

Johan Jarlbrink, Pelle Snickars & Cristian Colliander

Maskinläsning

Om massdigitalisering, digitala metoder och svensk dagspress

Abstract

This article highlights the media historical possibilities to analyse linguistic patterns in massive amounts of texts using digital methods. Our starting point is the fact that The National Library of Sweden has made over 12 million newspaper pages available in digital format. An important question is how to research them. The article presents a media history of the Swedish newspaper digitisation, as well as new ways of conducting historical newspaper research using digital methods. A case study is presented where the conceptualisation of a new media technology (the internet) in newspapers from the 1990s is tracked with a digital tool searching for word co-occurrences. The possibilities of digital methods are often incredible, but we should not underestimate the problematic aspects of using digital tools to explore digitised newspapers. The poor quality of the OCR (Optical Character Recognition) is described as one of the major challenges facing historical newspaper research in a digital environment.

Keywords: media history, digitized newspapers, OCR, digital humanities, text analysis

Hösten 2014 samlades ett sjuttioal bibliotekari-er, forskare och representanter för tidnings-branschen på British Library i London för en konferens om digitalisering av äldre dagspress inom ramen för kulturarvsportalen Europeana. 'Europeana Newspapers' har ambitionen att digitalisera 18 miljoner tidningssidor, varav drygt hälften ska vara sökbara i fulltext genom så kallad 'optisk teckenigenkänning'. Just OCR, 'Optical Character Recognition', var ett av konferensen mest omdiskuterade teman. Det är inte förvånande eftersom OCR på flera sätt är den teknik som gör digitaliserad dagspress till en helt ny typ av forskningsmaterial. OCR är en process i flera steg som konverterar dagspressen som medium – från tidning till maskintext. Det sker genom en närmast cirkulär textuell migreringsprocess där tidningssida blir inskannad bild, vilken därefter omkodas

till maskinläsbar text. Ur den senare kan dels skapas en sorts inherent data, extrapolerad ur det inskannade materialet – vilket är en av digitaliseringens stora fördelar – dels möjliggör den maskinkodade texten storskaliga kvantitativa och jämförande analyser av exempelvis ords samförekomster.

Samtidigt är OCR-tekniken fortsatt behäftad med problem, vilket London-konferensen också resonerade kring. Har vi nått 'peak OCR' var det flera deltagare som undrade. Särskilt tröstlöst tycktes läget vara i länder där frakturstilen dominerat. 1930 publicerades exempelvis fortfarande omkring 60 procent av alla trycksaker i Tyskland i fraktur, och frakturstilen var även vanlig i Sverige fram till 1800-talets slut. Men även den latinska tryckstilen, antikva, orsakar igenkänningsproblem menade andra. Det marknadsledande

företaget ProQuest, och deras inskanning av 472 brittiska tidskrifter mellan 1681 och sent 1930-tal, ger exempelvis fantastisk access till ett omfattande källmaterial. Samtidigt har forskare påpekat att till och med ProQuest sofistikerade OCR-algoritmer fortsatt inte riktigt går att lita på. I en genomgång av viktoriaiska texter om Kuba i det inskannade tidskriftsmaterialet, har Albert D. Pionke bland annat visat på betydande svårigheter för Proquests algoritmer att urskilja ordet 'Cuba' från ord som: 'cash', 'Cheap', 'Colours', 'cube' eller 'cure' (Pionke 2014: 391). Det säger sig självt att sådana 'bad matches' kan leda till felaktiga slutsatser och forskningsresultat. OCR-kodad text är med andra ord synnerligen användbar – men långt ifrån felfri.

Den här artikeln innehåller dels en mediehistoria kring den svenska dagspressens digitala förhistoria (med fokus på mikrofilm), dels resonerar den kring de möjligheter som nya digitala metoder ger den mediehistoriska forskningen. Utifrån en mediehistorisk infallsvinkel uppmärksammar texten den digitala teknikens möjligheter att på ett generellt plan analysera stora textmängder på jakt efter lingvistiska mönster, bland annat genom så kallad distansläsning, en sorts data-metodologisk läsning av stora textmassor som populariserats av bland andra litteraturvetaren Franco Moretti med fokus på 1800-talets skönlitteratur (Moretti 2013). För att illustrera metodernas möjligheter och problem görs en teststudie på ett lättillgängligt material från 1990-talet.

Dagspressen remedierad – från mikrofilm till OCR

I bibliotekskretsar, kunde man läsa i 1949 års direktiv rörande biblioteks- och arkivfilmning, uppmärksammas numera den så kallade 'mikrofilmningen' vilken innebär "att man i stark förminskning fotograferar den handling, som man önskar kopiera. ... Vad biblioteken beträffar är i detta sammanhang spørgsmålet om bevarandet av de dagliga tidningarna särskilt betydelsefullt." Möjligheten att kopiera dagstidningar var kring 1950 svaret på två brännande frågeställningar när det

gällde informationshantering: dels vållade det tråhaltiga pappret i dagstidningar att de (i regel) hade mycket kort livslängd, dels utgjorde tillväxten av antalet tidningar ett skrymmande utrymmesproblem (SOU 1951: 36: 7).

Mikrofilmning av svenska dagstidningar startade i privat regi redan i slutet av 1940-talet genom företaget Rekolid/Cefab (Centrala filmarkivet AB). Verksamheten var delvis ett sätt att svara upp mot tidningsbranschens behov, men även Kungliga biblioteket började snart att beställa filmade tidningskopior. Rekolid/Cefab kom under 1950-talet att samarbeta med både KB och Lunds universitetsbibliotek. Sammantaget mikrofilmade Rekolid/Cefab cirka 360 tidningstitlar fram till 1970-talet (Widholm 2011). I en rapport från Statskontoret 1969 påtalades att mellan 1850 till 1950 hade cirka 30 procent av alla tidningssidor mikrofilmats, och för perioden därefter uppemot 45 procent (Statskontoret 1969). När en ny utredning gjordes vid mitten av 1970-talet bedömdes att ungefär hälften av alla dittills utkomna tidningar hade filmats. En del av de befintliga mikrofilmerna var dock av så dålig kvalitet att omfilmning krävdes. Eftersom tidningarna var inbundna var det ibland svårt att få hela sidan läslig på bild. Ofta var det innerspalten som var oläslig, och där placerades bland annat ledarartiklar och presskommentarer, av tradition centrala texter i presshistorisk forskning. Överföring från ett analogt medium till ett annat innebär i princip alltid förlust av data, även om överföringen i övrigt förlöpt som det är tänkt. Vad gäller mikrofilmningen av dagspress anfördes också att den premierade texter på bildernas bekostnad; egenskaper som storlek och papperskvalitet gick naturligtvis också förlorade (Tidningsfilmningskommittén 1976).

Såväl mediehistoriskt som informationshanteringsmässigt kan man notera att om mikrofilmtekniken redan kring 1950 framför allt knöts till tidningsmediet – både som ett lovande bevarandeinstrument och som ett sätt att tillgängliggöra presshistorien – så förknippades automatisk databehandling (ADB) och tidiga digitaliseringsinsatser inom bibliotekssektorn under 1970-talet snarare till

bokmediet. Anledningen har delvis att göra med bristfällig OCR-teknik där presslägggen innehöll alltför mycket brus, för att tala med en informationsteoretisk vokabulär. Boksidor var grafiskt renare, och därför enklare att digitalisera (åtminstone framstod de så i teorin). Mikrofilm och digitalisering är med andra ord relaterade till varandra på flera sätt, och under lång tid existerade de följaktligen som parallella informationshanteringstekniker. Inte förrän i den så kallade KB-utredningen 2003 påtalades att KB borde ges ”uppdraget att kartlägga fördelar och nackdelar med att ersätta mikrofilmningen med digital leverans och lagring” (SOU 2003: 129: 142).

KB avslutade sin reguljära mikrofilmning av svensk dagspress så sent som 2013. Det var en verksamhet som i egen regi hade påbörjats 1979 (genom att delvis införliva Rekolid/Cefab), detta parallellt med att Arkivet för ljud och bild påbörjade sin verksamhet – baserad på en utvidgad pliktexemplarslag för audiovisuella medier – där framför allt etermedier kopierades på video- och magnetband. *Medietransfer* var här den gemensamma nämnaren. Mediernas materialitet nedprioriterades, samtidigt som innehållet uppgraderades till att bli vad som primärt värdesattes för forskning och som framtida kulturarv. Det är en underskattad förändring i mentalitetsarkivariskt hänseende – innehåll ersatte form. Omställningen i den tankemässiga och praktiska hanteringen av medialt kulturarv som verksamheten på ALB introducerade förklarar också varför den institutionen (vilken sedermera bytte namn till Statens ljud- och bildarkiv) var långt mer framgångsrik (än KB) på att ställa om sin verksamhet i digital riktning. Om innehåll kopierades till magnet- och videoband eller till digitala bärare var strängt taget av mindre betydelse – innehållet var allt, formatet intet.

Att tidningar och böcker skulle digitaliseras och bibliotekens kortkataloger göras tillgängliga över nätverk kan från dagens horisont förefalla vara en självklarhet. Men teknikomställningen inom bibliotekssektorn har varit långsam. Under 1960-talet var det snarare genom datoriseringen av bank- och försäkringsbranschen som digitaliseringens

möjligheter gradvis uppenbarades. Just maskinläsning – som enligt Computer Swedens datalogiska språkwebb bör vara den svenska översättningen av OCR – har en i sammanhanget intressant medie-ekonomisk historia. Under 1960-talet började nämligen checkar och inbetalningskort att OCR-behandlas. För det krävdes att speciella typsnitt användes för att tekniken skulle fungera, först i mitten på 1970-talet utvecklade datavisionären Ray Kurzweil en mjukvara som kunde OCR-läsa i princip alla typsnitt. Det var en produkt som kopieringsföretaget Xerox köpte 1978 för att underlätta överföring av elektronisk information till papper, och därefter har tekniken kommersialiserats på bred front (Schantz 1982).

Att just Xerox övertog patent på tekniken är inte konstigt eftersom all OCR-behandling innebär ett slags omkopiering, där algoritmen i ett första steg omformar färg eller gråskala i originaldokumentet till bi-tonala bilder i svart-vitt – och därefter läser om den analoga texten till binär data. Det är den processen som ofta orsakar problem då alla tecken inte uppfattas på exakt rätt sätt av programvaran. Beträffande dagspress är det i regel så att ju äldre tidningsmaterialet är desto sämre fungerar OCR-tolkningen. En äldre tidningssida kan exempelvis ha blivit nött av användning eller missfärgad, vilket inte sällan resulterar i en ojämn mörkerskala på satsytan. Den senare är svår att exakt OCR-läsa, även om programvaror som exempelvis Tesseract använder sig av algoritmisk efterbearbetning för skannade ark. ‘Optical Character Recognition’ handlar som termen antyder om *igenkännbarhet* på teckenbildsnivå, vad som på svenska kallas ‘glyf’. När teckenbilderna är grafiskt svårläsliga – som i äldre presslägg med frakturstil – blir de också svåra att maskinläsa. Det gäller naturligtvis också om tidningens satsyta hamnat snett i skannern; OCR-algoritmen är helt beroende av raka, linjära rader. OCR-algorithmens svårigheter gäller därför dels *urskiljandet* av specifika glyfer, dels *överensstämmelser* mellan olika teckenbilder, som exempelvis *a* respektive *α* – vilka är två glyfer som representerar samma tecken. Att grafisk igenkänning kan

vara knepig är egentligen uppenbart för var och en; skillnaden mellan glyfen ‘T’ och ‘!’, ‘J’, ‘I’ eller ‘↑’ är trots allt inte så stor – speciellt inte för en dator. Under det senaste två decennierna har både kommersiella företag liksom olika nationalbibliotek därför strävat efter att optimera OCR-kodningen, inte minst genom införandet av dynamiska begreppsordböcker (tesaurus) som OCR-tolkningen matchas mot. Generellt brukar det för äldre dagspress hävdas att om mer än tio procent av innehållet är felkodat så är OCR-tolkningen av dålig kvalitet. Över 90 procent OCR-överensstämmelse är däremot för de flesta aktörer ett acceptabelt resultat (Holley 2009).

KB digitaliserar

Genom en tilltagande datorisering av biblioteks- och arkivväsendet under 1970- och 1980-talet uppenbarades digitaliseringens potential för alla inom ABM-sektorn. I Sverige tog den digitala omställningen dock lång tid – inte förrän 1997 introducerade KB exempelvis genom Libris sitt så kallade webbsök. Ett år senare påbörjades projektet, ‘Tiden – Digitalisering av historiska tidningar i Norden’, ett skandinaviskt samarbetsprojekt för att utveckla metoder för digitalisering av tidningar. I projektet prövades olika OCR-läsningstekniker av både *antikva* och *fraktur* för att göra tidningarna sökbara. Men verksamheten syftade främst till att höja kompetensnivån kring tidningsdigitalisering. Intern och tekniskt orienterad metodutveckling har på många sätt varit ledstjärna för i princip samtliga av KBs digitaliseringsinsatser – snarare än resultat i form av digitalt tillgängligt material för forskare och andra intresserade. Efter projektet ‘Tiden’ följde under 00-talet ytterligare två försök att få fart på tidningsdigitaliseringen: ‘Digitalisering av svensk dagspress’, samt ett EU-finansierat projekt ‘TELplus’. Resultaten var ånyo magra; inalles skannades bara omkring 200 000 sidor tidningstext från 24 svenska dagstidningar i ett försök att visa en helhetsbild av svensk dagspress. KB tog under denna period föga intryck av Googles omtalade bokskanningsprojekt.

Snarare präglades även de två nya digitaliseringsprojekten av ett internt sökande efter lika optimala som utsiktslösa metoder och tekniker, där KB bland annat provade att skanna dagspress från mikrofilm. Eftersom pressmaterialet till största delen bestod av äldre lägg från 1700- och tidigt 1800-tal (då frakturstilen dominerade) blev OCR-läsningen därefter. Stickprov ger vid handen att knappt en enda mening tolkades rätt.

På detta minst sagt valhända sätt fortsatte KB att hantera frågan om digitalisering av dagspress, ett av nationalbibliotekets allra mest använda forskningsmaterial. Så sent som i årsredovisningen för år 2008 kunde riksbibliotekarie Gunnar Sahlin påpeka att ”dags- och veckotidningar behöver digitaliseras i stor skala. KB har inlett ett arbete men ytterligare resurser krävs” (KB årsredovisning 2008: 5). Vid denna tidpunkt hade de flesta nationalbibliotek i Europa redan digitaliserat stora delar av sin dagspress. I digitaliseringshänseende var Sveriges nationalbiblioteket kring 2010 ett av Europas minst utvecklade. Genom en beviljad ansökan om medel ur EU:s strukturfonder kopplat till Riksarkivets mediekonverteringscentral (MKC) i Västernorrland lyckade KB emellertid få till stånd ett pilotprojekt – återigen dock med syfte att ”utveckla kompetens, arbetsrutiner och teknik för digitalisering av dagstidningslägg.” Projektet, påtalades det i årsredovisningen 2009, ”lägger grunden för framtida digitalisering av svenska dagstidningar i stor skala” (KB årsredovisning 2009: 11). Genom att även andra finansörer gick med, bland annat Schibstedtkoncernen, lades grunden för digitaliseringsprojektet Digidaily.

Digidaily har under de senaste åren (tillsammans med Kungliga bibliotekets migrering av audiovisuella medier) varit nationalbibliotekets digitala flaggskepp. Efter snart två decennier av idel tveksamheter har Digidaily faktiskt inte enbart resulterat i tekniska metoder och kravspecifikationer, utan också i konkreta resultat i form av massdigitaliserade tidningslägg. I skrivande stund är mängden sökbara tidningssidor strax över 12 miljoner – men majoriteten av dem kommer från nutida tidningar, från 2012 och framåt. Sök-

tjänsten på tidningar.kb.se, 'Sök bland svenska dagstidningar', innehåller för närvarande hela *Svenska Dagbladet* och *Expressen* och nästan hela *Dagens Nyheter* och *Aftonbladet*. Därtill finns *Dagens Industri* från 1983 till idag samt *Göteborgs Aftonblad* mellan 1888 och 1918. KB tolkar emellertid upphovsrätten strängt, och material publicerat senare efter 1900 är upphovsrättsskyddat och kan endast läsas lokalt i Stockholm.

Digidaily är inte minst ett lyckat projekt eftersom kostnaden per digitaliserad tidningssida ligger långt under den internationella prisbilden. 2010 redovisade Collection Trust till EU-kommissionen en kostnadsrapport (Poole 2010) av vilken det framgår att kostnaderna per digitaliserad tidningssida i projekt som bedrivs vid British Library ligger på cirka £1. En genomsnittssida i Digidaily produceras däremot för sju svenska kronor. Resultaten från Digidaily-projektet visar dock att priset är starkt beroende av tidningssidans beskaffenhet; ju sämre kondition en sida befinner sig i desto högre kostnad och desto sämre kvalitet på OCR-resultatet.

Genom Digidaily är de fysiska förutsättningarna för digitaliseringsprocessen färdiggetablerade när det gäller historiska dagstidningar. KB:s tidningsarkiv skiljer sig från många andra nationalbiblioteks eftersom man i början av 2000-talet beslöt att ta hand om Uppsala universitetsbiblioteks tidningssamling. Det betyder att KB idag förfogar över en nästan komplett dubbeltsamling, vilken med fördel kan användas för digitalisering genom att läggen på intern jargong 'slaktas' – bara så kan digitaliseringspriser pressas och en kritisk massa av skannad dagspress uppnås. Noterbart är att KB vid flera tillfällen äskat extra medel från regeringen (bland annat till 2012 års forskningsproposition) för att försätta att digitalisera merparten av den svenska dagspressen inom ramen för Digidaily. Den totala dagstidningssamlingen på KB omfattar 122 miljoner sidor, och nationalbiblioteket har kalkylerat att det reguljära anslaget skulle behöva ökas med omkring tio miljoner per år (i tio år) för att digitalisera det för forskningen mest värdefulla tidningsmaterialet.

Digitala metoder – från närläsning till distansläsning

Vad kan ni göra med en miljon inskannade böcker? Eller en miljon tidningssidor? Sådana frågor har sedan 2009 ställts av ett antal anglosaxiska forskningsstiftelser i utlysningen om forskningsmedel inom den så kallade 'Digging into Data Challenge'. En rad utlysningar har gjorts, och den övergripande frågan har gällt hur ny tillgång till enorma mängder digitaliserat (kulturarvs)material påverkar den humanistiska forskningen – praktiskt såväl som teoretiskt. Vilka nya frågor (och svar) kan genereras av datorkraft applicerad på ett slags kvantitativ empirisk skalbarhet? Termen 'data mining' anförs ofta i sammanhanget, men metodfrågor har också varit vanliga. Givetvis är det utmärkt att digitalisering förenklar tillgängligheten till forskningsmaterial; för mer human-tekniskt orienterade forskare ställer digitaliseringsprocessen dock framför allt frågor kring nya digitala metoder liksom en uppdaterad digital forskningspraktik (Rogers 2013). Från ett forskningsperspektiv ligger digitaliseringens stora betydelse i att vetenskapligt kvantitativa forskningsperspektiv på allvar kan tas i bruk inom humaniora, och i så måtto både utveckla och komplettera de traditionellt kvalitativa hermeneutiska läsningarna utifrån förmodade nyckeltexter.

Att utföra sökningar av stora datamängder var länge förbehållet naturvetare och samtidsorienterade samhällsvetare, men med historisk dagspress i massdigitaliserad form kan även humanister (tillsammans med kulturarvsinstitutioner) utföra storskalig analys av enorma mängder data; det gäller både 'hårda' mätbara uppgifter kring diverse ekonomiska förhållanden såväl som 'mjuka' kring begreppsanvändning och diskursiva frågor. Inom det växande forskningsfältet kring digital humaniora kan man just skönja hur modern datateknik och växande digitala samlingar av historiskt material håller på att förändra (delar av) historievetenskapen i en storskalig och kvantitativ riktning (Svensson 2011, Svensson 2012). I korthet, handlar det om att bedriva historisk forskning som datorbaserad vetenskap, snarare än som mer eller

mindre godtycklig hermeneutisk praktik. Vad som förr räknades som heroiska insatser med år av mikrofilmsrullande kan nu göras av en dator på nolltid – åtminstone i teorin. Frågeställningar har givetvis skiftat, men en betydande fördel med digital humaniora-inspirerad metod och teknik är att nya storskaliga frågor kan ställas och besvaras.

Den presshistoriska forskningen i Sverige och internationellt har framförallt dominerats av studier som ägnat sig åt enskilda genrer, skribenter eller texter knutna till specifika händelser. Om större grepp har tagits har det ofta varit tidningsföretagen snarare än innehållet som stått i fokus. Undantagen utgörs framförallt av ett antal kvantitativa studier från 1970-talet. Symptomatiskt är att de använde sin tids datormöjligheter för att söka och kvantifiera. Eftersom OCR-tekniken inte fått något bredare genomslag kvävdes emellertid betydande mänskliga resurser för att excerpera och koda (Jarlbrink 2015). I övrigt har de kvalitativa studierna dominerat. En viktig anledning till detta är att presshistorisk forskning ofta haft rötterna i litteraturvetenskapliga traditioner. Att närgranska kanoniska texter såsom följetonger, ledare och kulturartiklar, skrivna av tongivande skribenter, ligger i linje med traditionella litteraturvetenskapliga angreppssätt. Ofta har forskningen kretsat kring 'bra' artiklar, där det funnits ett behov av att visa att även den flyktiga dagspressen innehållit texter av litteraturhistoriskt värde. En annan förklaring till detta något snäva perspektiv är att det har varit svårt att orientera sig i materialet. Antalet tidningar och texter har varit överväldigande. Ingen har kunnat läsa allt, och den som sökt efter något specifikt har sällan haft några register att gå efter. Vad som har funnits är bibliografier för några få titlar och år, men de har haft flera begränsningar. Bibliografiernas upplägg har ofta verkat styrande för vad som varit möjligt att hitta och beforska. *Svenskt pressregister*, som en gång initierades på Litteraturvetenskapliga institutionen i Lund, har exempelvis förtecknat poesi och prosa, kulturartiklar och signerade texter, det vill säga sådant som intresserar den litteraturhistoriskt orienterade pressforskaren.

Osignerade texter i andra genrer – som alltid dominerat i spalterna – har valts bort och osynliggjorts. I det läget har många forskare valt att koncentrera sig på några få kanoniska tidningstitlar, på enskilda genrer eller författarskap, eller på texter som varit möjliga att hitta eftersom de publicerats i samband med någon specifik händelse.

Mot bakgrund av ovanstående kritik av traditionerna inom den presshistoriska forskningen ska det samtidigt medges att den forskning som försökt analysera presshistoriskt material med hjälp av digitala metoder än så länge inte alltid har varit mycket bättre. I flera studier har forskare experimenterat med metoder, men resultaten har ofta varit triviala. En forskargrupp har exempelvis använt 'topic modeling', en metod för att analysera ords samförekomster, för att bearbeta 232 000 sidor ur en rad tidningar från Texas mellan 1865 till 1930. Ett resultat är att det skrivs mycket om bomull i dessa tidningar – vilket de flesta presshistoriker hade kunnat räkna ut utan avancerad programvara (Yang, Torget & Michalcea 2011).

En anledning till att den här typen av forskningsresultat än så länge ofta framstår som närmast banala ur ett presshistoriskt perspektiv, är att forskningsprojekten i regel drivs av informations- och datavetare. De är mer intresserade av att testa och utveckla programvara än att åstadkomma mediehistoriskt intressanta analyser (t.ex. Hendrickx et al. 2013; Kushkuley 2012). Men givetvis existerar det undantag. Som en mer presshistoriskt driven pionjär framstår exempelvis Sharon Block, som tillsammans med datavetaren David Newman använt så kallad 'pLSA' (Probabilistic Latent Semantic Analysis) för att kartlägga samförekomster av ord och de ordkluster – 'topics' – som samförekomster av högfrekventa ord bildar. Materialet bestod av 80 000 artiklar och annonser i *Pennsylvania Gazette*, från 1728 till 1800. Block och Newman kunde visa hur enskilda ordkluster (exempelvis rörande statstyre, religion och brottslighet) ökade eller minskade över tid, hur enskilda ordkluster relaterade till varandra, och vilka kluster ett enskilt ord med störst sannolikhet

ingick i. Exempelvis visade det sig att ämnet 'statsstyre' blev mer och mer marginellt från 1730-tal till 1750-tal, för att därefter öka kraftigt igen under 1700-talets andra hälft. Ämnet 'textilier' gjorde en motsatt resa. Analyserna av vilka ordkluster ett enskilt ord med störst sannolikhet ingick i, kunde dessutom synliggöra hur marginaliserade vissa grupper av människor var. Block och Newman kunde till exempel visa att ordet 'Negro' nästan enbart förekom inom klustret tjänare/slav – och även ordet 'woman' återfanns oftast inom samma kluster (Newman & Block 2006).

Block och Newman har visat att det är fullt möjligt att använda digitala metoder för att synliggöra mönster som hade varit mycket svåra att upptäcka med traditionella metoder. Men de har också påpekat att metoderna mycket väl kan kombineras, att den digitala analysen kan visa på samband som kan undersökas vidare med andra metoder (Newman & Block 2006, Block 2006). Få senare studier har visat de digitala metodernas fördelar lika framgångsrikt som Block och Newman, även om en del forskare försökt. Bob Nicholson (2012) har med enkla sökningar i British Librarys digitala tidningsarkiv visat hur samförekomsterna av orden 'konkurrens' och 'Amerika', respektive 'Frankrike' och 'Tyskland', varierar under 1800-talets andra halva – och att artiklarna om utländsk konkurrens tycks samvariera med spannmålsskördarna. Colella (2013) och tidigare anförda Pionke (2014) har använt något mer avancerade metoder för att undersöka så kallade 'nyckelord i kontext' ('Key Word in Context'). I Collellas fall handlar det om hur affärslivet beskrivs i brittisk 1800-talspress. I Pionkes fall om handel och slavarbete i artiklar om Kuba. Huijnen et. al (2014) och Eijnatten et. al. (2014) använder en liknande metod för att kartlägga diverse ämnen i framförallt holländsk press. Verktuget Texcavator erbjuder här användaren att ange söktermer, precis som en vanlig online-databas, men presenterar sökresultatet i ett ordmoln som visar vilka frekvent använda ord som förekommer tillsammans med sökordet. Detta möjliggör komparationer mellan olika tidsperioder, tidningar och länder. En del av

deras forskning relaterar till medicinhistoriska frågor, och här har Huijnen et. al. (2014) bland annat visat hur eugenik går från att tillhöra en medicinsk diskurs under 1800-talets mitt, till att bli en del av en rasdiskurs i början av 1900-talet. Detta förutsätter dock att forskarna får direkt tillgång till tidningsmaterialet utan att behöva gå via biblioteksportalernas gränssnitt. Förmodligen är det en viktig anledning till att antalet presshistoriska studier som använt digitala metoder än så länge är få till antalet.

Inom delar av det litteraturvetenskapliga forskningsfältet har man hunnit betydligt längre. Orsaken är att de skönlitterära textmängderna är mer begränsade, att upphovsrättsproblematiken är enklare att hantera, samt att skönlitterära texter (vilket tidigare påtalats) som publicerats i bokform är renare, enklare att digitalisera och analysera jämfört med tidningslägg. Matthew Jockers, en av de tongivande forskarna på fältet, presenterar i boken *Macroanalysis* (2013) en rad analyser av engelskspråkiga 1800-talsromaner. Traditionellt har den engelskspråkiga litteraturvetenskapen ägnat sig åt en kanon som bara utgör en mindre del av allt som publicerats. Jockers tar ett större grepp och behandlar sammanlagt 3 346 romaner. Genom att använda verktyg för topic modeling visar han hur olika ämnen ökar och minskar i popularitet, hur ämnesvalen skiljer sig mellan olika länder, och mellan kvinnliga och manliga författare. Genom att mäta och jämföra ord- och ämnesval har Jockers även studerat avståndet mellan olika romaner och författare. Här visar det sig exempelvis att kvinnliga och manliga författare beskriver helt olika världar, och använder olika språk (Jockers 2013, kap. 8-9). De digitala metoder som används gör det möjligt att fånga mycket övergripande mönster, men också att visa likheter och skillnader på detaljnivå. Mänskliga läsare missar förmodligen den relativt höga frekvensen av ordet 'like' i bildningsromaner. När sådana mönster på mikronivå synliggjorts med digitala redskap kan en mänsklig läsare ta vid och försöka förklara vilken betydelse ordvalet har i texten (Jockers 2013, 89ff). Det är såtillvida i samspelet mellan människa och

maskin som den verkliga kunskapsutvecklingen sker.

Digitala lägg

I projektet, 'Digitala lägg. Om pressens gränssnitt 1800' (finansierat av Torsten Söderbergs stiftelse) avser vi att testa de metoder och forskningsperspektiv som beskrivits ovan. Ambitionen är att både utveckla metoderna och att generera mediehistoriskt intressanta forskningsresultat. Som projekttiteln anger är det i första hand 1800-talspressen som kommer att stå i centrum. På grund av strikt tolkad upphovsrättslagstiftning är det tidningsmaterial från 1800-talet som det är lättast att få tillgång till. Detta äldre material är emellertid svårast att arbeta med rent tekniskt. De flesta metoder för topic modeling och kartläggning av samförekomster kräver att varje text är tydligt avgränsad från de omgivande texterna på tidningssidan. När tidningssidor digitaliseras används en programvara som genomför en automatisk segmentering. Eftersom denna i flera fall är bristfällig krävs en manuell rättning av felsegmenterade texter. Digitaliseringen som KB stått för har emellertid hoppat över detta steg, vilket innebär att så gott som varje sida innehåller fel: halva följetongen har klumpats ihop med annonsen för galoscher, ledarartikeln har styckat upp i två separata texter, etcetera.

Materialet kan ändå användas, men kräver speciallösningar. Istället för att utgå ifrån en befintlig – segmenterad – artikelenhet (som alltså ofta är felaktig) kan man exempelvis avgränsa analysen till textsjok eller 'fönster' som innehåller specifika nyckelord. För att prova möjliga tillvägagångssätt har vi genomfört en mindre förstudie på ett mer lättarbetat tidningsmaterial. Vår tanke har varit att denna studie ska kunna tjäna som metodologisk modell för senare undersökningar, samtidigt som resultatet också ska kunna ingå i komparativa analyser. Den övergripande frågeställningen som väglett arbetet har varit hur ett nytt medium språkliggörs när det introduceras och diskuteras offentligt inför en publik av tidningsläsare. Mer konkret har vi velat

undersöka vilka ord som används frekvent i tidningstexter om ett nytt medium, samt hur dessa ord relaterar till varandra. Frågeställningen undersöks med hjälp av ett program som kartlägger ords samförekomster.

Tidigare forskning om nya medier i historien framhäver just introduktionsfasen som särskilt intressant. I ett sådant initialt skede är det inte givet hur mediet ska användas, vilken betydelse det ska få, vem som ska (tillåtas) använda det, hur det ska regleras (Thorburn & Jenkins 2003). När mediet är nytt pågår en sorts *förhandling* om dess betydelse och användning (Marvin 1990). Sådant som senare uppfattas som självklart måste förklaras, den 'svarta lådan' har ännu inte stängts igen (Latour 1987). Tekniker och experter får berätta om mediets möjligheter (Gitelman 2006), samtidigt som kritiker varnar för dess negativa konsekvenser (Gustafsson & Arnberg 2013).

För att undersöka kontinuitet och förändring i hur nya medier beskrivs och diskuteras är vår ambition att göra ett antal nedslag i tidningsmaterial som kan belysa introduktionen av en rad nya medietekniker: dagerrotypi kring 1840, den elektriska telegrafan under 1850-talet, telefonen kring 1880, kinematografi under 1890-tal. Ett första metodförsök gäller emellertid ett betydligt senare medium: internet under 1990-talet. Metoden som använts inbegriper ett antal steg, från insamling och urval till bearbetning och analys. För att åskådliggöra processen kommer den att presenteras kortfattat. Syftet är inte att leverera en slutgiltig analys av insamlade data utan mer att åskådliggöra metodens huvuddrag.

Samtliga artiklar ur *Expressen* och *Göteborgs-Posten* från år 1994 och 1997, tillgängliga i databasen Retriver, samlades in och indexerades med programvaran Lucene (McCandless et al. 2010). Programmet erbjuder avancerade sökmöjligheter såsom booleska sökningar, trunkering och så kallad fuzzy search. De två tidningarna utgör givetvis bara en liten del av det totala tidningsbeståndet, men representerar ändå olika tidningstyper med olika utgivningsort. Det första året har valts för att det representerar ett relativt tidigt skede i nätets (och webbens) publika historia. Mediet

var 1994 inte helt nytt – men ändå en nyhet för de flesta läsare. Det senare året, 1997, har internet och webben fått långt fler användare; då var nätet inte längre en teknik för några få entusiaster.

I de insamlade texterna – 102 616 stycken från 1994 och 131 880 från 1997 – sökte vi efter texter innehållande nyckelordet ‘internet*’. Dessa texter annoterades för att möjliggöra filtrering med avseende på ordklasser. För att kunna slå ihop olika böjningar av ett och samma ord till ett enda ord användes den fritt tillgängliga lingvistikapplikationen Stagger (Östling 2013). Stagger nyttjades i huvudsak till ordklasstagning, och detta material låg sedan till grund för sökningar av ords samförekomst. Länken mellan två ord i graferna (figur 1 och 2) utgörs av det totala antalet samförekomster mellan orden inom ett givet fönster i texterna. I detta fall har fönstret utgjort tre ord: ett ord samförekommer med ett annat om det förekommer tre ord före eller efter ett annat. Styrkan i samförekomsten definieras av en normalisering baserad på frekvensen av två ords samförekomst i relation till den totala förekomsten av dessa ord i textmaterialet, enligt det så kallade Jaccard indexet (Salton & McGill 1982). Om två ord alltid förekommer tillsammans är sambandet alltså starkt, om de två orden förekommer ofta, men bara samförekommer några enstaka gånger är sambandet svagt. För att undvika brus har småord sorterats bort. Listan över stoppord innehåller sextiofem småord i stil med: ‘i’, ‘att’, ‘och’, ‘det’. Beroende på materialet och forskningsintresset kan ord läggas till eller tas bort från stoppordlistan. Våra inledande sökningar i tidningsmaterialet visade att paratextuella ord som bildtext, teckning, Expressen och Göteborgs-Posten också behövde läggas till på listan över stoppord.

Antalet texter som innehöll nyckelordet ‘internet*’ var 101 från 1994 och 2107 från 1997. Samförekomster har sökts bland ord inom ordklasserna substantiv, egennamn och utländska ord. För att inte ytterligare reducera det mindre materialet från 1994, och göra det större materialet från 1997 svårbehandlat, har en minimigräns satts vid tre förekomster 1994

och sju förekomster 1997. Detta innebär att ordpar som endast samförekommer enstaka gånger (färre än tre respektive sju gånger) har sällats bort. Med dessa förutsättningar genererade sökningen 420 ordpar i materialet från 1994, och 2 662 par 1997. Samförekomstanalysen resulterar i data vilken kan representeras som ett nätverk – som i sin tur kan importeras i program för statistisk analys och grafisk visualisering. Här har vi använt visualiseringsverktyget Gephi. Visualiserat som nätverk har våra sökningar i materialet från 1994 genererat 243 noder och 410 länkar – respektive 1165 noder och 2602 länkar för år 1997.

Även med stoppordlistor, ett begränsat fönster och en spärr för lågfrekventa ordpar innehåller nätverken en hel del brus. Ofta utgörs detta brus av fristående nodkonstellationer som består av namn – typ: Rolling och Stones, Kurt och Cobain – som inte länkar till några andra noder. Dessa är det ofta önskvärt att filtrera bort. Samtidigt kan sådana fristående subnätverk vara intressanta i sig: här finns aktörer och företeelser som är en del av materialet men som samtidigt har svag koppling till nätverkets övriga delar. Att noderna ‘Ines-Uusman-kommunikationsminister’ till exempel helt saknar koppling till det övriga nätverket 1997 tyder förmodligen på att hon hade en ytterst begränsad roll i de texter som skrevs om internet detta år, ett resultat som inte är oviktigt i sammanhanget. Ett annat belysande exempel från 1994 är det fristående subnätverket ‘William-Gibson-roman-Neuromanser-begrepp-cyberspace’. Beroende på fönstrets storlek fångar sökningen in också det som är mer eller mindre perifert. Och att såväl kommunikationsministern som den framburna cyberpunkförfattaren endast förekommer i marginalen när det skrivs om internet kan som sagt vara värdefulla resultat i sig.

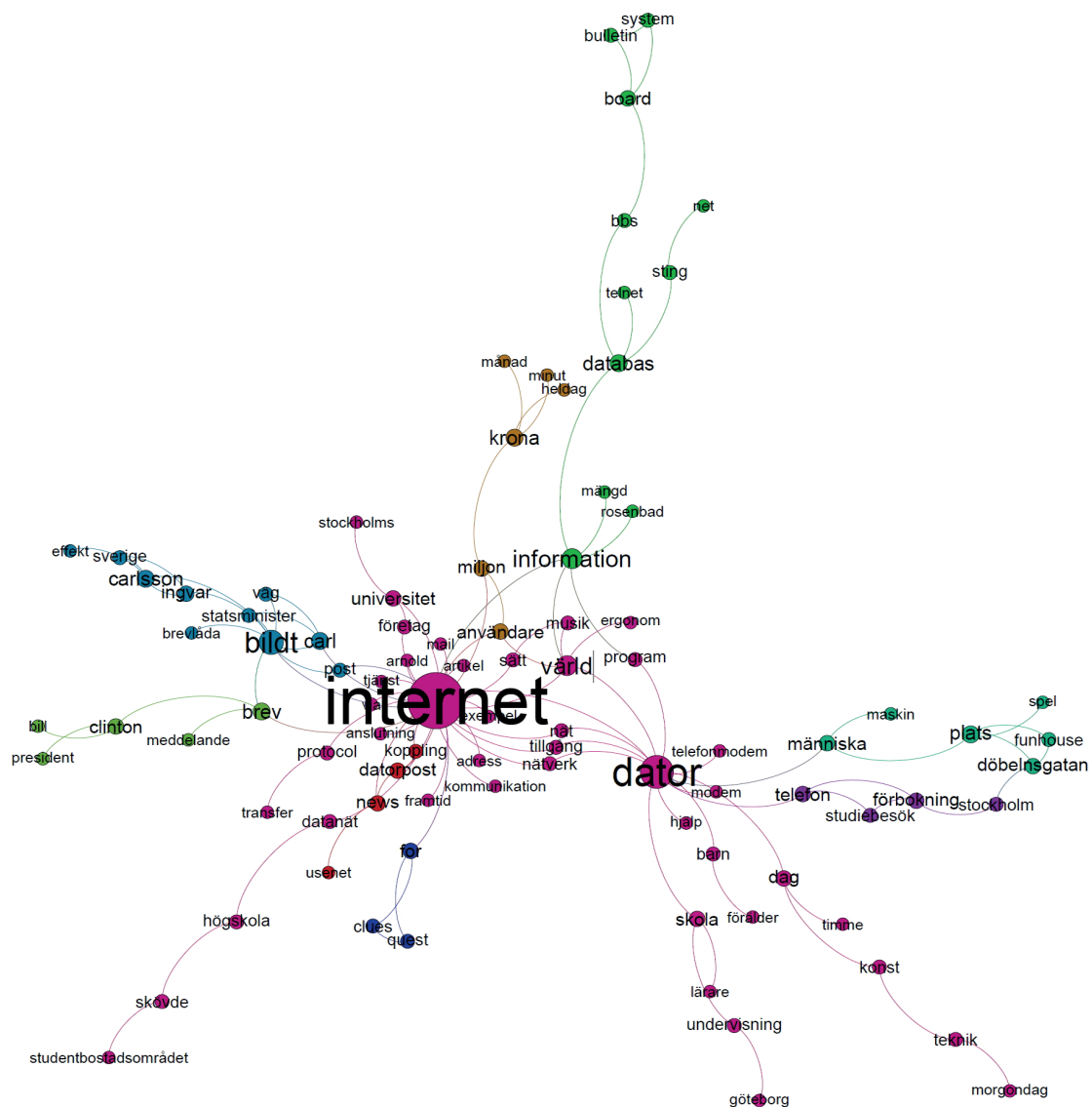
Centrala noder – i den meningen att de knyter an till flera andra noder som tillsammans bildar kluster – är det i regel klokt att framhäva. Noder med många länkar kan göras grafiskt större och olika kluster kan färgläggas för att tydliggöra vilka ord som ofta förekommer tillsammans. Klustren kan påminna om de ‘topics’ som skapas med andra digitala ana-

lysverktyg. De kan också ses som olika typer av gestaltningar (frames) av internetfenomenet – eller som internets diskursiva kontexter (DiMaggio et al. 2013).

De noder som dominerar är i vissa fall samma för båda år: ord som information, dag, dator, miljon, värld, et cetera. Men det finns också mycket som skiljer de båda åren åt. År 1994 har en enskild aktör en central plats: Carl Bildt. I februari detta år skickade han ett e-postmeddelande till president Bill Clinton vilket fick mycket uppmärksamhet. När de borgerliga partierna förlorade valet i september och Bildt efterträddes av Ingvar Carlsson uppstod frågan om den senare skulle få en ny 'elektronisk brevlåda' eller behålla

den som Bildt hade haft. Några få enskilda personer sätter såtillvida ett tydligt avtryck i internetrapporteringen, men de tongivande politikerna är ändå avskilda från andra delar av nätverket.

Noterbart är att majoriteten av de analyserade texterna 1994 behandlar praktiska och tekniska aspekter av internetanvändningen: Vad det är, hur man gör, vad som finns, vad det kallas, vad det kostar. Texterna lanserar och förklarar nya begrepp och företeelser som 'bbs', 'databas' eller 'modem'. Här finns också spår av diskussioner om relationen människa-maskin och datorer i skolundervisningen. År 1994 sprider sig internetanvändningen över världen – men internet är samtidigt en



Figur 1. Internet i *Expressen* och *Göteborgs-Posten* 1994

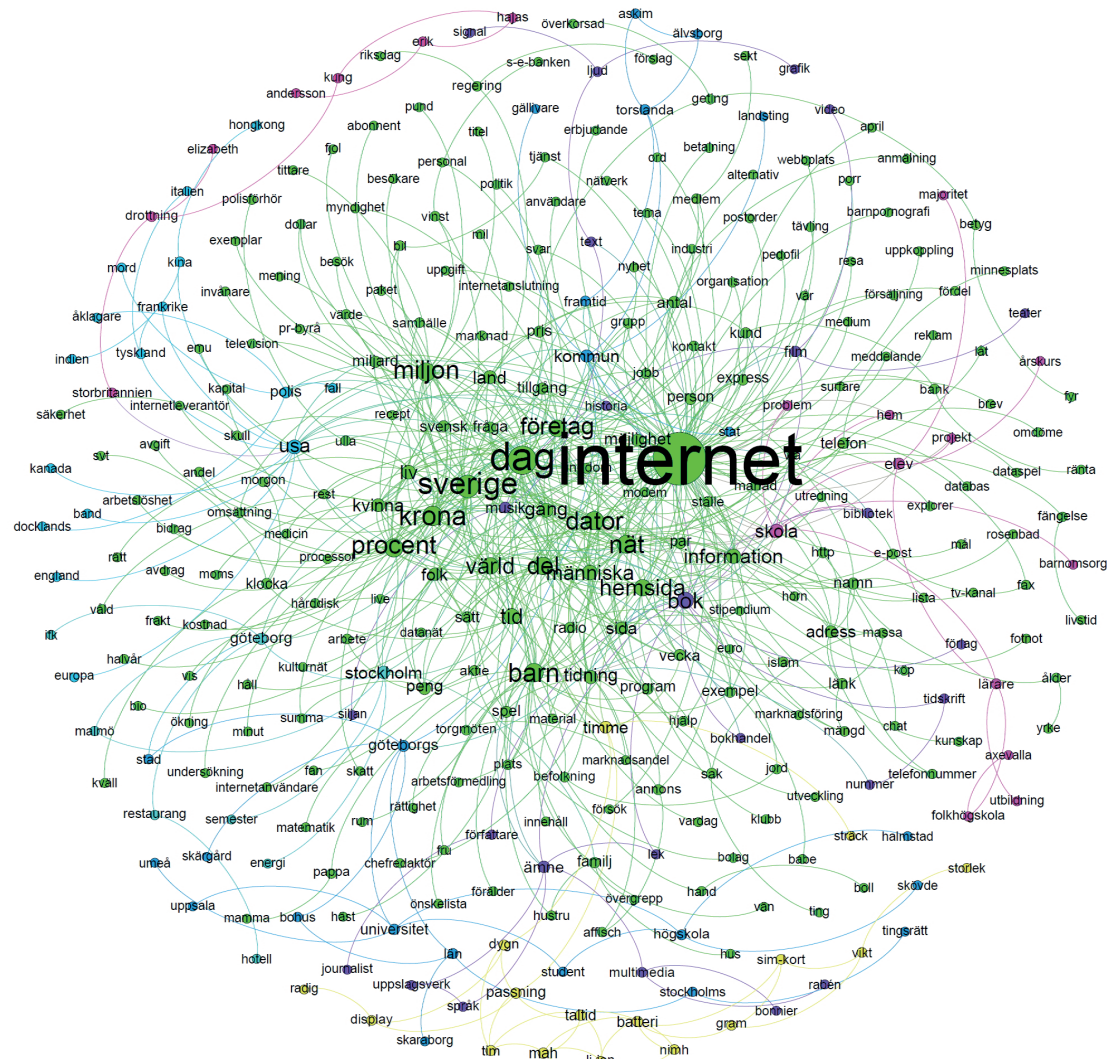
(språklig) värld i sig. Denna digitala värld är emellertid ännu ganska glesbefolkad. Enskilda exempel och verksamheter får illustrera möjligheterna: 'universitet', 'studentbostadsområdet', 'futurekids', spelhallen 'Funhouse'.

I nätverket förekommer flera ord som är synonyma: 'mail', 'post', 'brev', 'datorpost', 'brevlåda'. Om syftet är att studera kommunikationsformen snarare än ordvalen kan dessa enkelt slås ihop till en enda nod. I en analys av hur nya medier språkliggörs är det emellertid intressant att notera begreppsmångfalden. År 1994 har de journalister som skriver inte bestämt sig för en enskild benämning utan rör sig med flera olika termer.

I materialet från 1997 är Sverige en viktig del av rapporteringen. Barn har också en

central plats, liksom begrepp som 'hemsida', 'USA', 'tid' och 'bok'. Inga enskilda aktörer dominerar på samma sätt som Carl Bildt 1994. Istället förekommer en lång rad personer och institutioner. Här finns kultur och förlag, kvinnor och arbete, framtid och möjligheter, men också barnpornografi och brottslighet. Internets värld är större och språkligt mer mångfasetterad 1997.

Denna mångfald kan visserligen förblinda, men det finns flera sätt att bryta ner nätverket i mindre delar, exempelvis genom att studera ett enskilt ord och dess närmast omgivande nätverk. Barn omges exempelvis av 'föräldrar', 'spel', 'möjlighet', 'önskelista', 'tid', 'liv' och 'rättighet' – men också av 'övergrepp' och 'matematik'. Människa länkar på motsvarande sätt



Figur 2. Internet i *Expressen* och *Göteborgs-Posten* 1997

till 'värld', 'möjlighet', 'utveckling', 'kontakt'. Ord som har många länkar är mer centrala än ord med få länkar. 'Företag' väger med detta sätt att räkna 23 gånger mer än 'politik', 'kommun' tio gånger mer än 'riksdag'. Bland statsråden är internet något som berör "kommunikationsminister", "kulturminister", "skatteminister" och "stadsminister", men dessa aktörer länkar endast till några enstaka noder. USA:s inflytande är desto mer påtagligt, med länkar till 17 andra noder. Landets centrala ställning framgår också av att "Europa" inte har någon direkt länk till 'internet' i nätverkets centrum – nätverksanslutningen går via USA.

Avslutning

Med utgångspunkt i de preliminära resultat som ovanstående förstudie presenterar är det viktigt att inte övertolka de samförekomster som identifierats. Frekvenser och länkar är som påtalats resultat av sökfönstrets storlek. Med en annan optik kan bilden förändras – precis som vid all annan vetenskaplig kartläggning. Fördelen gentemot manuell kvantitativ kodning är att den digitala sökningen enkelt kan göras om på nytt om ingångsvärdena visar sig vara problematiska. Själva ordklassningen och samförekomstsökningen går i regel mycket snabbt och tar ofta mindre än en minut. Att sökverktygen är enkla att använda innebär att det finns goda möjligheter att pröva sig fram. Det är inte minst experimenterandet som gör att forskaren lär känna texterna och börjar upptäcka mönstren.

Ett alternativ till den enklare metod som valts för detta test är 'topic modeling'. Ett topic kan påminna om det begrepps nätverk som omger ett enskilt ord i nätverksanalysen ovan. Fördelen med topic modeling är att programvaran kan placera ett och samma ord i flera olika topics. Den enklare samförekomstsökningen förmår ju faktiskt inte skilja mellan, säg, schlagerartisten Nick Borgens låt 'World Wide Web' och storföretaget Atlas Copcos användning av samma webb. Självklart finns det en poäng med att visa att den norske musikern och den globala industrikoncernen samlas kring samma medieteknik. Men eftersom

sammanhanget skiljer sig åt finns också en risk med att klumpa ihop de båda. En fördel med den enklare kartläggningen av samförekomster kan emellertid vara metodens transparens. Inga komplicerade sannolikhetskalkyler ligger bakom resultaten, metoden är enkel att förstå, förklara och utvärdera, och den gör det lätt att växla mellan distansläsning och närläsning. Risken med mer förfinade verktyg är att de blir metodologiska svarta lådor vars utdata är svårutvärderade.

En styrka med digitala metoder är att de på ett enkelt sätt visar vad som är centrum och vad som är periferi. Eftersom kvantifieringen inte kräver några fördefinierade kategorier är sökningarna mer förutsättningslösa jämfört med traditionella kvantitativa metoder. Detta gör att de digitala redskapen lämpar sig väl när man vill skaffa sig översikt – samt inte minst att orientera sig i stora material. När man väl identifierat de stora dragen kan man sedan kombinera distansläsningen med närläsning av traditionell sort. Belackare av digitala humaniora brukar ofta anklaga de forskare som är verksamma i detta fält för att vara antihermeneutiska nypositivister med en övertro på 'Big Data'. Vad de ofta glömmer är att storskalig analys av kulturella dataset som exempelvis äldre pressmaterial ofta just handlar om att kombinera kvantitativ och kvalitativ analys. Härvidlag finns stora möjligheter att ställa helt nya frågor, att arbeta explorativt och laborativt, samt att undersöka och tänka med 'det digitala'. Som den här artikeln visat är inte minst själva metodutvecklingen (kring de forskningsfrågor som riggas) synnerligen produktiv i sig själv (ibland nästintill oberoende av de resultat som åstadkoms). Det senare gör forskningsfältet kring digital humaniora till ett av de mer progressiva inom humanvetenskaperna för närvarande.

Här finns därför, menar vi avslutningsvis, betydande metodologiska impulser som den mer traditionella medieforskningen kunde hämta näring ur. För digitala metoder behöver inte med nödvändighet appliceras på pressmaterial från 1800- eller 1900-talet. Exempelvis finns ett underskott av digitala metoder inom medievetenskapen för mätning av samtida

mediebetenden. Ett metodologiskt problem för den medie- och kommunikationsvetenskapliga forskningen – liksom i ökande grad för de forskare som ägnar sig åt mediehistoriska frågeställningar med utgångspunkt i äldre digitaliserat mediematerial – är emellertid svårigheten att få tag på relevant data. Här handlar det inte med nödvändighet om att programmera fram nya sätt att förstå eller hantera medier som kod (även om det underlättar). Som forskningspraktik innebär digitala metoder framför allt att tänka med och förstå den mjukvara, applikationer och protokoll som idag hanterar nya (och äldre)

medier, exempelvis på exakt vilka sätt som analoga medier digitaliserats och vilka forskningsmöjligheter som då uppstår när det gäller hanteringen av filer och metadata som digitaliseringsprocessen genererar. Alla medier är sedan länge digitala. I takt med att digitaliseringen av kulturarvet fortsätter gäller det även äldre medieformer som dagspress. Att lära sig hur olika slags digitala metoder kan användas för att analysera *medier som data* för alternativa förklaringsmodeller och (ibland) ökad förståelse, framstår därför som en angelägen uppgift – både för den historiska och mer samtidsorienterade medieforskningen.

Referenser

- Arkiv- och biblioteksfilmning (1951). SOU 1951: 36. Stockholm
- Block, Sharon (2006). Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources. *Common-Place*, 6(2). (www.common-place.org/vol-06/no-02/tales)
- Colella, Silvana (2013). 'That Inscrutable Something': Business in the Periodical Press. *Victorian Periodicals Review*, 46(3): 317-342.
- DiMaggio, Paul, Nag, Manish & Blei, David (2013) Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding. *Poetics*, 41(6): 570-606.
- van Eijnatten, Joris; Pieters, Toine & Verheul, Jaap (2014). TS Tools: Using Texcavator to Map Public Discourse, *Tijdschrift voor Tijdschriftstudies*, 35: 59-65.
- Gitelman, Lisa (2006). *Always Already New: Media, History and the Data of Culture*. Cambridge, Mass: MIT Press.
- Gustafsson, Tommy & Arnberg, Klara (2013). *Moralpanik och lågkultur: Genus- och mediehistoriska analyser*. Stockholm: Atlas.
- Hendrickx Iris; Düring, Marten; Zervanou, Kalliopi & van den Bosch, Antal (2013), Searching and Finding Strikes in the New York Times, pp. 25-36 in Mambrini, Francesco; Passarotti, Marco & Sporleder, Caroline (eds.), *Proceedings of The Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3)*, Sofia: Bulgarian Academy of Sciences.
- Holley, Rose (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine* 15(3/4): (www.dlib.org/dlib/march09/holley/03holley.html)
- Huijnen, Pim; Laan, Fons; de Rijke, Maarten & Pieters, Toine (2014) A Digital Humanities Approach to the History of Science. *Social Informatics: Lecture Notes in Computer Science*, 8359: 71-85.
- Jarlbrink, Johan (2015). Historievetenskapens mediehantering. Hyvönen, Mats; Vesterlund, Per & Snickars, Pelle (eds.) *Massmedieproblem*. Lund: Mediehistoriskt arkiv.
- Jockers, Matthew (2013). *Macroanalysis: Digital Methods & Literary History*. Chicago: University of Illinois Press.
- KB – ett nav i kunskapssamhället (2003). SOU 2003: 129. Stockholm,
- KB årsredovisning (2008). Stockholm,
- KB årsredovisning (2009). Stockholm,
- Kushkuley, Sophie (2012). Trend Analysis in Harper's Bazaar, pp. 84-87 in *Workshop on Computational Linguistics for Literature, Montreal, Canada, June 8, 2012*. Stroudsburg, Pa.: Association for Computational Linguistics.
- Latour, Bruno (1987). *Science in Action: How to follow Scientists and Engineers through Society*. Cambridge, Mass: Harvard University Press.
- Marvin, Carolyn (1988). *When Old Technologies Were New: Thinking about Electric Communication in the Late Nineteenth Century*. New York: Oxford University Press.
- McCandless, Michael; Hatcher, Erik & Gospodnetić, Otis. (2010). *Lucene in Action*. Greenwich: Manning.
- Moretti, Franco (2013). *Distant Reading*. London: Verso.
- Newman, David J. & Sharon Block (2006). Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper, *Journal of the American Society for Information Science and Technology*, 57(6): 753-767.
- Nicholson, Bob (2012). Counting Culture; or, How to Read Victorian Newspapers from a Distance. *Journal of Victorian Culture*, 17(2): 238-246.
- Pionke, Albert D. (2014). Excavating Victorian Cuba in the British Periodicals Database. *Victorian Periodicals Review*, 47(3): 369-397.
- Poole, Nick (2010). *The Cost of Digitizing Europe's Cul-*

- tural Heritage. A Report for the Comité des Sages of the European Commission.* (http://nickpoole.org.uk/wp-content/uploads/2011/12/digiti_report.pdf)
- Rogers, Richard (2013). *Digital Methods*. Cambridge, Mass.; MIT Press.
- Salton, Gerard, & McGill, Michael J. (1982). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Schantz, Herbert F. (1982). *The History of OCR, Optical Character Recognition*. Manchester: Recognition Technologies Users Association.
- Svensson, Patrik (2011). From Optical Fiber to Conceptual Cyberinfrastructure. *Digital Humanities Quarterly*, 5(1).
- Svensson, Patrik (2012) Envisioning the Digital Humanities. *Digital Humanities Quarterly*, 6(1).
- Statskontoret (1969). Mikrofilmning av svensk dagspress. Förstudierapport. Riksarkivet – Dataarkiveringskommittén volym E1:3.
- Thorburn, David & Jenkins, Henry (2003). Introduction: Toward an Aesthetics of Transition, pp. 1-16 in Thorburn, David & Jenkins, Henry (eds.), *Rethinking Media Change: The Aesthetics of Transition*. Cambridge, Mass: MIT Press.
- Tidningsfilmningskommittén (1976: 14). *Dagspress på mikrofilm: Filmning – arkivering – läsning*. Stockholm.
- Widholm, Christian (2011). Från massmedium till mikrosurrogat Några reflektioner kring den mikrofilmade dagstidningen som källa. *Scandia*, 77(2): 155-170.
- Yang, Tze-I; Andrew J. Torget & Rada Mihalcea (2011). Topic Modeling of Historical Newspapers, *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 96-104.
- Östling, Robert (2013). Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3: 1-18.

Johan Jarlbrink är docent i mediehistoria och lektor i medie- och kommunikationsvetenskap vid Umeå universitet. Han forskar om 1800-talets tidningspress i alla dess skepnader: tidningen som informationsbärare i arkiv och klippböcker, som avfall och omslagspapper, som forskningsdata i digitala databaser.

Pelle Snickars är professor i medie- och kommunikationsvetenskap med inriktning mot digitala humaniora vid Umeå universitet. Han har bedrivit forskning om såväl äldre som nyare medier. För närvarande leder han två forskningsprojekt: om Spotify och strömmande medier, samt om digitala gestaltningar av samlingar på Tekniska museet.

Cristian Colliander är doktor i biblioteks- och informationsvetenskap och arbetar vid Umeå universitet som forskare på sociologiska institutionen samt som bibliometriker på universitetsbibliotek.