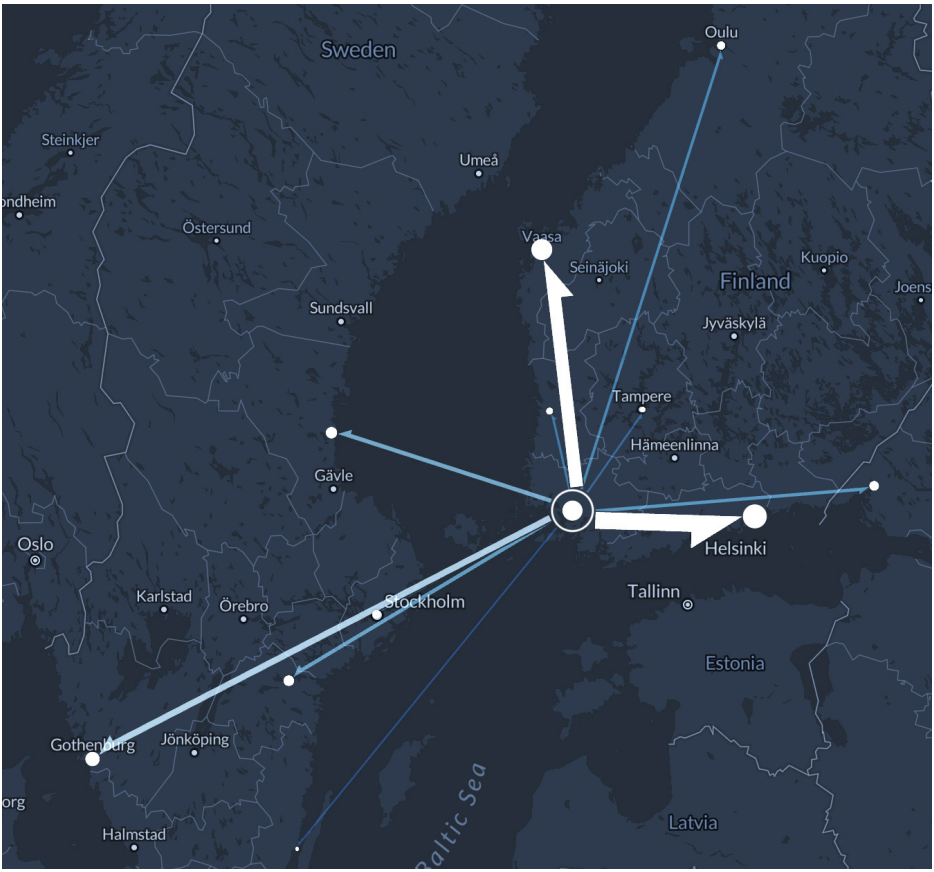


Figur 2: Diagram, absoluta tal. Här har användaren sökt fram nationsgränsöverskridande kluster innehållande ordet nykterhet under åren 1832 till 1899.* (Källa: https://textreuse.sls.fi.)

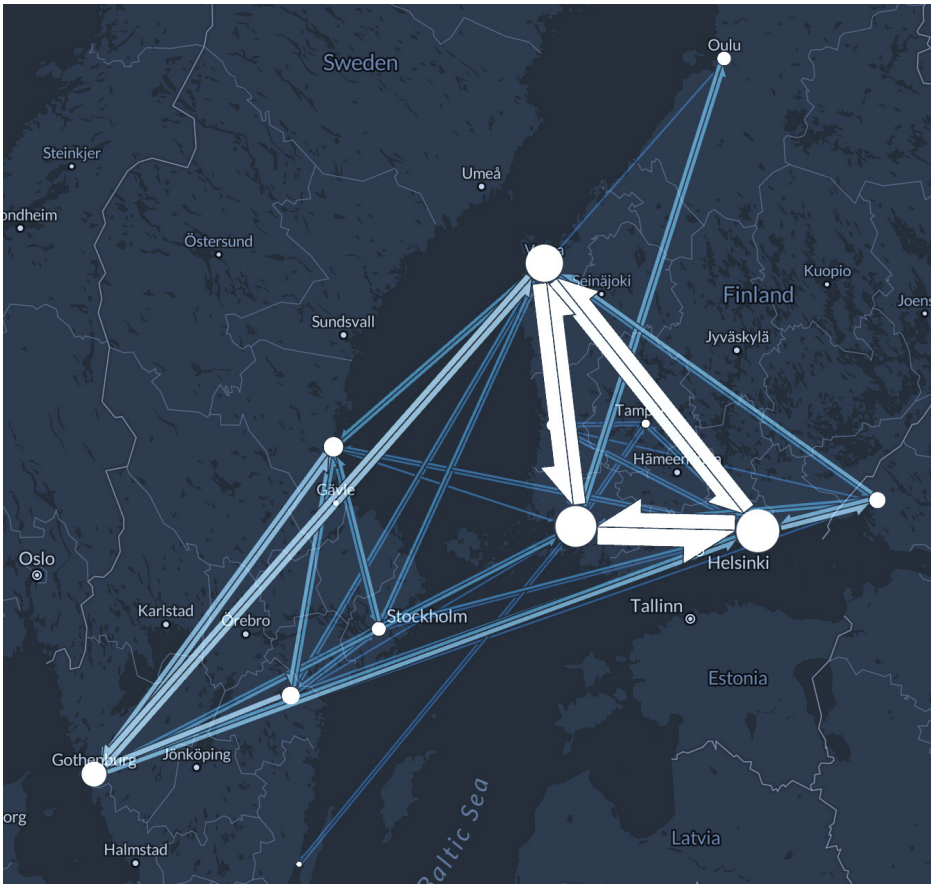


Figur 3: Diagram, normaliserade tal. Här har användaren sökt fram nationsgränsöverskridande kluster innehållande ordet nykterhet under åren 1832 till 1899. Till skillnad från i redovisningen av absoluta tal (figur 2) framträder här den stundtals även under 1830-, 1840- och 1850-talen relativt intensiva texttrafiken.* (Källa: https://textreuse.sls.fi.)

Genom att välja kartfunktionen kan användaren utforska till exempel ett gränsöverskridande textåteranvändningskluster geografiskt. Kartläggning kan endast göras med ett enskilt kluster i taget men på två olika sätt: *Origin to rest* visar textåteranvändningen som om den strålade ut från platsen för klustrets första träff. *Chain hits together* antar att klustret flödar linjärt i tiden från startplats till resten av klustrets platser i turordning (se figur 2). Kedjorna av faktisk, historisk textåteranvändning överensstämmer ofta inte med dessa mönster utan består snarare av en kombination av båda. De två alternativen är likafullt till stor hjälp för att spåra faktiska eller åtminstone uppskatta möjliga påverkanskedjor vid återanvändning av text.



Figur 4: Annons för Carlsbader Piller omtryckt under åren 1889–1915, i cluster_2270395, count 945 (karta: from origin to rest, https://textreuse.sls.fi).



Figur 5: Samma kluster som ovan (figur 2) men presenterat enligt den andra kartvarianten (karta: chain hits together, https://textreuse.sls.fi).

En viktig funktion i gränssnittet är möjligheten att ladda ned sökresultat i formatet .tsv, som kan importeras till de flesta kalkylbladsapplikationer. Nedladdade resultat kan bearbetas vidare och granskas med andra verktyg, till exempel programvara för nätverksanalys, eller användas för att skapa grafiska presentationer för att visualisera trender i sökresultaten.

Kritiska spörsmål

Digitala källor behöver källkritisk uppmärksamhet lika mycket som annat historiskt material. Digitaliseringsprocessen i sig aktualiserar flera kritiska frågor, till exempel avseende fel som skapats av OCR-bearbetningen.[21] Till yttermera visso väcker processen med att kombinera nationella tidningssamlingar till en gemensam korpus frågor som användaren av databasen bör vara medveten om. Det finns flera asymmetrier mellan de svenska och finska digitala tidningssamlingarna. Vissa av dem härrör från historiska skillnader mellan länderna medan andra är en produkt av olika beslut som fattats i digitaliseringsprocessen.

Några av asymmetrierna har redan berörts, som skillnaderna i digitaliseringsprocessens faser och i storleken på de två samlingarna samt den svenska tidningsutgivningens längre historia. Av historiska skäl skulle de första ”finländska” publikationerna förstås kunna räknas som svenska eftersom Finland tillhörde det svenska riket fram till år 1809. I vår databas har dock alla återanvändningsfall mellan Finland och Sverige klassificerats som ”gränsöverskridande” förflyttningar. Dessutom ingår de tidigare finska städerna Viborg och Sordavala (båda grundades under det svenska styret) i det historiska Finland även om de för närvarande är en del av Ryssland.

Den finländska pressens volym växte relativt långsamt före slutet av 1800-talet i jämförelse med svenska förhållanden. Den svenskspråkiga tidningssfären i Finland var fortfarande ganska blygsam under 1800-talets första hälft – från två till åtta samtidigt utgivna tidningar före 1850 – vilket innebär att det inte fanns särskilt många publikationer som kunde fånga upp och sprida informationen. När det gäller det finländska materialet bör man också komma ihåg att databasen endast berör återanvändningsfall inom den svenskspråkiga pressen. Det innebär med andra ord att inga finskspråkiga publikationer ingår i databasen. Därtill är det också nödvändigt att påminna om att tidningar och tidskrifter inte sällan har katalogiserats godtyckligt. Det vi i dag tydligt uppfattar som en tidskrift har ibland katalogiserats och

21. Johan Jarlbrink & Pelle Snickars, ’Cultural heritage as digital noise. Nineteenth century newspapers in the digital archive’, *Journal of Documentation* 73 (2017:6), s. 1228–1243, https://doi.org/10.1108/JD-09-2016-0106; Mark John Hill & Simon 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 (2019:4), s. 825–843, https://doi.org/10.1093/llc/fqz024.

därför digitaliserats som en tidning och återfinns därför i vår korpus, medan tidningar som i andra fall klassats som tidskrifter tyvärr fallit utanför. Eventuella närläsningar och fallstudier kan alltså dra nytta av en extra sökning i de nationella samlingarna.[22]

Digitaliseringen av tidningar har varit en lång och är fortfarande en pågående process under vilken de tekniska möjligheterna för skanning har förändrats. Ur ett skanningsperspektiv är gamla tidningar ett material med mycket brus. Det finns alla möjliga defekter – som revor och veck, svagt tryck i frakturstil och spill av trycksvärta – som har resulterat i varierande kvalitet på de skannade tidningsbilderna.[23] Dessutom har OCR-mjukvaran ändrats flera gånger under den långa skanningsprocessen.[24] Detta betyder att kvaliteten på OCR-läsningen varierar inom en och samma nationella samling, såväl som mellan den svenska och finska samlingen. Å andra sidan – och helt avgörande för det här projektets framgång – är textåteranvändnings-BLAST mycket tolerant mot brus och kan därför användas även i de fall där kvaliteten på materialet är varierande eller rent av undermålig.

Förutom dessa kritiska punkter som härrör från de nationella samlingarnas egenskaper, bör användaren av databasen vara medveten om några andra brister och begränsningar. I vår databas avser textåteranvändning olika former av textöverlappning, inklusive direkta citat och såväl avsiktliga som oavsiktliga lån. Upprepningens karaktär är dock specifik från fall till fall. Databasen visar denna överlappning i textåteranvändningskedjor (kluster) men det är viktigt att vara medveten om att vissa fel kan förekomma. Det redan nämnda problemet med fler redovisade än faktiska kluster innebär att BLAST ibland har delat upp återanvändningskedjor i flera kluster med identiskt eller

22. De nationella tidningssamlingarna återfinns på https://tidningar.kb.se (Sverige) och https://digi.kansalliskirjasto.fi/etusivu (Finland).
23. Mika Koistinen, Kimmo Kettunen & Tuula Pääkkönen, 'Improving optical character recognition of Finnish historical newspapers with a combination of fraktur & antiqua models and image preprocessing', *Proceedings of the 21st Nordic Conference on Computational Linguistics* (Göteborg 2017), s. 277–283, https://ep.liu.se/ecp/131/038/ecp17131038.pdf.
24. Eetu Mäkelä, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi & Terttu Nevalainen, 'Wrangling with non-standard data', Sanita Reinsone, Inguna Skadiņa, Anda Baklāne & Jānis Daugavietis (eds.), *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference* (CEUR-WS.org, Tyskland 2020), s. 91–96, http://ceur-ws.org/Vol-2612/paper6.pdf.

mycket liknande innehåll. I praktiken ger detta ytterligare kluster och resulterar i ett större klusterantal än faktiska fall av upprepning. Därför måste man ha en kritisk blick på det exakta antalet kluster och behandla det som ungefärligt. Det totala antalet kluster i databasen är dock cirka 17,8 miljoner, vilket tyder på att mängden upprepningskedjor är mycket stor.

Trots att databasen fungerar som ett sätt att studera hur texter spreds mellan orter och regioner avslöjar den egentligen ingenting om rörelsens faktiska kronologi och precisa väg genom rummet. Den utgår enbart från publiceringsdatumen och utgivningsorterna i en specifik kedja av återvunnen text. Det betyder inte nödvändigtvis att tidningarna faktiskt har citerat varandra i just den ordningen. För att illustrera svårigheten med att dra enkla slutsatser om enskilda klusterfall kan vi föreställa oss ett textstycke som såg dagens ljus först i Helsingfors den 30 juli 1850, därefter i Stockholm den 10 augusti och sedan i Göteborg den 20 augusti. Detta påhittade men realistiska exempel kan innebära 1) att samtliga tidningar trycker om ett brev eller en annons som skickats till dem, 2) att alla tre publikationerna oberoende av varandra citerar en fjärde tidning (som inte ingår i datasetet), 3) att Stockholms- och Göteborgsredaktionen båda prenumererar på Helsingforstidningen och citerar direkt ur den, 4) att Göteborgstidningen citerar Stockholmstidningen som i sin tur citerar den i Helsingfors, eller 5) att en kombination av dessa scenarier försiggick. Med andra ord krävs mer information för att fastställa den faktiska fysiska (eller elektriska) transporten av innehåll. Ibland finns denna information i de texter som dessa passager ingår i (de anger helt enkelt sin källa mer eller mindre explicit), ibland behöver annat material konsulteras, men oftast är det omöjligt att avgöra frågan med bestämdhet. Å andra sidan, om det aktuella fallet ses i termer av en immateriell spridning av innehåll, ja, då förflyttade sig verkligen något i tid och rum, från Helsingfors via Stockholm till Göteborg. Därför måste även termer som hamnstad (*port city*), inkommande stad och viralitetspoäng, som gränssnittet använder, brukas med både försiktighet och fantasi.

Historievetenskaplig betydelse

Även om tidningar vanligtvis inte studeras genom enskilda författare, belyser studier av återbruk av text i tidningar ytterligare hur texter inte bara skrevs av enskilda författare utan samlades in och återpublicerades för olika syften av andra aktörer. Ryan Cordell analyserar detta som ”the network author”.[25] Men textåteranvändning belyser mer än en omtolkning av författarskapet, nämligen hur mottagande och återpublicering fungerat som historiskt aktörskap på ett sätt som historievetenskapen vanligen inte uppmärksammat. En text kan ha varit mycket viktig – en omedelbar klassiker – från allra första början, men ibland har dess betydelse cementerats först efter ett aktivt mottagande. Det är alltså lätt att uppfatta databasen och dess diagram och kartor som att de kort och gott handlar om spridning, men det oftast mer korrekta och produktivare synsättet betonar i minst lika hög grad själva mottagandet. Samtidigt vill vi betona vikten av att hålla båda dessa synsätt aktuella. Man kan till exempel fråga sig vilka strategier olika aktörer använt sig av för att befordra mottagandet av deras texter.

Med databasens hjälp går det att studera textåteranvändningspraktiker både kvalitativt och kvantitativt. Kvalitativa fall kan innefatta att studera en viss text och dess återanvändnings- och rekontextualiseringshistoria i olika tidningar i Finland och Sverige. Men de kan också utgå från ett visst tema eller ämne som definieras genom en nyckelordssökning eller fokusera på särskilda städer eller tidningar definierade utifrån tillgängliga metadata. Alla sökresultat kan sedan laddas ner och analyseras ytterligare med andra verktyg. En studie som analyserar textåteranvändning mellan Göteborg och Viborg kan till exempel ladda ner sökresultat och använda ett korpuslingvistiskt verktyg för vidare analyser.

Man bör heller inte underskatta de insikter om själva tidningspressen som kunskapen om återbruket ger. Att det vi i vanliga fall betraktar som journalistik återanvändes med eller utan tillstånd, har tidigare forskning ofta uppmärksammat.[26] Men att detta kvantitativt sett inte utgjorde huvudsaken i tidningarna utan både det som de inne-

25. Cordell, ’Reprinting, circulation, and the network author in antebellum newspapers’.
26. Se exempelvis Erik Edoff, ’Stockholm mot landsorten: Stockholmsbrev och 1800-talets landsortspress som nätverk’, Alf Arvidsson (red.), *Spaningar i kultursektorn* (Umeå 2019), s. 77–92.

höll mest av och det som verkligen flödade mellan dem var en typ av generiska texter, eller schabloner, som anpassades från fall till fall. Den typen av systematisk klipp-och-klistra-praktik framträder tydligt i databasens resultat.

Sökningarna ger med andra ord även data som kan användas vid mer kvantitativa ingångar till textåteranvändningsmönstren. Som nämns ovan erbjuder de identifierade klustren inte alltid ett-till-ett-matchningar med historiska återpubliceringar av texter på grund av komplexiteten i layoutförändringarna, men siffrorna ger en bra indikation på hur många texter som cirkulerade, varifrån de härstammade och vilka tidningar som återpublicerade dem. Databasen ger till exempel siffror på hur vanligt det var att texter återanvändes över Bottniska viken jämfört med inrikes svensk textåteranvändning. Siffrorna kan ytterligare begränsas geografiskt och efter datum, vilket gör det möjligt att studera regionala påverkansmönster över tid. Eventuella tematiska intressen, definierade genom en nyckelordssökning, kan också kvantifieras genom de siffror som gränssnittet tillhandahåller.

Både de kvalitativa och kvantitativa fallen vittnar om de kulturella asymmetrier som fanns mellan Finland och Sverige. Å ena sidan dominerar svensk press textåteranvändningsklustren antalsmässigt. Å andra sidan komplicerar enskilda fall bilden, med exempel där påverkan är omvänd eller helt enkelt inte så klar som den vid en första anblick verkar. Det finns en betydande agens i mottagandet och återanvändningen av texter, och återanvändningen av svenska texter i Finland bör ses både som ett bevis på svenskt inflytande – och som ett finskt val att upprätthålla en kulturell anknytning. Denna koppling tycks ha blivit viktigare mot slutet av 1800-talet när språkrelationerna i Finland blev mer ansträngda.[27] På den tiden hämtade svenskspråkiga tidningar i Finland betydande symbolisk kraft genom att låna från tidningar i Sverige.

Sammanfattning

Denna text presenterar en databas och dess tillhörande gränssnitt som ett verktyg för att studera transnationella informationsflöden. De kan vara till nytta för forskare från olika ämnestraditioner, men förutsätter

27. Max Engman, *Språkfrågan. Finlandssvenskhetens uppkomst 1812–1922* (Helsingfors 2016).

goda kunskaper i svenska språket och om tidningsutgivningens historia i Sverige och Finland. De är ett resultat av ett forskningsprojekt som historiskt vill förstå informationsflöden mellan Finland och Sverige under det långa 1800-talet. Vi har emellertid vinnlagt oss om att användbarheten ska vara mycket bredare än så. De specifika parametrarna och funktionerna arbetades fram stegvis samtidigt som vi först lärde av en testdatabas innan vi gick vidare med den större databasen. Allt eftersom vi har gått vidare har nya idéer och behov uppstått, och vi räknar med att utveckla gränssnittet ytterligare. I presentationen här har vi betonat fördelarna med databasens funktioner och även lyft problem med den. Användaren bör ha dess olika begränsningar i åtanke, och ett sätt att göra det är att regelmässigt kontrollera sina resultat i de ursprungliga, nationella digitala tidningssamlingarna.

Enligt våra inledande undersökningar är databasen väl lämpad för att upptäcka transnationella textflöden, vilket ger stora möjligheter att filtrera resultat och visualisera återanvändningskluster. Det fanns till exempel storskaliga transnationella marknadsföringskampanjer som databasen bidrar till att synliggöra. Detta är möjligt i de fall där annonsen innehåller tillräckligt med oformaterad text (mer än 300 tecken). Likaså fungerar databasen väl vid studiet av nyheter som sprids snabbt eller viralt i tidningsnätverket. På grund av dess långa tidsspann, från 1600-talet till det tidiga 1900-talet, är databasen användbar även för att utforska hur texter reste i tiden och hur de trycktes om senare i historien. Databasen *Text Reuse in the Swedish-language Press, 1645–1918* omfattar med andra ord flera nivåer av historisk representation. För det första fungerar databasens gränssnitt som en representation av de kluster som hittades under själva den datorstödda processen att identifiera textåterbruk. Klustren i sig är bara kedjor av textpassager, varför de måste berikas genom tillgängliga metadata så att deras dimensioner kan utforskas. För det andra syftar gränssnittet till att representera det förflutnas textflöden. Detta handlar inte bara om klustren utan om den faktiska migrationen av texter i den historiska världen från 1600-talet till 1900-talet. I denna mening fungerar databasens gränssnitt som en historisk tolkning, även om man måste komma ihåg att kluster inte utan förbehåll kan betraktas som verkliga textåteranvändningskedjor. En kedja i den historiska världen kan i databasen delas upp i flera kluster på grund av datans kvalitet och processens natur. Utöver dessa två representationslager finns det också ett tredje: I det förflutna såg och

förstod historiska aktörer själva att det förekom textåteranvändning. Deras möjligheter att förstå omfattningen av tidningsnätverket eller registrera dess vindlingar ur ett fågelperspektiv var emellertid mycket begränsade. Samtidigt låter databasen användaren också undersöka aktörernas position, särskilt i situationer då de utnyttjade nätverket genom marknadsförings- och informationskampanjer.

Vi har strävat efter att hitta ett nytt sätt att förstå gränsöverskridande informationsflöden. Det finns många användbara databaser för historisk forskning, men det som är nytt i detta fall är att vi, utifrån datorstödda metoder, har konstruerat en databas som kopplar samman och kombinerar digitaliserat innehåll från flera länder och möjliggör studier av textflöden.[28] Databasen har utvecklats under ledning av forskarna själva och innehåller funktioner som är skräddarsydda för att utforska textåteranvändning och informationsflöden. Detta är viktigt även ur en mer generell synvinkel. Digitaliserade tidningssamlingar har i regel byggts på nationell grund, genom initiativ från nationalbibliotek världen över. Detta har utan tvekan varit ett värdefullt arbete, men samtidigt har de digitaliserade samlingarna begränsats till nationella domäner. Vår databas och dess gränssnitt erbjuder en väg att gå bortom tidningsarkivens nationellt lagrade historia och mer brett bidra till en mer transnationell förståelse av historien.

Forskningsprojektet Informationsflöden över Östersjön: Svenskspråkig press som kulturförmedlare, 1771–1918 (2020–2023) är finansierat av Svenska litteratursällskapet i Finland.

28. Se till exempel Jari Eloranta, Pasi Nevalainen & Jari Ojala, 'Towards big data. Digitising economic and business history', Mats Fridlund, Mila Oiva & Petri Paju (eds.), *Digital Histories. Emergent Approaches within the new Digital History* (Helsingfors 2020), s. 45–67, https://doi.org/10.33134/HUP-5-3.