Utgivare: *Institutionen för kulturvetenskaper, Lunds universitet*

DOI: https://doi.org/10.37852/oblu.348.c804

Att skriva idéhistoria
ISBN: 978-91-8104-800-1 (tryck)
ISBN: 978-91-8104-801-8 (e-bok)

Ugglan 21
Lund Studies in the History of Ideas and Sciences
ISSN 1102-4313 (tryck)
ISSN 2004-867X (e-bok)

Vid citering: Mathias Johansson & Emil Stjernholm, ”Digitala metoder och verktyg”, i *Att skriva idéhistoria* (Lund: Lunds universitet, 2026), https://doi.org/10.37852/oblu.348.c804

Information om *Ugglan: Lund Studies in the History of Ideas and Sciences* finns här: https://www.ht.lu.se/serie/41/

Digitala metoder och verktyg

Mathias Johansson & Emil Stjernholm[1]

Hur kan digitala metoder för textanalys användas inom idéhistoriska studier? Denna fråga ställer sig Peter de Bolla, idéhistoriker verksam vid Cambridge, i den nyutkomna antologin *Explorations in the Digital History of Ideas: New Methods and Computational Approaches* (2023). Kultur- och idéhistoriker bygger ofta sina analyser på hermeneutisk tolkning och noggrann analys av texter, där historiskt material placeras i en större kontext.[2] Även om historiker använt sig av digitala metoder i över ett halvt sekel har den idéhistoriska disciplinen varit mer avogt inställd till denna utveckling. Som idéhistorikern Benjamin G. Martin konstaterar i en reflektion om de digitala metodernas löften och utmaningar: ”Idéer verkade lämpa sig dåligt för kvantitativ undersökning.”[3] De Bolla och Martin är dock ense om att det har skett en förändring. De senaste åren märks en snabb tillväxt av idéhistorisk forskning som tar sin utgångspunkt i digitala metoder, såväl i Sverige som internationellt. Särskilt tydligt har detta blivit inom politisk idéhistoria, där digital textanalys har använts för att studera trender i användningen av politiska begrepp i stora korpus (textsamlingar),

1 Arbetet med detta kapitel har stötts av Riksbankens jubileumsfond (P21-0012) och är en del av forskningsprojektet *Moderna Tider 1936* (https://www.modernatider1936.se)

2 Peter De Bolla, *Explorations in the Digital History of Ideas: New Methods and Computational Approaches* (Cambridge: Cambridge University Press, 2023).

3 Benjamin Martin, ”De digitala metodernas löften och utmaningar: ur kulturpolitikens internationella historia”, i *Perspektiv på politisk idéhistoria*, red. Hjalmar Falk, My Klockar Linder och Petter Tistedt (Södertörns högskola, 2023), 188–208.

som exempelvis parlamentariska debatter, digitaliserade dagstidningar, och andra digitaliserade arkiv.[4]

”Digitalhistoria” är ett mångtydigt begrepp som rymmer både idéhistoriska och historiska perspektiv. I detta kapitel ger vi en kort introduktion till begreppet och en kartläggning av viktiga koncept och källor för vidare läsning inom storskalig textbehandling. Exakt var gränsen mellan *historia* och *digitalhistoria* går är svårt att sätta fingret på. Dels för att det är en gräns som hela tiden rör på sig, dels för att *digitalhistoria* är ett heterogent koncept som blandar två olika extremer: å ena sidan finns *traditionella* digitalhistoriker som huvudsakligen förlitar sig på hermeneutik för att studera olika aspekter av digitalisering såsom internets historia. Å andra sidan finns det historiker som använder (huvudsakligen) digitala verktyg och källor för att studera frågeställningar som vore svåra eller rent omöjliga att besvara utan digitala hjälpmedel. Nyckelskillnaden mellan traditionella hermeneutiker och denna senare typ av digitalhistoriker ligger inte i om eller hur digitala verktyg används och inte heller är det ämnet eller perioden som avgör. Snarare kan frågan huruvida något bör räknas som digitalhistoria besvaras genom att ställa frågan: ”hur mycket bidrar det digitala arbetssättet/verktyget?” Så länge bidraget är signifikant är det, i vår mening, digitalhistoria – men denna lösa definition är öppen för debatt. I slutet av kapitlet rekommenderar vi några texter som för mer utförliga reflektioner om definitionen av digitalhistoria.

Eftersom den första sortens digitalhistoria är vinklad mot forskning och den andra är mer metodorienterad går det utmärkt att kombinera dem och studera digitalisering med hjälp av digitala verktyg – något som troligtvis kommer bli ett måste när dagens historia skall skrivas genom internetarkiven.[5] I detta kapitel kommer vi fokusera på den metoddrivna sortens digitalhistoria.

4 Se *Lychnos* temanummer om digital historia, Pelle Snickars, red. (2022).

5 Niels Brügger, *The Archived Web: Doing History in the Digital Age* (Cambridge: MIT Press, 2018).

Tre sorters läsning

Fjärrläsning (Distant reading) är ett koncept som ofta tillskrivs Franco Moretti och hans inflytelserika bok med samma namn,[6] och har sedermera blivit en paraplyterm för alla möjliga sorters textaggregeringsmetoder, inklusive *text-mining*, *Natural Language Processing* (NLP) och diverse AI-system. Till skillnad från traditionell *närläsning (close reading)*, där man noggrant läser sitt källmaterial, är fjärrläsning ett sätt att abstrahera (och ofta aggregera) innehållet i en korpus för att studera större trender som är svårare att upptäcka vid närläsning. Den enklaste varianten av fjärrläsning (både i teori och praktik) är att arbeta med ordfrekvenser. Det handlar helt enkelt om att räkna antalet gånger varje ord förekommer i ens korpus. Resultaten presenterar man i en tabell eller som en visualisering – exempelvis för att visa utvecklingen av förekomsten av orden över tid. Ett populärt verktyg för att visualisera ordfrekvenser i Googles digitaliserade boksamlingar är Google Ngram (se bild 1). Fjärrläsning ger givetvis inte samma detaljerade insikter som att närläsa texterna, men bara en så enkel fråga som vilka ord som är de (o)vanligaste kan ändå bidra med nya insikter kring materialet.

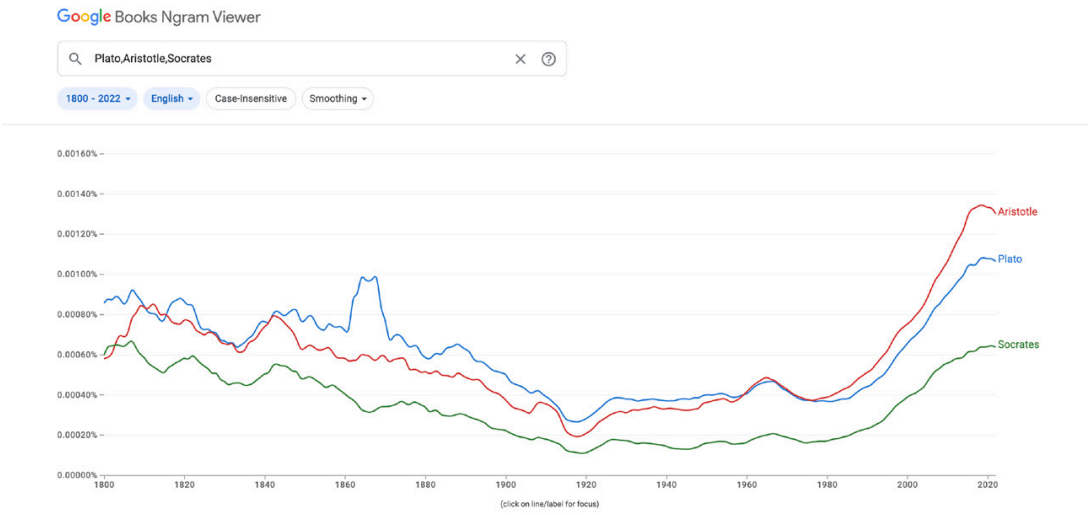

Bild 1. En jämförelse mellan hur frekvent namnen Platon, Aristoteles och Sokrates förekommer i Googles digitaliserade boksamling. En sådan jämförelse väcker exempelvis frågor kring vad som ligger bakom den skarpa ökningen under 1990-talet och huruvida denna hänger samman med dygdetikens renässans.

6 Franco Moretti, *Distant Reading* (London: Verso Books, 2013).

Få historiker skulle nöja sig med att endast studera abstraherade mönster av ett material och därefter skriva en analys. Det finns en vilja att lära känna materialet mer intimt innan det analyseras. Fjärrläsning kan dock vara ett effektivt sätt att snabbt identifiera övergripande mönster och med hjälp av dessa identifiera ett mer specifikt material som är särskilt intressant att närläsa. Närläsningen kan i sin tur forma ytterligare försök till fjärrläsning. Att skifta mellan fjärr- och närläsning på detta sätt kallas ibland för *skalbar (scalable) läsning*.[7]

Fjärrläsning och idéhistoria

Under det senaste decenniet har viktiga och stora källmaterial digitaliserats vilket öppnat för digital textanalys inom idéhistoria. Inflytelserika projekt som Mapping the Republic of Letters, ett projekt som kartlägger tidigmoderna lärda nätverk och deras utveckling, eller The Concept Lab, ett projekt som studerar hur begrepp formas, symboliserar denna utveckling.[8] För att mer konkret exemplifiera hur digital textanalys kan användas på ett mer handfast, småskaligt vis i en nordisk kontext – vilket är av större intresse för studenter än dessa storskaliga, internationella program – kan man vända sig till historikern Matti La Melas arbeten. I en studie undersöker La Mela allemansrättens begreppshistoria i en finsk kontext med utgångspunkt i digitaliserade parlamentsdata. Allemansrätten innebär att alla har rätt att vistas i naturen, och lagen finns bara i en nordisk kontext (i Sverige och i liknande form i exempelvis Finland och Norge).[9] Genom att studera hur

7 Max Odsbjerg Pedersen, Josephine Møller Jensen, Victor Harbo Johnston, Ulrich Thygesen, Alexander Ulrich och Helle Strandgaard Jensen, ”Scalable Reading of Structured Data”, *The Programming Historian*, October 4, 2022, https://programminghistorian.org/en/lessons/scalable-reading-of-structured-data#scalable-reading-a-gateway-for-newcomers-to-digital-methods/.

8 Se ”Mapping the Republic of Letters”, besökt 1 juli 2024, http://republicofletters.stanford.edu/; Centre for Research in the Arts, Social Sciences and Humanities (CRASSH), ”The Concept Lab | Cambridge Centre for Digital Knowledge”, *CRASSH*, besökt 1 juli 2024, https://www.crassh.cam.ac.uk/research/projects-centres/the-concept-lab-cambridge-centre-for-digital-knowledge.

9 Matti La Mela,”Tracing the Emergence of Nordic Allemansrätten Through Digitised Parliamentary Sources”, i *Digital Histories: Emergent Approaches Within the New Digital History,* red. Mats Fridlund, Mila Oiva och Petri Paju (Helsingfors: Helsinki University

och när allemansrätten började diskuteras i Finland ämnar författaren visa att begreppet använts på ett flexibelt vis i relation till en rad politiska diskurser. Med hjälp av temamodellering (*topic modelling*) visar La Mela hur allemansrätten först blev del av en finsk vokabulär under 1950-talet, hänvisande till en allmän tillgång till naturen; under 1970-talet blev allemansrätten ett frekvent använt begrepp i politiska diskussioner – men inte enbart i relation till naturen, utan i en rad olika lagstiftningsdebatter; och slutligen, under 1990-talet, menar La Mela att begreppet främst användes för att försvara en särskilt nordisk livsstil i en tid av fördjupat europeiskt samarbete.[10]

Med hjälp av den här typen av digitala metoder kan forskare lyfta fram mönster i korpus och göra breda påståenden om diskurserna under perioden i fråga, vilket ofta framställs som en förtjänst. Det är dock värt att poängtera att den här sortens fjärrläsningar inte är lika rika i sin kontext, och detaljer fångas inte på samma sätt som i traditionella idéhistoriska studier. Detta är en metodologisk avvägning som man måste göra innan man tar sig ett uppsatsämne med digitalhistorisk vinkel.

Medan vissa digitalhistoriska forskare främst sysslar med fjärrläsningar och makroanalyser är det även många som använder skalbar läsning. Som Benjamin G. Martin påpekat i en metodreflekterande essä finns det faktiskt flera teoretiska beröringspunkter mellan digital textanalys och begreppshistoria i Reinhart Kosellecks anda.[11] Exempelvis lyfter Martin fram synen på mening som historiskt betingad och kontextberoende. Ett exempel på hur detta kan fungera i praktiken finner vi inom ramen för Martins eget projekt International Ideas at UNESCO: Digital Approaches to Global Conceptual History. I en studie diskuterar Martin och medförfattaren Fredrik Mohammadi Norén hur begreppsparet ”natur” och ”kultur” diskuterades inom ramen för tidskriften *The UNESCO Courier*.[12]

Press, 2020), 181–197.

10 La Mela, ”Tracing the Emergence of Nordic Allemansrätten”, 192.

11 Martin, ”De digitala metodernas löften och utmaningar”, 191.

12 Benjamin Martin och Fredrik Mohammadi Norén, ”Nature and Culture in the Age of Environmental Crisis: Digital Analysis of a Global Debate in The UNESCO Courier, 1948–2020”, i Annika Rockenberger, Sofie Gilbert, Juliane Tiemann och Elisa Pierfederici, red., *Digital Humanities in the Nordic and Baltic Countries Publications* (Oslo: Universitetet

Metodologiskt tillämpar de skalbar läsning med utgångspunkt i temamodellering för att sedan titta närmre på specifika sidor som relaterar till ett tema. Å ena sidan kan iakttagelser om individuella teman kontextualiseras och sättas i ett större perspektiv av tillgången till kvantitativa data. Å andra sidan kan viktiga frågor väckas genom att titta närmre på individuella artiklar: i vilka kontexter förekom begreppsparet? Hur förändrades dessa kontexter över tid? Och vilka drev på debatten? Växlandet mellan mikro- och makroanalyser beskrivs ofta som produktivt men krävande. Det förutsätter nämligen en nära förtrogenhet med både den specifika domänen som med digitala metoder och verktyg, något som kan vara särskilt utmanande att ta sig an inom ramen för en uppsatskurs.

Att välja digitala verktyg

Utbudet av verktyg för digitala metoder är stort och rörligt; nya verktyg släpps, gamla verktyg försvinner. De flesta uppsatser skrivs under en tillräckligt kort intervall för att risken att just ditt verktyg kommer sluta fungera är minimal. I längre forskningsprojekt är risken betydligt större, men fortfarande relativt liten. Det finns vissa tumregler för att sänka riskerna associerade med verktygens fortlevnad: 1) Prioritera etablerade och aktivt utvecklade program; 2) Öppen källkod är bättre än stängd, skulle utvecklarna av ett sådant projekt sluta finns alltid chansen att någon annan (kanske du?) fortsätter; 3) Undvik mjukvaror som inte kan exportera data till öppna format – denna export möjliggör nämligen att man enkelt kan byta mjukvara vid behov. I detta kapitel har vi prioriterat verktyg som inte kräver någon licens. Transkribus kräver ingen licens men har inte öppen källkod, dock får man 100 sidor/månad gratis, därefter kostar det ~2kr per sida. Studenter kan be om extra gratissidor för till exempel masteruppsatser.

Text som data

När man arbetar med digital text är det bara en tidsfråga innan man stöter på problemet att ens text inte kan utläsas ordentligt i något program, till

i Oslo, 2023), 274–286.

exempel att *å, ä,* och *ö* inte visas. Problemet är troligtvis att programmen och filerna inte använder sig av samma teckenkodning. För människor är texttecken skrivna, projicerade, utskrivna, inristade, renderade, med mera i rad på en yta som vi kan utläsa som ord som förhoppningsvis förmedlar någon betydelse. Relevant ”tecken” i detta sammanhang är relativt till läsaren. För svenska är det huvudsakligen det svenska alfabetet medan för japanska kan det vara en kombination av Hiragana, Katakana, Kanji och Romaji. För datorer är text bytes (8 bitar (ettor och nollor)) som översätts till tecken (bokstäver, siffror, skiljetecken, med mera). Processen att konvertera mellan binärkod och läsbara tecken kallas *encoding* när tecken översätts till bitar och *decoding* när bitarna översätts till tecken. Det finns en lång rad olika tabeller för denna översättning för olika språk och ändamål, men de två viktigaste att känna till är ASCII (American Standard Code for Information Interchange), som är en av de äldsta med brett användande, och UTF-8 som är det mest använda formatet idag (1 juli 2024 uppskattades det att 98,3 % av webben använde UTF-8).[13] Det är också värt att notera att ordningen på bokstäver inte alltid är som man kan förvänta sig, i UTF-8 följer a-z alfabetet, medan å-ä-ö ligger utspridda i senare delar i ordningen ä-å-ö. Besluten bakom vilka tecken som läggs, vad de ska kallas och deras ordning hanteras av non-profit organisationen Unicode Consortium.

Transkribering

Digitala metoder är bara produktiva när digitala data finns tillhanda. Därför blir det första steget i många digitalhistoriska projekt att transkribera källtexter eller att samla in redan digitaliserade texter. Att transkribera för hand är tidskrävande och således inte att rekommendera för kortare uppsatsarbeten, men det är ibland oundvikligt inom ramen för mer omfattande projekt. Det finns två huvudsakliga sorters tekniker för att digitalisera texter: OCR och HTR. OCR (Optical Character Recognition) är utvecklad för att i första hand känna igen maskinskrivna tecken men fungerar också på modernare, konsekventa, handstilar. OCR-maskiner och

13 W3Techs, ”Historical trends in the usage of character encodings for websites”, *W3Techs*, besökt Juli 2, 2024, https://w3techs.com/technologies/history_overview/character_encoding/ms/y.

mjukvaror har varit i utveckling sedan 1950-talet. Tidiga system kunde bara känna igen ett fåtal typsnitt och endast om tecknen var rakt uppradade. I god ingenjörstradition bygger många av dessa lösningar på ingenjörernas snillrika förmåga att upptäcka mönster. Stora framsteg inom ämnet gjordes genom introduktionen av maskinlärning.[14] Idag behärskar OCR nästan alla typsnitt, finns som öppet tillgänglig mjukvara,[15] och finns som standardfunktion i både skanners och telefoner. I ett nötskal fungerar OCR genom att först identifiera var på sidan olika tecken är – sedan tolkar den var tecken för sig i ordning. Blanksteg och liknande tecken känns igen genom avståndet mellan tecken – därigenom återskapas orden.

HTR (Handwritten Text Recognition) har inte lika långa anor och är mer komplicerat ur ett tekniskt perspektiv. Eftersom HTR behandlar kursiv stil använder den ett extra steg där individuella ord segmenteras innan de individuella bokstäverna kan kännas igen.[16]

Verktyg 1: Transkribus

Transkribus är ett verktyg som kan utföra både HTR och OCR samt publicera sin transkriberade korpus.[17] Det utvecklades inom ett EU Horizon 2020 program och styrs numera genom ett kooperativ av universitet och forskningsinstitutioner. Verktyget bygger på djupinlärningsmodeller (en sorts AI) för transkriptionerna. Eftersom allt arbete utförs på deras servrar är allt man behöver ett gratis-konto och en webbläsare för att kunna transkribera dokument.[18] När man skapar ett konto skapas automatiskt ett exempelprojekt

14 Shunji Mori, Ching Y. Suen och Kazuhiko Yamamoto, ”Historical review of OCR research and development”, *Proceedings of the IEEE*, 80.7 (1992): 1029–1058, https://doi.org/0.1109/5.156468.

15 Tesseract OCR, ”tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (Main Repository)”, *GitHub*, besökt 1 december 2024, https://github.com/tesseract-ocr/tesseract.

16 I skrivande stund har en ny sorts metod börjat bli möjlig, nämligen LLMs (Large Language Models) kombinerade med ViT (Visual Transformers), men även om deras nuvarande prestanda är lovande är de fortfarande för opolerade för vi ska välja att täcka dessa i detta sammanhang.

17 Transkribus, ”Unlock the past with Transkribus”, *Transkribus*, besökt 3 juli 2024, https://www.transkribus.org/.

18 Att all behandling sker på deras servrar innebär att man borde undvika att transkri-

med en färdigbehandlad text. Transkribus och liknande HTR-transkriberingsprocesser är uppdelade i tre steg: 1) lista ut var på sidan texter finns, 2) lista ut hur textraderna går i dessa områden, och 3) transkribera textraderna.

Varje steg går att automatisera, men man kan också göra allt manuellt – eller så låter man maskinen göra grovjobbet och därefter korrigerar man resultatet. Även om algoritmerna och modellerna är väldigt bra är de inte felfria och det är alltid att rekommendera att kontrollera hur de fungerar på just ditt material innan allt transkriberas. Steg 1 och 2 går att få fram samtidigt genom en så kallad ”layout analysis” – prestandan på detta steg är ofta väldigt hög (särskilt på västerländska texter), vilket är ganska enkelt att förstå eftersom det är en relativt grov process. Transkribus brukar ha problem med väldigt tätt skrivna rader (till exempel när någon skrivit in en notis mellan raderna) eller när bläcket från baksidan lyser igenom. Samtidigt är verktygen ofta bra på att känna igen texter skrivna åt olika håll, såsom marginalia skrivet lodrätt.

Det tredje steget är det mest problematiska. Och för att förstå varför det är problematiskt måste vi tänka på hur datorer ser text – alltså inte som former med en mening, utan mer som kategorier (sekvenser, bitar) som översätts till tecken genom en tabell. När man tränar en modell att känna igen text lär den sig att kartlägga från olika former till olika kategorier som sedan tolkas om till tecken. Modellen gör alltså inte en jämförelse mellan formerna på sidan och hur tecknet ser ut. Eftersom den är baserad på formen, inte på kontexten av tecken, kan den ibland blanda ihop tecken som ser lika ut (”e” och ”c” eller ”o”, ”O”, och ”0” eller ”-”, ”_”, ”–”) och de är inte alltid konsekventa. Kartläggningen är också begränsad till de tecken som finns i träningsmaterialet – 100 % engelska modeller är alltså inkapabla att känna igen ”ääöÅÄÖ”.

Transkribus erbjuder ett smörgåsbord av allmänt tillgängliga modeller på många olika språk – med en klar övervikt mot europeiska språk. Modellerna blir fler hela tiden eftersom olika forskningsprojekt blir klara och publicerar sina modeller. Ta till exempel The Swedish Lion I som är tränad på 15,6

bera känsliga dokument med Transkribus.

miljoner ord från ett urval ur riksarkivets texter.[19] Poängen med en sådan modell är inte att den direkt skall kunna transkribera alla svenska texter felfritt – utan att den ger oss en bra och solid startpunkt när vi tränar egna modeller. Denna modell har redan lärt sig grunderna i svenskt skrivande, och behöver därför bara se ett (relativt) fåtal exempel av en ny svensk handstil för att kunna transkribera den väl. Men perfekt lär det inte bli, även The Swedish Lion I har en *Character Error Rate* (CER) på 4 % – alltså, den transkriberar fel på cirka 1 tecken av 25. Det är imponerande bra, men kräver lite finputsning.

När man inte behöver digitalisera

Hittills har vi mest pratat om hur man kan extrahera texter från existerande material. Men det finns så klart tillfällen när andra har gjort digitaliseringen åt oss och när vi arbetar med nyare material kan det rentav handla om data som är *born digital*. Alltså, textens originalform är ett digitalt format, såsom exempelvis hemsidor, databaser eller digitala manuskript.

När materialet redan är digitaliserat är den svåraste och mest arbetsintensiva delen av arbetet gjort. Dock är det alltid en bra idé att först inspektera materialet för att observera vilka sorters fel eller problem som kan vara kvar – och kanske städa bort dessa. När vi använder material som är *born digital* kan vi också behöva förhålla oss till etiska frågor angående datainsamling, datahantering och frågor kring huruvida vi kan dela med oss av den insamlade datan.

I dagens ekosystem finns det många ställen att hämta färdig data. Det finns hemsidor med sökgränssnitt, API:er (Application Programming Interface), men också plattformar där man kan exportera data manuellt, och i bästa fall i form av ett mindre urval data. Men det finns också fall där all data bara dumpas någonstans och det är upp till användaren att göra ordning i den. Två svenska exempel på det tidigare, sökgränssnitt och API, är DigitaltMuseum och KB:s tidningsdatabas.[20] Ett (svenskt) exempel på

19 Swedish National Archives, ”Swedish Handwriting Model”, Transkribus, senast uppdaterad 14 april 2022, https://app.transkribus.org/models/public/text/55158.

20 DigitaltMuseum, besökt 28 november 2024, https://digitaltmuseum.se/. Kungliga biblioteket, ”API:er och öppna data”, *Kungliga biblioteket*, besökt 18 augusti 2025, https://

det senare är sökgränssnittet riksdagsdebatter.se,[21] vilket erbjuder ett tillgängligt grafiskt gränssnitt för att söka igenom de svenska riksdagsdebatterna och göra exporter av relevanta delar. Dessutom är alla transkriptionerna bakom riksdagsdebatter.se öppet delade för vem som helst att behandla och studera.[22] Även KB:s samling med digitaliserade offentliga utredningar utgör ett spännande material.[23]

Verktyg 2: RegEx – reguljära uttryck

Få verktyg är användbara i så många olika sammanhang att man, efter att ha lärt sig använda det, undrar hur man överlevt utan det. RegEx är ett sådant verktyg. RegEx (Regular Expression, reguljära uttryck) är ett verktyg för *pattern matching* i textfiler och är ett sätt att känna igen och extrahera specifika mönster eller teckenföljder i text. Det är ett särskilt användbart verktyg i tidiga skeden av projekt där stora mängder text behandlas. Det är såpass flexibelt, kraftfullt och snabbt att man kan bygga hela uppsatsens metod på den: från att utforska och städa korpusen, till att utföra enkla analyser såsom ordfrekvenser, konkordanser och KWIC (Key Words in Context).[24]

Till skillnad från Transkribus är det inte alltid så användarvänligt, det finns ingen säkerhetskontroll som ser till att ditt mönster stämmer överens med din avsikt. Samtidigt tillåter RegEx skrivandet av väldigt komplexa och kondenserade uttryck som kan göra det väldigt svårt för människor att tolka. Men, det räcker ofta med enkla uttryck för att man ska kunna göra underverk. RegEx kommer i olika dialekter, som skiljer sig väldigt lite åt, och är ofta inbyggd i text/code-editors (såsom Sublime Text, VS Code) och

---

data.kb.se/api.

21 *Riksdagsdebatter.se*, besökt 13 april 2025, https://riksdagsdebatter.se/public/index.html#/.

22 Swerik Project, ”swerik-project/the-swedish-parliament-corpus”, *GitHub*, besökt 18 augusti 2025, https://github.com/swerik-project/the-swedish-parliament-corpus.

23 ”SOU – Statens offentliga utredningar (digitaliserad samling)”, *Kungliga biblioteket*, besökt 19 augusti 2025, https://regina.kb.se/sou.

24 För en storskalig öppen tillämpning se Språkbanken Text, ”Korp”, *Språkbanken Text*, besökt 18 augusti 2025, https://spraakbanken.gu.se/korp.

programmet grep (General RegEx) och sed (stream editor) brukar komma förinstallerade i MacOS och de flesta Linuxdistributioner.

När vi skriver RegEx-mönster kan vi använda alla de vanliga tecken vi hittar på tangentbordet: bokstäver, siffror, specialtecken, mellanslag, tab, retur och så vidare, är alla till vårt förfogande. Vissa tecken har en specialstatus som kräver att de leds av en backslash för att söka för just det tecknet. Men RegExs kraft kommer från möjligheten att specificera både bredare och snävare sökparametrar. En speciellt intressant markör är \b som används för att söka för ”word boundaries”. Det matchar alltså inte något specifikt tecken, utan gränsen till ord. Med hjälp av \b kan man urskilja *ordet* ”i” från ord som innehåller bokstaven i – en mycket viktig distinktion. För att se en mer detaljerad genomgång av hur RegEx kan tillämpas för historiebruk, rekommenderas boken *Doing Digital History*.[25] Det finns även flertalet webbsidor där man kan öva på och testa RegEx uttryck.[26]

Verktyg 3: Voyant

Det finns många olika verktyg med lättillgängliga grafiska gränssnitt för att analysera större mängder text, som dessutom är fria att använda. Ett sådant program är Voyant – som kommer i två versioner – en som är tillgänglig webbläsaren från Voyants egna servrar[27] och en som man kan ladda ner för att köra sin egen Voyant-server.[28] Om man arbetar med ett potentiellt känsligt material är det senare att föredra, även om det kräver mer förarbete.[29] Voyant innehåller en bred samling mindre verktyg för att utforska korpus. Den enklaste funktionen är troligtvis att räkna ordfrekvenser, hur många

25 Blaney, Jonathan, Jane Winters, Sarah Milligan och Martin Steer, *Doing Digital History: A Beginner's Guide to Working with Text as Data* (Manchester: Manchester University Press, 2021).

26 ”Firas Dib, ”Regex101 – Online Regex Editor and Debugger”, *Regex101*, besökt 18 augusti 2025, https://regex101.com/.

27 Stéfan Sinclair och Geoffrey Rockwell, ”Voyant Tools”, *Voyant Tools*, besökt 4 januari 2025, https://voyant-tools.org/.

28 Voyant Tools, ”Voyanttools/VoyantServer”, *GitHub*, besökt 4 januari 2025, https://github.com/voyanttools/VoyantServer.

29 Voyant Tools, ”Voyant Tools”, *GitHub*, besökt 18 augusti 2025, https://github.com/voyanttools.

gånger varje ord använts i korpusen. Frekvensen uttrycks ibland som relativa frekvenser vilket kan vara speciellt användbart när man jämför ordanvändningen i olika stora korpus. Men vissa ord är så vanliga att man knappt behöver räkna dem, såsom ”i” och ”och”. I *textmining* kallas dessa ord ”stoppord” och brukar vanligtvis filtreras bort ur korpusen innan analysen. I Voyant finns en generell stoppordslista för svenska. Dock är det viktigt att komma ihåg att vad som anses vara ett stoppord beror på sammanhanget så vi rekommenderar starkt att man inspekterar och justerar den vid behov.

LDA – Topic Modelling

Latent Dirichlet Allocation (LDA)[30] är en väldigt vanlig algoritm för temamodellering (Topic Modelling); låt datorn lista ut vilka teman som finns i korpusen. Det finns många öppet tillgängliga implementeringar av LDA, här använder vi den i Voyant. Temamodellering ämnar sig speciellt för *utforskandet* av stora korpus. Det går till så att man tränar en (tema) modell att känna igen mönster mellan texterna, och räkna ut de parvisa sambanden mellan varje ord och vardera av de *x* teman man ber den hitta. Det vill säga, frågan är inte *om* ett ord tillhör ett ämne – det är *hur starkt* det är kopplat till det – en hårfin men viktig distinktion. Första steget i analysen är ofta att gå igenom de *n starkast* associerade orden till varje tema och namnge dem.[31] Det är i sådana sammanhang en bra stoppordslista kan göra underverk. Utan den skulle var och en av dessa listor av topp-ord vara full av dessa mycket vanliga ord.

---

30 David Blei, Andrew Ng och Michael Jordan, ”Latent Dirichlet Allocation”, *Advances in Neural Information Processing Systems* 14 (2001): 993–1022.

31 Jo Guldi, ”Parliament’s Debates about Infrastructure: An Exercise in Using Dynamic Topic Models to Synthesize Historical Change”, *Technology and Culture* 60, nr 1 (2019): 1–33, https://doi.org/10.1353/tech.2019.0000; Pelle Snickars,”Modeling Media History: On Topic Models of Swedish Media Politics 1945–1989”, *Media History* 28, nr 3 (2022): 403–424, https://doi.org/10.1080/13688804.2022.2079484; Johan Jarlbrink, Fredrik Norén och Robin Saberi, ”Contextual Modelling of ’Propaganda,’ ’Information’ and ’Upplysning’ in Swedish Parliamentary Speeches, 1920–2019”, i *Digital Parliamentary Data in Action*, red. Matti La Mela och Fredrik Norén, 2022

När modellen är tränad på korpusen kan vi använda den för att räkna ut vilka teman texter innehåller. Kartläggningen kan i sig vara tillräcklig för att analysera korpusen men man kan också använda den för att identifiera vilka texter som är relevanta för närläsning. För att uppnå temamodelleringens fulla potential behöver man en stor korpus. Det finns ingen fastslagen miniminivå men erfarenhet visar att hundra till tusentals dokument (med ett större antal ord) kan behövas för att utvinna nya rön.

Temamodellering som metod har tre anmärkningsvärda svagheter: 1) stoppordslistan kan ha ett stort inflytande över vilka teman som hittas, så välj den med omsorg, 2) träningsprocessen innehåller ett slumpmoment och två körningar med exakt samma data ger lika, men inte identiska, teman, och 3) användaren bestämmer själv hur många teman modellen ska hitta (och den kommer alltid hitta exakt så många teman). Som exempel har vi med Voyant tränat två modeller på ett utkast av detta kapitel, en modell med fem teman och en med 100, resultaten visas i Bild 2. Med fem teman får man en någorlunda rimlig beskrivning av kapitlets innehåll. Med 100 teman är det många teman som inte ens har fem starkt associerade ord (de tomma rutorna) och ett flertal är associerade med osammanhängande ord. Men vi frågade efter 100 teman, så vi fick 100 teman.

Bild 2. En jämförelse mellan temamodeller med 5 teman (vänster) och 100 teman (höger). Tränade på ett utkast av denna text.

Avslutning

Med ökad tillgänglighet till storskaliga digitala källmaterial och digitala metoder för textanalys uppenbarar sig många möjligheter och vi hoppas att detta kapitel har kunnat illustrera några potentiella analytiska vinster med dessa metoder. För studenter som ska skriva uppsats, och därmed arbetar under en snäv tidsram, är det viktigt att påpeka att digital textanalys även kan vara utmanande och tidsödande. Exempelvis kan det ta lång tid att kurera och bearbeta källorna (även om dessa finns tillgängliga i maskinläsningsbar form) och det kan ta tid att sätta sig in i de digitala verktyg som finns tillgängliga för digitalhistoriker. Samtidigt kan arbetet med digitala metoder och verktyg vara en fin merit när man ger sig ut i arbetslivet. För uppsatsstudenten är det således viktigt att reflektera över såväl förtjänster som utmaningar tidigt i uppsatsarbetet.

Slutligen bör man påpeka att fjärrläsningar även kan fungera som ett sätt att väcka frågor om långsiktiga trender, som sedan kan undersökas närmare med traditionella närläsningsmetoder. Otaliga är de konferenser, workshops och symposier som inleds med ett diagram över hur vissa termer (som exempelvis ”informationssamhället”, ”polarisering” eller ”falska nyheter”) ökat i popularitet under vissa specifika tidsperioder. Sådana enkla digitalhistoriska övningar kan onekligen vara inspirerande och enkla att genomföra utifrån lättillgängliga digitala samlingar. För att ta ett konkret exempel: i antologin *Efterkrigstidens samhällskontakter* undersökte författarna det ofta traderade påståendet att begreppet propaganda blev ett nedsvärtat begrepp med negativa associationer i kölvattnet av andra världskriget och att det gradvis ersattes med en annorlunda vokabulär (public relations, information, kommunikation, etcetera) i en svensk kontext. En undersökning som vi genomförde av cirka 4 500 SOU:er (Statens offentliga utredningar) visade visserligen på ett sådant semantiskt skifte – däremot kunde flera kvalitativa bidrag i antologin, med utgångspunkt i historisk arkivforskning, visa att begreppet fortsatte att användas i en mer neutral form långt in i efterkrigstiden och att de kommunikativa praktiker som associerades med begreppen inte ändrades alls lika häftigt som graferna antydde.[32] Som Martin poängterar, med hänvisning till litteraturvetaren Ted

32 Fredrik Norén och Emil Stjernholm, red., *Efterkrigstidens samhällskontakter* (Lund:

Underwood, behöver digitala metoder alltså ”inte ses som en ersättning för humanistiska metoder, utan snarare som ett komplement.”[33]

Föreslagen vidare läsning

- Salmi *What is Digital History?* (2020)
- Orrje *Vad är digital historia?* (2021)
- Jarlbrink och Norén, *Digitala metoder i humaniora och samhällsvetenskap* (2021)
- Fridlund, *Digital History 1.5: A Middle Way between Normal and Paradigmic Digital Historic Research* (2020)
- Brügger, *The Archived Web: Doing History in the Digital Age* (2018)
- Blaney, *Doing Digital History: A Beginner's Guide to Working with Text as Data* (2021)
- Zebtgraf, *What Every Programmer Absolutely, Positively Needs to Know About Encodings and Character Sets to Work With Text* (2015)
- *Programming Historian*

Referenser

Blaney, Jonathan, Jane Winters, Sarah Milligan och Martin Steer. *Doing Digital History: A Beginner's Guide to Working with Text as Data*. Manchester: Manchester University Press, 2021.

Blei, David Andrew Ng och Michael Jordan. ”Latent Dirichlet Allocation.” *Advances in Neural Information Processing Systems* 14 (2001): 993–1022.

Brügger, Niels. *The Archived Web: Doing History in the Digital Age*. Cambridge: MIT Press, 2018.

Centre for Research in the Arts, Social Sciences and Humanities (CRASSH). ”The Concept Lab | Cambridge Centre for Digital Knowledge.” *CRASSH*. Besökt 1 juli 2024. https://www.crassh.cam.ac.uk/research/projects-centres/the-concept-lab-cambridge-centre-for-digital-knowledge.

De Bolla, Peter. *Explorations in the Digital History of Ideas: New Methods and Computational Approaches*. Cambridge: Cambridge University Press, 2023.

Dib, Firas. ”Regex101 – Online Regex Editor and Debugger.” *Regex101*. Besökt 18 augusti 2025. https://regex101.com/.

Mediehistoriskt arkiv, 2019).

33 Martin, ”De digitala metodernas löften och utmaningar”, 207; Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: University of Chicago Press, 2019).

DigitaltMuseum. ”DigitaltMuseum.” *DigitaltMuseum*. Besökt 28 november 2024. https://digitaltmuseum.se/.

Fridlund, Mats. ”Digital History 1.5: A Middle Way Between Normal and Paradigmic Digital Historic Research.” I *Digital Histories: Emergent Approaches Within the New Digital History*, red. Mats Fridlund, Mila Oiva och Patri Paju, 69–88. Helsingfors: Helsinki University Press, 2020. https://doi.org/10.33134/HUP-5.

Guldi, Jo. ”Parliament’s Debates about Infrastructure: An Exercise in Using Dynamic Topic Models to Synthesize Historical Change”, *Technology and Culture* 60.1 (2019): 1–33, https://muse.jhu.edu/pub/1/article/719944/summary

Jarlbrink, Johan och Fredrik Norén. *Digitala metoder i humaniora och samhällsvetenskap*. Lund: Studentlitteratur, 2021.

Jarlbrink, Johan, Fredrik Norén och Robin Saberi. ”Contextual Modelling of ’Propaganda,’ ’Information’ and ’Upplysning’ in Swedish Parliamentary Speeches, 1920–2019”, i *Digital Parliamentary Data in Action*, red. Matti La Mela och Fredrik Norén, 2022
https://www.diva-portal.org/smash/get/diva2:1655635/FULLTEXT01.pdf

Kungliga biblioteket. ”API:er och öppna data.” *Kungliga biblioteket*. Besökt 18 augusti 2025. https://data.kb.se/api.

Kungliga biblioteket. ”SOU – Statens offentliga utredningar (digitaliserad samling).” *Kungliga biblioteket*. Besökt 19 augusti 2025. https://regina.kb.se/sou.

La Mela, Matti. ”Tracing the Emergence of Nordic Allemansrätten Through Digitised Parliamentary Sources”, i *Digital Histories: Emergent Approaches Within the New Digital History*, red. Mats Fridlund, Mila Oiva och Petri Paju, , 181–197. Helsingfors: Helsinki University Press, 2020.

Mapping the Republic of Letters. ”Mapping the Republic of Letters.” *Mapping the Republic of Letters*. Besökt 1 juli 2024. http://republicofletters.stanford.edu/.

Martin, Benjamin. ”De digitala metodernas löften och utmaningar: ur kulturpolitikens internationella historia.” I *Perspektiv på politisk idéhistoria*, red. Hjalmar Falk, My Klockar Linder och Petter Tistedt, 188–208. Södertörns högskola, 2023.

Martin, Benjamin och Fredrik Mohammadi Norén. ”Nature and Culture in the Age of Environmental Crisis: Digital Analysis of a Global Debate in The UNESCO Courier, 1948–2020.” I Annika Rockenberger, Sofie Gilbert, Juliane Tiemann och Elisa Pierfederici, red., 274–286. *Digital Humanities in the Nordic and Baltic Countries Publications*. Oslo: Universitetet i Oslo, 2023. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1809055.

Moretti, Franco. *Distant Reading*. London: Verso Books, 2013.

Mori, Shunji, Ching Y. Suen och Kazuhiko Yamamoto. ”Historical review of OCR research and development.” *Proceedings of the IEEE* 80, nr 7 (1992): 1029–1058. https://doi.org/10.1109/5.156468.

Norén, Fredrik och Emil Stjernholm, red. *Efterkrigstidens samhällskontakter*. Lund: Mediehistoriskt arkiv, 2019.

Odsbjerg Pedersen, Max, Josephine Møller Jensen, Victor Harbo Johnston, Ulrich Thygesen, Alexander Ulrich och Helle Strandgaard Jensen. ”Scalable Reading of Structured Data.” *The Programming Historian*, October 4, 2022, https://programminghistorian.org/en/lessons/scalable-reading-of-structured-data#scalable-reading-a-gateway-for-newcomers-to-digital-methods/.

Orrje, Jacob. ”Vad är digital historia?”, *Historisk Tidskrift* 141, nr 4 (2021): 723–732.

Riksdagsdebatter.se. ”Riksdagsdebatter.” *Riksdagsdebatter.se*. Besökt 13 april 2025. https://riksdagsdebatter.se/public/index.html#/.

Salmi, Hannu. *What is Digital History?* Cambridge: Polity Press, 2021.

Sinclair, Stéfan och Geoffrey Rockwell. ”Voyant Tools.” *Voyant Tools*. Besökt 4 januari 2025. https://voyant-tools.org/.

Snickars, Pelle. ”Modeling Media History: On Topic Models of Swedish Media Politics 1945–1989.” *Media History* 28, nr 3 (2022), 403–424. https://doi.org/10.1080/13688804.2022.2079484.

Språkbanken Text. ”Korp.” Språkbanken Text. Besökt 18 augusti 2025. https://spraakbanken.gu.se/korp.

Swedish National Archives. ”Swedish Handwriting Model.” *Transkribus*. Senast uppdaterad 14 april 2022. Besökt 24 september 2025. https://app.transkribus.org/models/public/text/55158.

Swerik Project. ”swerik-project/the-swedish-parliament-corpus.” *GitHub*, besökt 18 augusti 2025. https://github.com/swerik-project/the-swedish-parliament-corpus.

Tesseract OCR. ”tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (Main Repository).” GitHub. Besökt 1 december 2024. https://github.com/tesseract-ocr/tesseract.

Transkribus. ”Unlock the Past with Transkribus.” *Transkribus*. Besökt 3 juli 2024. https://www.transkribus.org/.

Underwood, Ted, *Distant Horizons: Digital Evidence and Literary Change.* Chicago: University of Chicago Press, 2019.

Voyant Tools. ”Voyant Tools.” *GitHub*. Besökt 18 augusti 2025. https://github.com/voyanttools.

Voyant Tools. ”Voyanttools/VoyantServer.” *GitHub*. Besökt 4 januari 2025. https://github.com/voyanttools/VoyantServer.

W3Techs. ”Historical Trends in the Usage of Character Encodings for Websites.” *W3Techs*. Besökt 2 juli 2024. https://w3techs.com/technologies/history_overview/character_encoding/ms/y.

Zentgra, David C. ”What Every Programmer Absolutely, Positively Needs to Know About Encodings and Character Sets to Work With Text.” Kunstube (blog), april 27, 2015, besökt juli 1, 2024. https://kunststube.net/encoding/.