100 miljoner ord

Reflektioner kring forskningsarbete med storskaliga textuella dataset som historisk empiri

PELLE SNICKARS *Lunds universitet*

I forskningsprojektet "Välfärdsstaten analyserad" (1945–1989) arbetar vi med olika typer av algoritmisk textanalys av storskalig empiri från politikens sfär (digitaliserat riksdagstryck och statliga offentliga utredningar under perioden), dagspress och skönlitteratur. Svensk efterkrigstid är en väl utforskad period, men genom att applicera digitala metoder på kurerade dataset kan politikens, nyhetsmediernas och kulturens sfärer granskas på nytt. I artikeln presenteras vår digitala historieforskning, men fokus ligger främst på erfarenheter och reflektioner kring det praktiska hantverket med att arbeta med storskalig empiri, på iordningställande av dataset och datakurering, samt de möjligheter och tillkortakommanden som sådana forskningspraktiker inbegriper.

"Empiri är datainsamling genom vetenskapliga undersökningar av verkligheten", kan man läsa på Wikipedia. Men hur är egentligen begrepp som empiri och data relaterade? Det är förstås en omtvistad fråga. Men den kan undersökas empiriskt – exempelvis genom att studera versionshistoriken av artikelposten om just empiri på Wikipedia. Även på vårt svenska språk är data om posten omfattande: under fliken "Visa historik" återfinns hundratals ändringar publicerade under snart 20 års tid. "Empiri är vetenskapliga undersökningar av verkligheten", hette det i den första artikelversionen från 2002. Några år senare hade den inledande meningen ändrats till: "Empiri innebär när människan bygger en slutsats utav våra sinnen", en något ålderdomlig formulering, men inte uppseendeväckande eftersom svenskt innehåll på Wikipedia ibland baserar sig på

Artikeln har granskats av två externa lektörer enligt modellen *double blind peer review*.

Pelle Snickars (f. 1971) är professor i digitala kulturer vid Lunds universitet. Han leder forskningsprojekten "Välfärdsstaten analyserad. Textanalys och modellering av svensk politik, media och kultur, 1945–1989" (Vetenskapsrådet) och "Moderna Tider 1936" (Riksbankens Jubileumsfond).

E-post: pelle.snickars@kultur.lu.se

(mer eller mindre) automatgenererad, aggregerad text från inskannade äldre källor på webben, i detta fall Uggleupplagan av *Nordisk familjebok*. Och det är förstås inte enbart text som autogenereras, en betydande del av ändringarna i artikelposten om empiri är utförda av botar, korta städalgoritmer som friserar (i regel) ovidkommande ändringar. Loggen för artikelposten vimlar av sådana ändringar: ”robot tar bort, gör ogjord, rullade tillbaka redigeringar” etcetera. Dataflöden av text går alltså i olika riktningar, men all data om posten empiri går enkelt att studera online på Wikipedia. Vem som gjort ändringar framgår, och det är även möjligt att grafiskt via applikationen RevisionSlider jämföra olika poster med varandra, ”bläddra interaktivt i historiken”, där val av publiceringstidpunkt kan specificeras. Sidstorlek på artikelposten (i antal byte) framgår då, liksom positiv eller negativ ändringsstorlek, det vill säga om det lagts till text eller tagits bort. Och är man lite fingerfärdig går empirin – om posten empiri – att sätta samman och ladda ned som ett dataset.[1]

Att digitalt källmaterial ställer historikern inför nya utmaningar är välbekant. Men vilken roll spelar skala och storlek på den empiri som historikern arbetar med? Kan ett dataset på flera hundra miljoner ord betraktas som historisk empiri? Om ja – vad är det då för sorts forskningspraktik som etableras när en sådan massiv empiri inte längre går att överblicka utan datorers hjälp? Syftet med den här artikeln är att resonera kring den typen av frågeställningar med utgångspunkt i ett forskningsprojekt som jag leder: ”Välfärdsstaten analyserad. Textanalys och modellering av svensk politik, media och kultur 1945–1989”. Inom ramen för detta projekt arbetar vi med olika typer av algoritmisk textanalys av storskalig empiri från politikens sfär (allt digitaliserat riksdagstryck och alla statliga offentliga utredningar under perioden), digitaliserad dagspress (det rör sig om tiotalet dagstidningar) och skönlitteratur (alla svenska romaner utgivna under perioden håller på att digitaliseras), därtill har periodens mest prestigefulla kulturtidskrift, *Bonniers litterära magasin* digitaliserats och iordningställts som dataset. Det rör sig om en historisk empiri grupperad i flera dataset, vilka sammantagna omfattar hundratals miljoner ord.[2]

1. Artikelposten om empiri på svenska Wikipedia återfinns på <https://sv.wikipedia.org/wiki/Empiri> (1/4 2022).

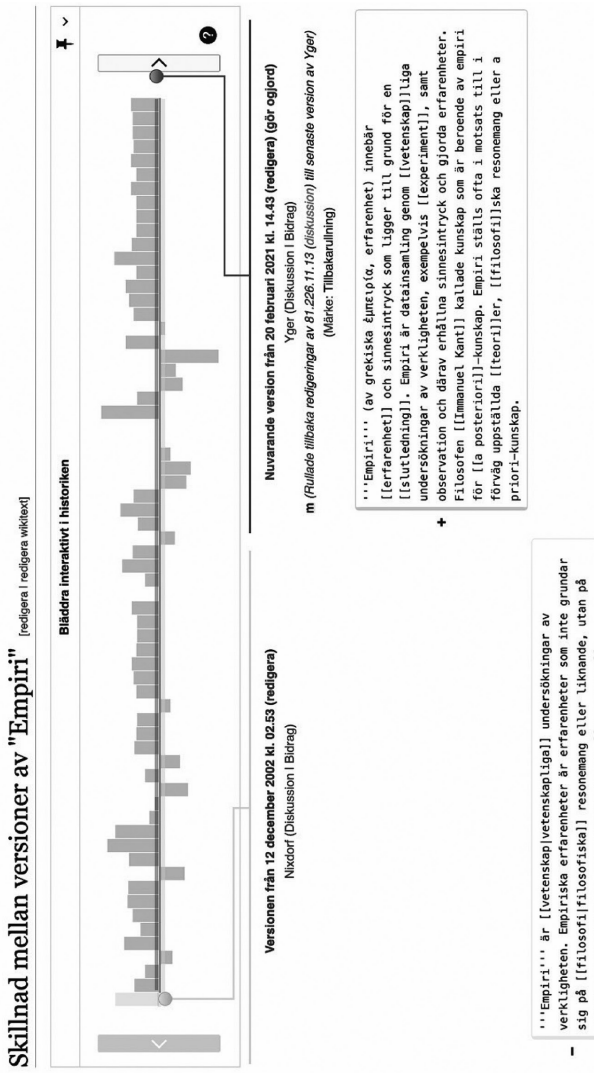
2. Forskningsprojektet ”Välfärdsstaten analyserad. Textanalys och modellering av svensk politik, media och kultur 1945–1989” är finansierat av Vetenskapsrådet (2019–2024) inom ra-

Lingvister och språkteknologer har under lång tid arbetat med snarlika, omfattande textkorpusar. I vårt projekt använder vi både enklare språkteknologiska metoder som ordfrekvenser och kollokationer, *bigrams* (ordpar som uppträder tillsammans), liksom mer sofistikerade algoritmer som temamodeller eller NER-algoritmer (*Named-entity recognition*) där personnamn, geografi och tidsmarkörer automatiskt kan taggas upp. Men vi är inte språkvetare, det är inte språkets form som intresserar oss utan texters historiespecifika innehåll.[3] "Välfärdsstaten analyserad" är ett digitalt historievetenskapligt forskningsprojekt, grundtanken är att studera diskursiva förändringar på makronivå inom tre olika samhällssfärer. Svensk efterkrigstid är en väl utforskad period, men genom att applicera digitala metoder på dessa kurerade dataset kommer politikens, nyhetsmediernas och kulturens sfärer att kunna granskas på nya sätt.

I den här artikeln kommer jag att presentera en del av den forskning vi ägnat oss åt inom vårt projekt – samt några tentativa resultat. Men mitt fokus ligger främst på det praktiska hantverket med storskalig empiri, på iordningställande och arbete med datakurering, samt de möjligheter och tillkortakommanden som sådana forskningspraktiker inbegriper. En fråga jag intresserar mig för är relationen mellan data och empiri. I vårt projekt har vi dels arbetat med redan existerande dataset (hämtade från Riksdagens öppna data) men även skapat nya dataset utifrån äldre textsamlingar som digitaliserats. All digitalisering innebär en sorts medietransfer, och som mediehistoriker – med ett intresse för forskningsarbete med digitala metoder – har jag återkommande funderat på vad som egentligen utgör vår empiri, och vilken relation termer som data och empiri egentligen har till varandra. Artikeln är ställvis

men för forskningsprogrammet Digarv. För mer information, se <https://www.westac.se> (1/4 2022).

3. Språkbanken på Göteborgs universitet erbjuder via webben analysmöjligheter av stora textkorpusar – såväl äldre som mer moderna – genom applikationen Sparv eller konkordansverktyget Korp. Fokus är dock främst språkvetenskapligt: nyckelord, ordbild, ordattribut, ordformer etcetera. Nackdelen med Korp är också att flera av de historiska textkorpusar som används har bristfällig OCR. Samtidigt är det enkelt att få fram den kontext (i originaltexter) som ett ord förekommer inom. Språkbankens verksamhet är därtill kopplad till den europeiska forskningsinfrastrukturen för språkteknologi, CLARIN (Common Language Resources and Technology Infrastructure). Genom den nationella noden Swe-Clarin erbjuds språkteknologiska verktyg och resurser till forskare inom humaniora och samhällsvetenskap i syfte att främja forskning med digitala verktyg (bortom den strikt språkvetenskapliga). Se exempelvis Swe-Clarins digitala handbok, <https://sweclarin.se/swe/handbok> (1/4 2022).



FIGUR 1. *Textuell data om två olika versioner av artikelposten om empiri på svenska Wikipedia. Via applikationen RevisionSlider kan jämförelser göras mellan olika artikelversioner – i detta fall mellan den första från december 2002 och den sista (i skrivande stund aktuella) versionen från februari 2021. Höjden på staplarna anger mängden data (storlek i byte) för varje version av posten, som också kan vara negativ om ändringar inneburit att text raderats. Skärmdump från Wikipedia.*

personligt hållen och innehåller en rad reflektioner kring historikerns arbete med storskalig textuell empiri. En erfarenhet är att historikerns förhållande till empiri ändrar karaktär när denna blir så omfattande att den inte längre går att överblicka (utan datorers hjälp). En annan är att själva forskningspraktiken förändras, där exempelvis utfall av körningar gör att frågeställningar måste korrigeras. Det handlar inte bara om lära sig att förstå kod (någorlunda), eller att samarbeta med programmerare. Som historiker måste man också konsultera statistiker och maskininlärningsexperter för att iordningställa och kurera dataset, liksom att arbeta med data på nya sätt: att ladda olika modeller, att laborera med varierande ingångsvärden, att ändra parametrar och iordningställa utfall på ett sätt som ofta skiljer sig betydligt från traditionell historievetenskaplig praktik.

Att temamodellera det förflutna

För en människa är det inte möjligt att läsa ett dataset med flera hundra miljoner ord – men med mjukvara går det. Termen läsning har i datalogiska sammanhang dock en annan innebörd än den gängse. Den datalogiska funktionen *read/write* utgör exempelvis elementa för de flesta filer, hårddiskar och operativsystem. Att läsa in data i Python är alltså inte samma sak som att låta ett kortkommando för ett ljudprogram ”läsa upp” text på ens skärm. Olika former av *text mining* handlar därför strängt taget inte om att läsa – utan snarare om att mjukvara upptäcker mönster i textmassor, samband som kan vara av olika karaktär beroende på vilken programvara som används. En återkommande frågeställning inom kunskapsfältet digital historia är därför *vad* mjukvara kan uppfatta när algoritmisk textanalys appliceras på ett omfattande empiriskt material. Jo Guldi har i en analys av politiska diskussioner om brittisk infrastruktur under 1800-talet, baserade på *the Hansard* – transkriberade debatter från Storbritanniens parlament, ett dataset på flera hundra miljoner ord – påpekat att storskalig textanalys gör det möjligt att spåra ”the invisible categories that structure mind, language, priorities, or prejudice in a given corpus”.[4] Andra forskare har närmat sig en snarlik omfattande empiri utifrån ett begreppshistoriskt perspektiv. Matti La Mela

4. Jo Guldi, ”Parliament’s debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change”, *Technology and Culture* 60:1 (2019). Textkorpuset för Hansard finns öppet tillgängligt på <https://www.english-corpora.org/hansard> (1/4 2022).

har till exempel studerat hur ett begrepp som allemansrätten förändras i diskussioner och debatter i den finska riksdagen under 1900-talet.[5] En annan metod som blivit vanlig bland digitala historiker är så kallad *topic modeling* – temamodellering på svenska – ett samlingsnamn på en rad algoritmer som kan klassificera stora textkorpusar baserat på tematiska strukturer i textmassan. Genom temamodellering kan diskurser och begrepp urskiljas, baserade på ords statistiska sammanhang i ett mycket omfattande empiriskt material.[6]

Ett dataset brukar ses som en större textsamling av strukturerad data. Själv brukar jag använda beteckningen historiska eller kulturella dataset som ett sätt att antyda att data kommer från en kulturhistorisk sfär, exempelvis bestående av äldre dagspress, tidskrifter eller utredningar. Som med all källkritik är det viktigt att känna till hur ett dataset är sammansatt, vad dess delar består av, samt hur det eventuellt reproducerar förutfattade meningar eller fördomar. Kulturella dataset har inte sällan sin egen specifika etnografi. Ibland går dataset med en språkteknologisk terminologi under beteckningen textkorpusar, men den mediehistoriska forskning jag bedriver handlar inte enbart om text utan även om andra modaliteter, varför jag föredrar termen dataset.[7] Ett historiskt dataset som jag själv återkommande arbetar med innehåller alla statliga offentliga utredningar mellan 1945 och 1989 – det rör sig om fler än 3 000 utredningar.[8] Använder man den algoritmiska temamodellen

5. Matti La Mela, "Tracing the emergence of Nordic *allemansrätten* through digitised parliamentary sources", i Mats Fridlund, Mila Oiva & Petri Paju (red.), *Digital histories: Emergent approaches within the new digital history* (Helsinki 2020) s. 181–197.

6. För en introduktion till och applicering av temamodellering på SOU-data, se Pelle Snickars, "Från chiffer till klartext? Temamodellering av statliga offentliga utredningar 1945–1989", *Scandia* (kommande).

7. I ett annat forskningsprojekt på Humlab vid Umeå universitet, European History Reloaded, intresserar vi oss till exempel för hur audiovisuellt kulturarv i dag återanvänds på webben genom olika former av remixkulturer på videoplattformar som Youtube eller Vimeo. De dataset vi arbetar med består här alltså av rörlig bild. I projektet analyseras videoåterbruk genom en en programvara, Video Reuse Detector (som vi utvecklat) vilken algoritmiskt kan spåra hur ett visst bildmaterial återanvänts. För en vidare diskussion, se Maria Eriksson, Tomas Skotare & Pelle Snickars, "Understanding Gardar Sahlberg with Neural Nets: On Algorithmic Reuse of the Swedish SF-archive", *Journal of Scandinavian Cinema* (kommande). För mer info om projektet, se <https://www.cadeah.eu> (1/4 2022).

8. Det svenska kommittéväsendet har inom historievetenskaplig forskning ofta betraktats som en kunskapsgenererande verksamhet, liksom ett gott källmaterial "för historisk forskning", som Klas Åmark påpekat i *Hundra år av välfärdspolitik* (Umeå 2005) s. 32. Men sammansatta som ett enda dataset kan SOU-datan naturligtvis också ställvis vara partisk: att besvärliga samhällspolitiska frågor parkerades i endera utredning har förstås hänt – utredningar kan i sådana fall betraktas som statens dåliga samvete.

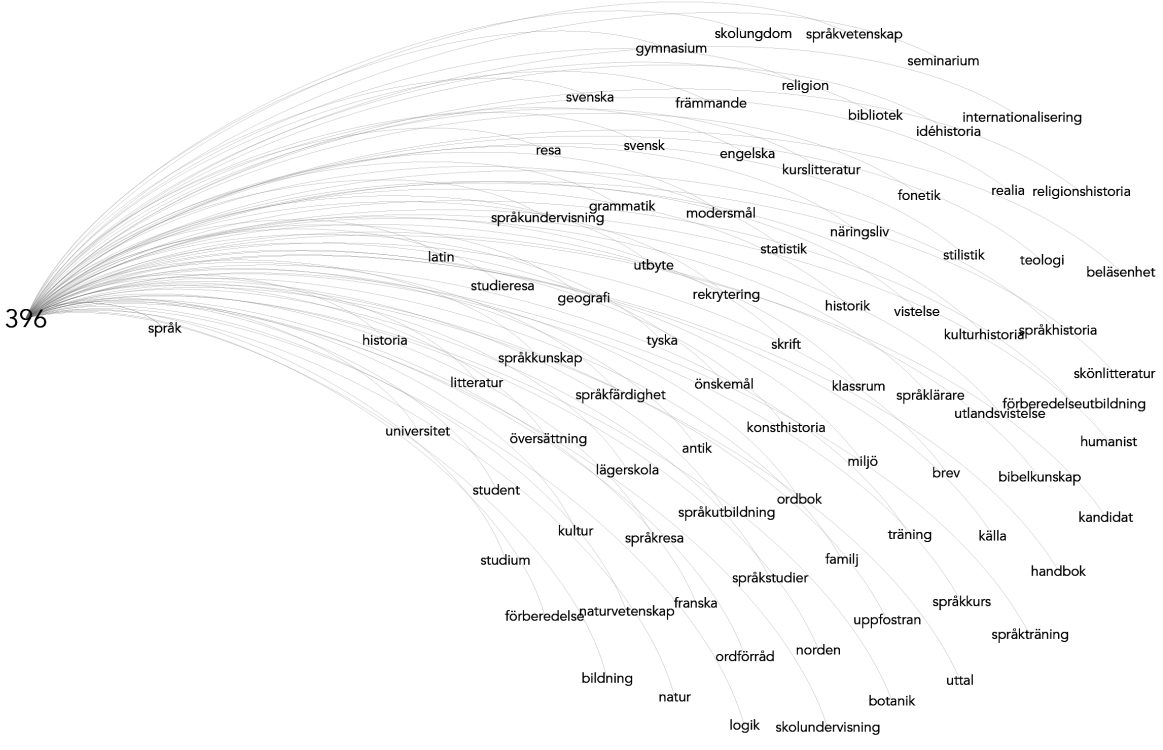
latent dirichlet allocation (lda) och i just detta fall sorterar datan i 500 teman, så framträder ett som tydligt handlar om humaniora, historia och språk. I ett flertal SOU:er återfinns alltså ett humanioratema – i varierande procentuell styrka. I temamodellen urskiljs de ord som ingår, exakt vilka utredningar som innehåller detta tema, hur det fördelar sig över tid liksom hur det anknyter till andra teman och utredningar. Humanioratemat (godtyckligt numrerat som 396) är som starkast fram till omkring 1970 – det innehåller frekventa ord som språk, historia, universitet, latin och litteratur, men även termer som beläsenhet och humanist. Temat kan i sin tur sättas i relation till andra teman som det delar ord med, till exempel ett universitetstema (med nummer 330), där de mest frekventa termerna är fakultet, universitet, högskola, ämne och undervisning, eller med ett kulturarvstema (nummer 386) med de mest prominenta orden byggnad, museum, kulturminnesvård, landsantikvarie och riksantikvarieämbetet.

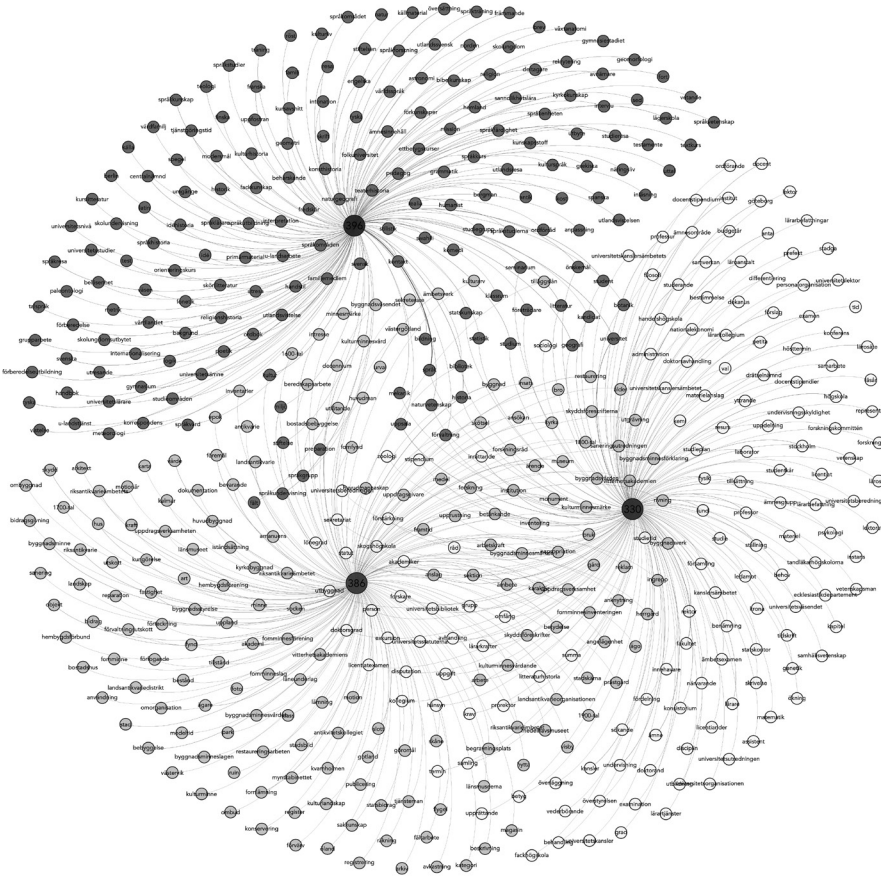
Figur 2–4 ger en antydan om att statliga utredningar som sammansatt dataset är ett rikt forskningsmaterial, ett slags omfattande textarkiv som kan analyseras, läsas och visualiseras på många olika sätt, inbegripet såväl tematiska som kronologiska nedslag. Grundprincipen för all temamodellering är att ord samförekommer. Temamodeller arbetar med statistisk analys av distribution av ord i form av sannolikhetsfördelning över alla ord i en textmassa. I fallet med humanioratema 396 så förekommer det i knappt två procent av de inalles 3 154 utredningar från perioden. Det låter kanske lite, men det gäller enbart detta mycket specifika tema. SOU-data från perioden inbegriper 87 miljoner ord, en historisk empiri som är betydligt mer omfattande än den som historievetenskapen i regel använder sig av.

När en beläsen humanist (för att nu anknyta till temat) som Reinhart Koselleck i sin *Vergangene Zukunft* (1979) metodologiskt redogör för sin begreppsorienterade empiri heter det att den baserar sig på läsning under 20 års tid av tryckta tidsupplevelser formulerade av politiker, filosofer, teologer och diktare, liksom mer obekanta skrifter och ordböcker – samt förstås historikers egna utsagor. ”Solche Texte wurden gesucht und befragt”, som det står i original.[9] Empirin har efter betydande forskarmödor tålmodigt letats fram i bibliotek och arkiv, ibland systematiskt

9. Reinhart Koselleck, *Vergangene Zukunft: Zur Semantik geschichtlicher Zeiten* (Frankfurt 1989) s. 11.

FIGUR 2. *Tema humaniora i statliga offentliga utredningar från en så kallad 500-gensim-lda-modell som urskiljer 500 teman i ett dataset av fler än 3 000 SOU:er mellan 1945 och 1989. Skärmdumpen av humanioratema 396 är tagen från en Jupyter Lab-miljö – utvecklad vid Humlab på Umeå universitet – och visar de 100 vanligast förekommande orden i temat; språk är mest frekvent följt av historia.*


396
språk
historia
latin
studieresa
geografi
litteratur
universitet
student
studium
förberedelse
bildning
natur
logik
skolundervisning
botanik
uttal
norden
ordförråd
uppfostran
naturvetenskap
franska
språkkurs
språkträning
handbok
familj
språkstudier
kultur
språkresa
språkutbildning
ordbok
kandidat
träning
källa
lägerskola
antik
översättning
konsthistoria
miljö
brev
bibelkunskap
humanist
förberedelseutbildning
utlandsvistelse
språklärare
klassrum
önskemål
språkfärdighet
språkkunskap
tyska
skrift
kulturhistoria
språkhistoria
skönlitteratur
historik
vistelse
rekrytering
utbyte
statistik
stilistik
teologi
beläsenhet
näringsliv
modersmål
grammatik
språkundervisning
fonetik
realia
religionshistoria
resa
svensk
engelska
kurslitteratur
svenska
främmande
bibliotek
internationalisering
idéhistoria
religion
gymnasium
seminarium
skolungdom
språkvetenskap



FIGUR 3–4. *Nätverken är gjorda i visualiseringsverktyget Gephi: humanioratema 396 plottat mer överskådligt med de cirkla 80 vanligast förekommande orden (från vänster till höger), termer som återkommer i det cirkulära nätverket (med algoritmen Fruchterman Reingold) med koppling till två andra teman – 330 (om universitet) och 386 (om kulturarv). Nätverket innehåller drygt 500 ordnoder där centrala ordnoder som bibliotek, historia och förvaltning (i mörkare grått) förbinder samtliga teman. Skärmdump från Jupyter Lab-miljö vid Humlab, Umeå universitet.*

får man förmoda, men stundtals också mer eller mindre godtyckligt. För det är ju så många historiker ofta arbetar, även jag själv – vi kan inte läsa allt. Devisen ”gesucht und befragt” gäller förstås inte för alla historiska studier, men ganska många skulle jag vilja hävda. Som mediehistoriker har jag metodologiskt arbetat ungefär på det sättet under mer än två decennier.

Filologer har i hundratals år ägnat sig åt texttolkning med fokus på det innehåll som ett språkligt dokument rymmer. Men i takt med att allt fler sådana dokument – som historisk empiri, förstådd i vid bemärkelse – blivit digitala har samma innehåll både blivit sökbar på nytt, liksom möjligt att analysera med olika typer av språkteknologisk mjukvara. Begreppsanvändning kan nu analyseras på andra sätt, och på grundval av en empiri som är betydligt mer omfattande än tidigare – även om det i fallet med SOU-datan härrör sig från en enda materialkategori. Den traditionella metoden ”gesucht und befragt” har i så måtto ändrat karaktär när arkivdriven historievetenskap kompletterats med datadriven *digital history*, det senare ett kunskapfält som kommit att omfatta flera olika humanistiska discipliner.

> Digital history is an approach to examining and representing the past; it uses new communication technologies and media applications and experiments with computational methods for the analyses, production and dissemination of historical knowledge.[10]

Ofta – men inte alltid – innebär det en metodologisk förskjutning från kvalitativa till kvantitativa studier där empirin ökat rejält i omfång. Andrew Piper har i en syrlig text om litteraturvetenskapens behov av att gå från när- till fjärrläsning, *distant reading*, exemplifierat med en annan tysk lärdomsgigant, Erich Auerbach. Piper har frågat sig om ”verklighetsframställningen i den västerländska litteraturen” – undertiteln till *Mimesis* (1946) – verkligen går att analysera utifrån en skönlitterär empiri bestående av ett tjugotal kanoniserade böcker. Att Auerbach (liksom Koselleck) var en flitig läsare betvivlar ingen: ”Who would ever presume to have read more than Auerbach? But what if he actually hadn’t read *enough*?”[11]

10. Hannu Salmi, *What is digital history?* (Cambridge 2021) s. 7.
11. Andrew Piper, ”There Will Be Numbers”, *Journal of Cultural Analytics* 1:1 (2016).

Lockelsen med historiens Big Data

Vari består egentligen den historiska lockelsen i att arbeta med ett mycket omfattande empiriskt material? Är det den illustrativa kapacitet som storskaliga dataset ofta har för att tjäna som empiriskt underlag för en historieskrivning à la *longue durée*? Är det tjusningen med nya digitala metoder och datorers kraftfullhet som drar? Eller handlar det snarare om ökad tillförlitlighet hos en partikulär undersökning, alternativt förbättrad representativitet för ens forskning rent generellt – eller kanske rentav om validering och ökad reproducerbarhet, där mer eller mindre subjektiva tolkningar kan styrkas och verifieras genom att öppet redovisa den data och mjukvara som använts?

Svaret beror förstås på vem man frågar. Somliga digitala humanister som arbetar med storskalig, historisk empiri skulle hävda att digitala metoder både ska användas för att bekräfta tidigare forskningsresultat (och visa att metoderna fungerar) liksom att ge nya insikter. Andra går längre och menar att det i dag inte längre räcker med en tre, fyra historiska exempel för att dra mer generella slutsatser, ”the anecdotal nature of evidence”, som Piper tillspetsat formulerat saken. ”What precisely does computation allow one to claim that has not been seen before or that was uncertain in the world of anecdotalism?”[12] Ofta handlar det om att konfirmera, ibland om nya upptäckter och inte sällan om en kombination. I ett av de dataset vi arbetat med från nyhetssfären – alla textblock som innehåller *bigram* med ”politisk/politiska” i *Aftonbladet* och *Dagens Nyheter* 1945–1989, ett dataset med cirka 400 000 textblock och 27 miljoner ord – så visade det sig att mot slutet av 1960-talet så var ”politisk teater” en nästan lika vanlig kombination som ”politiskt parti”.[13] Åter andra DH-forskare menar att det just är arbetet med sådana dataset som är det väsentliga. I den nystartade *Journal of Digital History* – ”promoting a new form of data-driven scholarship” – är frågan om empiri den centrala. För varje artikel som publiceras krävs ”a data layer providing access to data and code by means of a professional infrastructure”.[14]

12. Ibid.

13. Fredrik Norén, Johan Jarlbrink, Alexandra Borg, Erik Edoff & Måns Magnusson, ”The transformation of 'the political' in post-war Sweden”, i Estelle Bunout, Maud Ehrmann & Frédéric Clavert (red.), *Digitised newspapers: A new eldorado for historians? Tools, methodology, epistemology, and the changing practices of writing history in the context of historical newspapers mass digitization* (kommande).

14. Tidskriften *Journal of Digital History* drivs av Luxembourg Centre for Contemporary and Digital History (C2DH) i samarbete med förlaget De Gruyter: jag sitter i tidskriftens re-

I linje med sådana direktiv accentuerar ytterligare andra DH-forskare att även humaniora nu mer formellt kan definiera själva forskningsprocessen: vilka frågor som ska ställas, vilket dataset som ska användas, hur data ska kureras och modelleras, vilken programvara som ska användas, vilka algoritmer som ska nyttjas och hur parametersättningen av kod ska styras – i akt och mening att verifiera resultat och göra dem reproducerbara. Och vänder man sig till DH-forskare inom mer estetiskt orienterade ämnen, ofta baserade på en väletablerad kanon, så har frågan om en utökad empiri i regel handlat om att gå (långt) bortom denna. Lev Manovich har exempelvis i sin bok *Cultural Analytics* – som ägnar sig åt mer samtida, storskalig bildanalys – ironiskt påpekat att den kulturella kanon är lika ålderdomlig som seglivad: ”the canon corresponds to the vision of culture articulated in 1875 by Matthew Arnold – culture as 'the best'”. Men kunskap om en sådan smal kulturell kanon hjälper i dag föga, enligt honom, för att begripa en lika världsomspännande som skalbar digital kultur, ”the asthonishing scale of digital culture”.[15] Skriven i skärningspunkten mellan datavetenskap och mediestudier ger Manovichs bok en rad inblickar i analysmetoder för storskaliga kulturella dataset, från *new media* till *more media*. I hans fall handlar det främst om samtidsanalyser, men givet den exponentiellt ökande mängden digitaliserat källmaterial kan i princip samma typ av datadrivna metoder användas på äldre bildmaterial.

Det dataset av statliga utredningar mellan 1945 och 1989 som jag diskuterat ovan kan illustrera den datadrivna forskningens dragningskraft (åtminstone på mig). Jag är disputerad filmvetare och har under många år arbetat mediehistoriskt med statliga utredningar som empiri. I likhet med andra kvalitativt skolade humanister har utredningsmaterial i sådana sammanhang i regel handlat om läsning av ett fåtal SOU:er – ibland något fler, men sällan mer än trettiotalet utredningar.[16] Roger Blomgren

daktionsråd. Grundtanken är att publicera artiklar samt alla dataset som forskningen bygger på. Man kan dock notera att nationella dataset (från ett litet språkområde) ofta är knepiga i en internationell kontext. Att beskriva dem på engelska är en sak, men att översätta ett helt dataset är i regel omöjligt. För mer info, se <https://journalofdigitalhistory.org/en/about> (1/4 2022).

15. Lev Manovich, *Cultural Analytics* (Cambridge, Mass. 2020) s. 1, 119,

16. Vid sidan av de statliga utredningarna bör det framhållas att utredningsarkiven ofta utgör ett rikhaltigt empiriskt material – som mycket sällan har digitaliserats. Jag har själv bland annat arbetat med både Folkminneskommitténs arkiv (från tidigt 1920-talet) och framför allt Dataarkiveringskommitténs efterlämnade arkiv (från 1967–76), det senare samlat i 37 volymer på Riksarkivet. För en vidare diskussion, se Pelle Snickars, ”Mediestudiets infra-

baserade till exempel sin avhandling *Staten och filmen: Svensk filmpolitik 1909–1993* på ett knappt femtontal filmrelaterade utredningar, och i My Klockar Linders avhandling *Kulturpolitik: Formeringen av en modern kategori* används utredningar återkommande som empiri, även om bara tre återfinns separat i referenslistan. Som bekant har många historiker använt detta källmaterial på ett snarlikt sätt: Yvonne Hirdman nyttjade exempelvis ”det unika SOU-materialet” och citerade flitigt från det i sin bok *Att lägga livet tillrätta*. I hennes litteraturlista återfinns 33 utredningar.[17]

Forskningsfrågorna i exemplen skiftar förstås, men utredningsprosa utgör gemensam empirisk nämnare. Den kan läsas individuellt – eller betraktas som ett dataset. I mitt fall har SOU-datan gjorts tillgänglig i utvecklingsmiljön Jupyter Lab där man via ett användarvänligt gränssnitt kan köra Pythonkod (som bearbetar data på olika sätt) direkt i webbläsaren. Om alla SOU:er analyseras som en enda enorm text som staten skriver, vilka teman i denna text kan programvara då läsa och uppfatta? Ja, några av de teman som framträder allra tydligast har just med befolkning och familj att göra, för att knyta an till Hirdmans bok. Det är naturligtvis inte på något sätt rättvist att metodologiskt göra en sådan jämförelse. Hirdmans analys gjordes under en analog tidsperiod och baserade sig på familjeutredningar från 1930-talet, men den tematik hon spårade var riktig, och den växer i styrka i SOU-datan mellan 1945 och 1989. Till och med när man delar upp den i endast 50 teman – modeller med 50 teman fångar upp breda samhälleliga frågeställningar – så återfinns ett tydligt familjerelaterat tema som växer i statistisk styrka från mitten av 1970-talet.

Mitt syfte är här inte att vara polemisk, eller att upprätta någon slags motsättning mellan en äldre historievetenskaplig metodik och en av mer modernt digitalt snitt. Även om jag intresserar mig för digitala sätt att undersöka historisk empiri så bedriver jag också forskning i mer traditionell mediehistorisk bemärkelse. Likväl ska jag inte hymla med att det föreligger en viss skillnad i att empiriskt basera en historisk undersökning på personlig läsning av trettiotalet utredningar i jämförelse med

struktur: Om etableringen av Arkivet för ljud och bild”, i Mats Hyvönen, Pelle Snickars & Per Vesterlund (red.), *Massmedieproblem: Mediestudiets formering* (Lund 2015) s. 55–103.

17. Roger Blomgren, *Staten och filmen: Svensk filmpolitik 1909–1993* (Stockholm 1998); My Klockar Linder, *Kulturpolitik: Formeringen av en modern kategori* (Uppsala 2014); Yvonne Hirdman, *Att lägga livet tillrätta: Studier i svensk folkhemspolitik* (Stockholm 1990) s. 22.

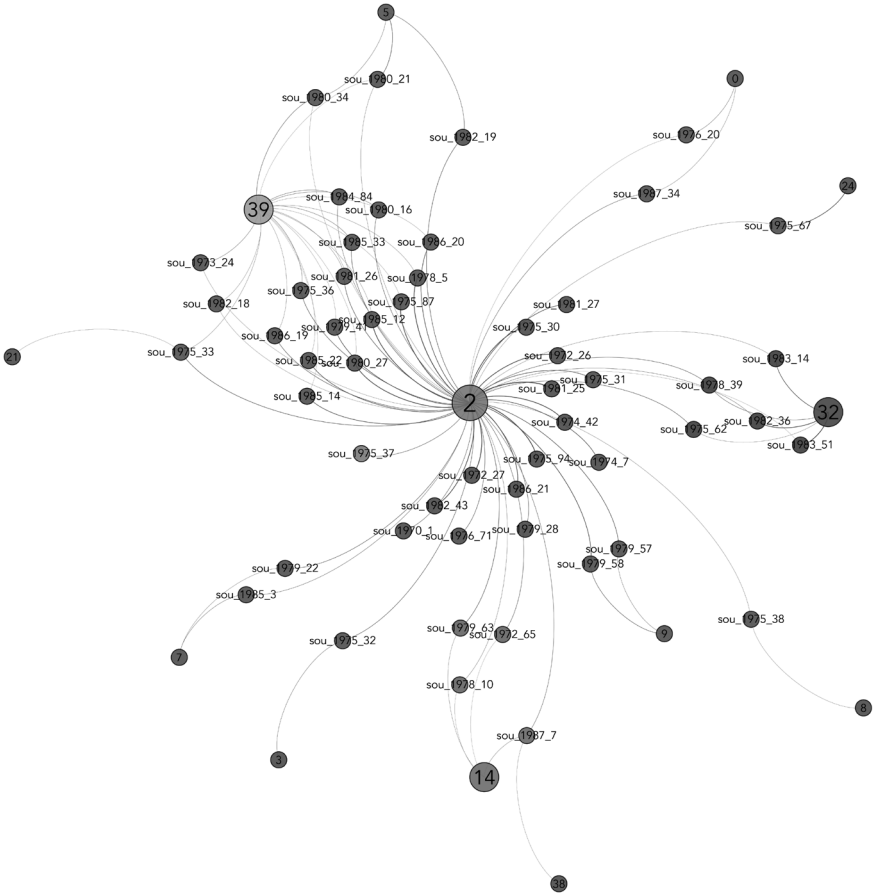
3 000 – låt gå att det är mjukvara som då tar sig an textmaterialet. Som sagt: de frågeställningar som styrt forskningen är väsensskilda. Men de baserar sig på samma empiriska utredningsmaterial, som i det ena fallet är småskaligt och i det andra enormt. Häri återfinns en betydande lockelse med en mer omfattande textuell empiri, i alla fall för mig. För om äldre begreppsanalytisk forskning använde sig av ett fåtal nyckeltexter i enlighet med metoden "gesucht und befragt", så är det i dag möjligt att komplettera en sådan, ofta ganska smal empirisk forskningstradition med mer storskalig textanalys förutsatt att empirin finns i digital form.

Ett annat enkelt tillgängligt dataset som vi arbetar med och bearbetat i vårt projekt är Sveriges riksdags protokoll. 2010 öppnade riksdagen sina databaser för fri användning, och under tiotalet så utökades data.riksdagen.se successivt med allt fler dataset: på data.kb.se återfinns bland annat äldre riksdagstryck (från 1521–1970) i varierande kvalitet och omfång. Med utgångspunkt i Riksdagsförvaltningen digitaliseringsarbete har vi i vårt projekt skapat ett kurerat dataset av alla protokoll från både första och andra kammaren under de senaste 100 åren. Drygt 12 000 protokoll av varierande längd gör att det innehåller omkring 400 miljoner ord. Vi har bland annat undersökt hur olika begrepp – som exempelvis information, upplysning och propaganda – förändrat sin betydelse över tid under 1900-talet i riksdagens debatter.[18] Men framför allt har äldre dataset annoterats med metadata för talare, geografisk hemvist, partitillhörighet och kön – ett arbete som kan beskrivas med en mening, men som varit mycket omfattande och tidskrävande.

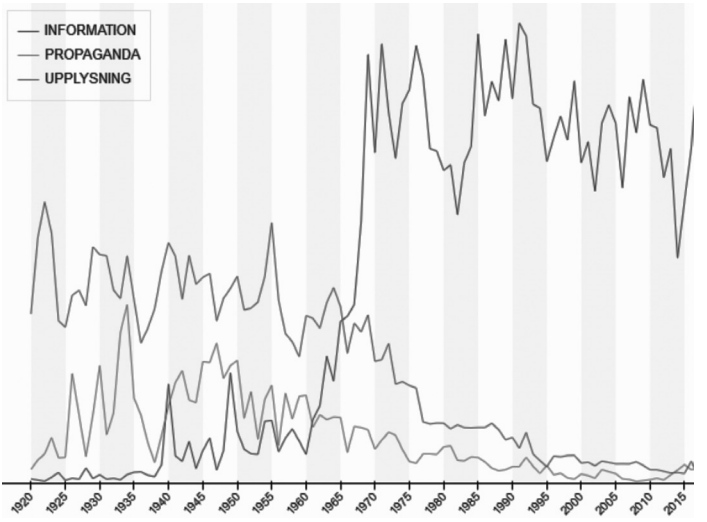
Liksom vid all digitalisering förekommer problem i riksdagsdatan med *optical character recognition* (OCR), och därtill har det funnits behov att strukturera textmassan av riksdagsprotokoll på olika sätt. Det säger sig självt att i ett dataset med cirka 750 000 anföranden så måste sådana (och andra) frågor lösas på algoritmisk väg. Även om det i protokollen i regel framgår vilken text som tillhör ett anförande och vad som är anteckningar (till och i protokollen), så är detta någonting vi behövt använda maskininlärning för att deducera och leda i bevis. För det har en algoritm tränats på ett mindre, manuellt annoterat dataset i syfte att kunna dra datadrivna slutsatser. Arbetet har inbegripit att

18. För en vidare diskussion, se Johan Jarlbrink & Fredrik Norén, "The rise and fall of legitimate propaganda: A digital reading of Swedish parliamentary records, 1867–2019", *Scandinavian Journal of History* (kommande).

FIGUR 5. *Skärmdump av stapeldiagram i Jupyter Lab-miljö för temamodell 2 om familj och barn (med 200 relaterade ord) i 50-mallet-lda-modellen av SOU-datan 1945–1989.*



FIGUR 6. *Nätverket i Gephi (med algoritmen Force Atlas 2) visar samma familjetema (ur samma 50-tema-modell) med ett tröskelvärde på 0,2 – det vill säga, familjetemat ska figurera till åtminstone 20 procent i de utredningar som listas. Kopplingar återfinns till ett femtiotal utredningar från perioden 1970–1989, samt till tema 39 (om statlig förvaltning och samhällelig planering), tema 32 (om inkomst och ersättning) samt till tema 14 (om lagstiftning). När utredningar skriver om familjefrågor så behandlar de under denna period också planerings- och ersättningsfrågor, liksom delvis juridiska spörsmål. Skärmdump från Jupyter Lab-miljö vid Humlab, Umeå universitet.*



FIGUR 7. *Mer information i Sveriges riksdag? Översiktliga begreppstrender i riksdagen baserade på en historisk empiri om 400 miljoner ord ur riksdagsprotokoll. Den lemmatiserade och normaliserade grafen visar begreppsfrekvens av termerna information, propaganda och upplysning i ett dataset som består av samtliga protokoll från första och andra kammaren under nästan 100 år. Skärmdump från Jupyter Lab-miljö vid Humlab, Umeå universitet.*

identifiera anförandens början och slut i protokollen, liksom den riksdagsledamot som talar (när det inte explicit framgår). I huvudsak har det gjorts genom att påvisa språkliga mönster i protokollen och att använda dem för att klassificera textstycken. Men problem uppstår exempelvis när det finns talare med samma efternamn; annan metadata såsom geografisk hemvist har då använts för att särskilja individer. Efter att ny metadata implementerats har vi genomfört manuella kontroller i syfte att undersöka kvalitén på datan, liksom att hitta nya sätt att ytterligare förbättra datans beskaffenhet, ett rekursivt arbete som resulterat i olika dataversioner av vår empiri. I praktiken har det inneburit annotering av ett dataurval där både korrekta och inkorrekta klassificeringar noterats, för att på så vis kunna identifiera felkällor i den algoritm som använts.

Datakurering genom maskininlärning är komplicerat – även för skolade statistiker. Jag vill därför fästa läsarens uppmärksamhet på att storskalig textanalys som forskningspraktik är mycket olik närläsning av ett

fåtal texter. Det ena utesluter som sagt inte det andra, båda inbegriper hermeneutiska övningar och ofta handlar det om en kombination, speciellt när resultat ska presenteras i artikel- eller bokform. Men den empiriska skalan är annorlunda – något som ett halvår av intensiv kurering och iordningställande av vår riksdagsdata vittnar om – fjärrläsningens kunskapsformer likaså: ”Distant reading [...] *is a condition of knowledge*”, som Franco Moretti slagit fast.[19] Inte minst blir det i en digital analysmiljö – i vårt fall via så kallade notebooks i Jupyter Lab – möjligt att vrida och vända på det dataset man arbetar med. Där kan man som forskare via kodceller med knappar och manuella reglage styra parametersättningen av koden och laborativt experimentera: att dela in och modellera data på olika sätt eller explorativt röra sig mellan ett storskaligt dataset och de specifika riksdagsprotokoll (eller statliga utredningar) som utgör grunden för empirin. Enskilda digitala protokoll går ju enkelt att läsa på webben och alla SOU:er finns öppet nedladdningsbara via Kungliga bibliotekets hemsida.

En erfarenhet av sådant forskningsarbete, där empiri och metod sammanförs i ett slags infrastrukturellt gränssnitt, är att det ibland är svårt att hålla isär vad som är vad. Körningar med empiri omvandlad till dataset framträder först när metoden man använder genererar ett resultat. Om själva iordningställandet av empiri till maskinläsbara dataset är ett första steg, och att ladda data – eller snarare modell, *load model* – utgör nästa, så blir den empiri som forskningen utgår från först tillgänglig som mjukvarans *output*. Själva utforskandet är därför helt beroende av en för ändamålet uppsatt infrastruktur, där det laborativa skärmarbetet också inbegriper andra typer av programvara vid sidan av Jupyter Lab-miljön. En sådan förändrad forskningspraktik innebär därför inte bara användningen av en specifik infrastruktur (i det lilla) och nya digitala metoder, den inbegriper också ett nytt slags förhållande till den empiri som undersöks. I fallet med SOU-datan (och i viss mån data från riksdagen) har arbetsgången i regel handlat om att söka fram begrepp eller teman,

19. Franco Moretti, *Distant reading* (London 2013) s. 48. Moretti är litteraturvetare, men inom DH-forskning som använder sig av omfattande textuella dataset har hans ursprungliga begrepp ”distant reading” – distans- eller fjärrläsning på svenska – blivit till ett slags övergripande term för storskalig empirisk textanalys. De (få) svenska litteraturvetare som använder sig av termen har på sistone översatt den till ”fjärrläsning”, se exempelvis Karl Berglund, ”Introduktion till fjärrläsning”, i Johan Jarlbrink & Fredrik Norén (red.), *Digitala metoder inom humaniora och samhällsvetenskap* (Lund 2021) s. 187–210.

se efter vilka utredningar som ingår, få korn på i vilka utredningar ett tema är som starkast, ladda ned tabelldata och därefter visualisera resultat i Gephi – standardprogramvara för nätverksvisualiseringar – för bättre överblick.

Det torde vara uppenbart för var och en att en sådan forskningspraktik är beroende av omfattande programmeringsinsatser. Digital humaniora inbegriper också i regel forskningspraktiker i flera led – vilka föregriper själva undersökningen. För historiker som är vana att arbeta på egen hand kan detta verka avskräckande. Men att fjärrläsningens kunskapsformer varit lockande är samtidigt uppenbart – forskning inom DH-fältet är ofta projektbaserad, sker i kluster och involverar olika kompetenser. Utveckling av digitala metoder förutsätter tvärvetenskapligt samarbete, liksom att programmerare tidigt kommer in i forskningsprocessen och driver den framåt i återkommande dialog.[20] Sådan humanistisk forskning kräver dock en annan infrastruktur än den gängse, vilket varit ett återkommande tema i forskningspolitiska inspel och diskussioner från senare år. I den allra senaste utredningen, *Stärkt fokus på framtidens forskningsinfrastruktur* ägnas visserligen humaniora minimalt med intresse – med undantag för just storskalig textanalys. Kungliga bibliotekets satsning på att etablera ett datalabb lyfts där fram som ett innovativt sätt ”för forskare att utföra storskaliga analyser av bibliotekets samlingar [...] Biblioteket utvecklar nya modeller, till exempel med metoder för artificiell intelligens, för att analysera samlingarna på ett nytt sätt”.[21] Några år tidigare ägnades frågan mer utrymme, i regeringens forskningsproposition 2016 påtalades nämligen behovet att stödja och främja ”datadriven forskning, särskilt inom humaniora och samhällsvetenskap.” Den typen av forskning och kunskapsproduktion utgår inte sällan från digitaliserat ”material som i dag finns vid kulturarvsinstitutionernas arkiv, samlingar och bibliotek [vilka både] ökar tillgängligheten och skapar nya möjligheter för forskning ... Ökad tillgång till digitala datavolymer öppnar för att besvara nya frågeställningar och för metodutveckling”.[22] Dessa for-

20. Pelle Snickars, ”Google Docs, digital humaniora & akademiskt samarbete”, i Per Olof Erixon & Julia Pennlert (red.), *Digital humaniora: Teori, metod & exempel* (Göteborg 2017) s. 121–145.

21. SOU: 2021:65 *Stärkt fokus på framtidens forskningsinfrastruktur* (Stockholm 2021) s. 217. KB-labb etablerades 2019, det var jag som arbetade med och skrev förstudien, *Datalabb på KB*, <https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-339> (1/4 2022).

22. Prop. 2016/17:50 *Kunskap i samverkan – för samhällets utmaningar och stärkt konkurrenskraft* (Stockholm 2016) s. 95.

muleringar resulterade i ett speciellt anslag till Vetenskapsrådet – som i samarbete med Riksbankens Jubileumsfond och Kungl. Vitterhetsakademien – i två omgångar därefter lyst ut forskningsmedel till Projektbidrag för digitalisering och tillgängliggörande av kulturarvssamlingar, i vad som kommit att kallas för Digarv-programmet. I dagsläget är det 14 större forskningsprojekt som finansierats inom denna satsning, bland annat vårt projekt ”Välfärdsstaten analyserad”. Det är visserligen långt ifrån alla Digarv-projekt som ägnar sig åt storskalig empirisk forskning, men flera projekt hämtar metodologiskt inspiration från snarlik DH-forskning – från maskininlärning av hällristningsdokumentation för att analysera, identifiera och klassificera hällristningar över datorbaserad ansiktsigenkänning i äldre fotografiska dataset till storskalig analys av ljudinspelningar av debatter i riksdagen.[23]

Kring den här typen av forskningspolitisk styrning är det förstås många som reser invändningar. För egen del betraktar jag utvecklingen som en tämligen naturlig fortsättning av diverse digitaliseringsinsatser inom kulturarvssektorn. I en första fas digitaliserades material i bevarandesyfte liksom för att öppet tillgängliggöras, i en andra fas kan forskning nu bedrivas på dessa ökade datavolymer med möjlighet att ställa – och ibland besvara – nya frågeställningar. Men för att utforska sådana dataset krävs andra kompetenser än rent historievetenskapliga – etableringen av en rad DH-lab på svenska lärosäten under senare år bär här syn för sägen. Metoder, maskinpark och mjukvara blir också ständigt mer sofistikerade och kraftfulla.

Jag är naturligtvis medveten om att historiker länge arbetat kvantitativt med en omfattande empiri – det är knappast något som digitaliseringen av kulturarv och källmaterial resulterat i. 1969 beklagade sig exempelvis Jörgen Weibull i en artikel i *Scandia* att moderna reproduktionstekniker gjort att mängden historiskt material ökat så till den grad ”att detta ofta blir det allt överskuggande problemet”.[24] Och redan då tänkte man sig att datamaskinerna skulle komma till undsättning. Som min kollega Johan Jarlbrink påpekat var det många som skrev om historievetenskapliga metodfrågor under 1960- och 1970-talen som ”uttryckte

23. Tillsammans med konstvetaren Anna Dahlgren koordinerar jag Digarv-programmet. För mer information om de forskningsprojekt som ingår, se <https://www.digarv.se/deltagande-forskningsprojekt> (1/4 2022).

24. Jörgen Weibull, ”Metodologiska problem i modern historia”, *Scandia* 35:1 (1969).

förhoppningar om datamaskinernas effektivisering av forskningen och att de skulle möjliggöra nya och storskaliga undersökningar".[25] Weibull var en av dem:

> med utnyttjande av de möjligheter, som de moderna datamaskinerna givit oss, står ett väldigt nytt källmaterial till historikernas förfogande: ett massmaterial av en typ som tidigare inte har kunnat utnyttjas helt enkelt på grund av sin mängd. Med datamaskinens hjälp kan forskaren av idag överblicka, systematisera och behandla nästan obegränsade mängder av dylika data.[26]

För mer än 50 år sedan formulerades med andra ord snarlika förhoppningar som idag inte sällan präglar retoriken kring digital humaniora. Själva lockelsen att med datorns hjälp ta sig an svällande historiska dataset har sin egen historiografi.

Att kurera empiri till data (eller tvärtom)

I sin idé- och mediehistoriska avhandling *Kretslopp av data* har Johan Fredrikzon undersökt vad han kallar för "den tidiga digitaliseringens" datapraktiker mellan cirka 1965 och 1985. Han hävdar att det framför allt var under 1970-talet som data kom att bli ett samhälleligt fenomen i Sverige. Statstjänstemän upphörde då att skriva för att istället mata in data, forskare övergick från att studera naturens förlopp till att bygga datamodeller, och ilskna medborgare krävde av integritetsskäl tillgång till sin personliga data med hjälp av kuponger de klippte ur dagspressen. Fredrikzon menar därför att det under denna period går att förstå vitt skilda områden, "ekologisk forskning, fastighetskartering, miljögiftlagstiftning, rapportskrivande, arkivvård – som 'dataproblem'". Bland annat intresserar han sig för Statistiska centralbyråns praktiska blankettarbete med Folk- och bostadsräkningen 1970. SCB:s digitaliseringsfabrik föreföll vara en "helautomatisk anordning där papperslappar från Sveriges befolkning flöt in på ena sidan och sökbara databaser presenterades i andra änden". Men så var inte riktigt fallet. Om Weibull något år tidigare inbillat sig att dåtidens datamaskiner skulle ha förmågan att behandla nästan obegränsade mängder av historiska data, så var

25. Johan Jarlbrink, "Historievetenskapens mediehantering", i Mats Hyvönen, Pelle Snickars & Per Vesterlund (red.), *Massmedieproblem: Mediestudiets formering* (Lund 2015) s. 225–248.
26. Weibull 1969.

man på SCB tidigt på det klara med att tidens modeord ADB – automatisk databehandling – var långt ifrån en realitet. Snarare utvecklade man flödesmodeller och algoritmiska scheman för den mänskliga hanteringen av det omfattande blankettmaterialet. Hundratals kodare anställdes inom ett arrangemang som SCB kort och gott kallade för MDB, manuell databehandling.[27]

Den forskning som jag och mina kollegor utfört på Humlab på Umeå universitet under de senaste åren baserade på storskaliga textuella dataset, har också *alltid* handlat om manuell databehandling. All digitalisering, datakurering och forskningsarbete med dataset innebär handpåläggning, det kan inte nog understrykas. Digital humaniora är därför åtminstone ställvis ett mycket praktiskt forskningsfält. Dataset talar aldrig för sig själva, de måste iordningställas och bearbetas, tvättas och kureras – och först därefter vidtar den egentliga forskningen. Detta manuella förarbete ger visserligen forskaren empirisk överblick, men precis som på SCB är manuell databehandling resurskrävande. I ett av de första forskningsprojekt vi arbetade med på Humlab som tog sig an en mer storskalig textuell empiri – samtliga årgångar av *Aftonbladet* mellan 1830 och 1862, ett dataset på 194 miljoner ord – intresserade vi oss för hur nya medier beskrevs och framställdes i dåtidens dagspress. Bland annat ville vi genom samförekomstanalys undersöka hur den elektriska telegrafen introducerades i *Aftonbladet*. En minst sagt bristande OCR-teknik när läggen digitaliserades gjorde emellertid arbetet tungrott. Variationsrikedomen av hur OCR-motorn tolkat stavningen av den elektriska telegrafen var betydande: tele{raf, tele«craf, telbgral, j7elegraf, ulegraf, fesegraf – och hundratals andra kombinationer. Vi identifierade faktiskt fler än 2 000 digitalt feltolkade varianter av ordet telegraf, och 600 varianter av ordet elektrisk i det dataset vi arbetade med.[28]

Felsökning och korrigering, manuell redigering och kurering av dataset kan bidra till ökad kännedom om ett empiriskt material, det kan

27. Johan Fredrikzon, *Kretslopp av data: Miljö, befolkning, förvaltning och den tidiga digitaliseringens kulturtekniker* (Lund 2021) s. 13, 188–189.

28. En konferenspresentation kallade vi stilenligt: Att hantera felegrafen: Textanalys av smutsiga tidningar. Forskningsprojektet Digitala lägg (2014–2016) finansierades av Torsten Söderbergs Stiftelse, min kollega Johan Jarlbrink utförde det mesta av det praktiska arbetet. För en vidare diskussion, se Johan Jarlbrink & Pelle Snickars, "Cultural heritage as digital noise: nineteenth century newspapers in the digital archive", *Journal of Documentation* 73:6 (2017), samt Johan Jarlbrink, "Telegrafen på distans: Ett digitalt metodexperiment", *Scandia* 84:1 (2018).

rentav leda till nya insikter. Men en något bitter erfarenhet av detta forskningsprojekt var att det (då) inte var möjligt att på ett någorlunda effektivt sätt genomföra storskalig textanalys av svensk dagspress från 1800-talet. En annan var att träffsäkerheten i Kungliga bibliotekets tidningsdatabas Svenska dagstidningar hade sina begränsningar, speciellt för äldre dagspress – det gäller även i dag. De dåtida tidningsläggens utseende (med ibland svagt tryck på färgat papper av dålig kvalitet) samt grafiska layout (med många flera kolumner än i dag) gjorde att OCR-tolkning och segmentering blev bristfällig – inte ens efter omfattande manuell datahantering blev resultaten tillförlitliga. Projektet genomfördes nu i dialog med Kungliga biblioteket, som ansvarar för digitaliseringen av svensk dagspress, vilken praktiskt utförs vid Riksarkivets digitiseringscentrum i Fränsta. Kungliga biblioteket och Riksarkivet har sedan dess förfinat de praktiker, metoder och algoritmer som används för digitalisering av dagspress, men från ett forskningsperspektiv kan det noteras att den allmänna beskaffenheten hos ett empiriskt textmaterial villkorar, ja rentav dikterar om det överhuvudtaget är möjligt att omvandla det till ett maskinläsbart dataset som uppfyller åtminstone vissa grundläggande kriterier för storskalig analys.[29] Handskrivna dokument är som bekant svåra att konvertera till dataset, tryckstilen fraktur likaså.

Vad som kan utgöra historievetenskaplig empiri i datadriven forskning är med andra ord ofta en påtagligt materiell fråga, en paradox om något. Givetvis är det specifika eller mer generella forskningsfrågor som bör driva akademiskt arbete, men jag ska inte sticka under stol med att när det gäller mitt och mina kollegors arbete med storskaliga textuella dataset så har andra faktorer varit nästan lika betydelsefulla – som exempelvis textmaterialets fysiska och grafiska kvalitet liksom upphovsrättsliga begränsningar. Data måste ju gå att exportera till en utvecklarmiljö bortom de arkiv eller bibliotek där digitaliserings utförts. En anledning till att vi arbetade med äldre lägg av *Aftonbladet* var att de tillhörde den kulturella allmänningen och som dataset därigenom fritt kunde skickas till Humlab från Kungliga biblioteket. Att SOU:er varit populära som

29. Äldre svensk dagspress finns numera tillgänglig som dataset via Språkbanken – Kubhist 1 innehåller tidningar perioden 1750–1950 och Kubhist 2 perioden 1645–1926. OCR-kvaliteten är dock fortsatt skral och tillsammans med forskare vid Språkbanken arbetar Kungliga biblioteket för att förbättra den. För en vidare diskussion, se Dana Dannéll, "The Kubhist corpus of Swedish newspapers", <https://spraakbanken.gu.se/blogg/index.php/2019/09/15/the-kubhist-corpus-of-swedish-newspapers> (1/4 2022).

empiri (både enskilda utredningar och som dataset) handlar om samma form av enkla access. Eftersom vi inom vår forskningsgrupp var intresserade av att utveckla våra metodologiska kunskaper, men insåg att äldre dagspress var för bökig att bearbeta, så påbörjade vi ett arbete med just statliga utredningar som dataset – med goda resultat.[30]

I ett nytt forskningsprojekt inriktade vi oss därför på ett snarlikt textmaterial, Tekniska museets årsbok *Daedalus*, en publikation med ett modernistiskt utseende men med klart och tydligt svart tryck på vita boksidor (för optimal bildfångst). I projektet Digitala modeller – som handlade om att digitalisera kulturarv på olika sätt och bedriva forskning på resulterande dataset – digitaliserades *Daedalus* (en tidskrift som Tekniska museet har rättigheterna till). Vi kurerade den till ett dataset om cirka tre miljoner ord och utförde i princip den forskning som inte varit möjlig att genomföra på äldre dagspress.[31] En fråga vi intresserade oss för var hur årsboksvolymerna förhöll sig till och beskrev teknikens historia, en annan var övergripande teknikhistoriska mönster som framträdde i *Daedalus* som dataset. Samtidigt blev det uppenbart att det var svårt att säga något mer generellt om svensk teknikhistoria – det var ju främst möjligt att uttala sig om synen på denna såsom den framträdde i årsboken. En generell iakttagelse av arbete med historiska dataset är därför att forskningen *också* blir begränsad. Dels i bemärkelsen att det inte alltid är så enkelt att dra mer generella slutsatser (då dataunderlaget är för litet), dels i mer praktiskt avseende eftersom det ofta är ett enda, specifikt dataset som utgör empiriskt underlag. Ibland kan historiska dataset vara så omfattande att de har mer generell relevans – som brittiska *the Hansard*, statliga utredningar eller riksdagsdata – men när de är av mindre omfång blir utfall och resultat av körningar lätt kringskurna. Att kombinera källor, vilket ju är en självklarhet inom historievetenskaplig forskning, är förstås inte på något sätt förbjudet inom DH-forskning av

30. Digitaliseringen av statliga offentliga utredningar är ett Kungliga bibliotekets allra mest lyckade digitaliseringsprojekt, ett arbete som avslutades 2015. En första, tentativ analys av utredningarna som dataset återfinns i Fredrik Norén & Pelle Snickars, "Distant reading the history of Swedish film politics in 4500 governmental SOU reports", *Journal of Scandinavian Cinema* 7:2 (2017).

31. Forskningsprojektet "Digitala modeller. Teknikhistoriens samlingar, digital humaniora & industrialismens berättelser" var ett samarbete mellan Tekniska museet och Umeå universitet 2016–2019, finansierat av Riksbankens jubileumsfond och Kungl. Vitterhetsakademien. För mer information, se Jenny Attemark-Gillgren & Pelle Snickars (red.), *Digitala modeller: Teknikhistoria och digitaliseringens specificitet* (Lund 2019).

denna art, men metodologiskt uppstår ofta problem när empiri i olika skalor jämförs. En erfarenhet från de forskningsprojekt vi genomfört är att publicerade resultat blir som bäst om man lyckas att kombinera analys av större mönster i ett dataset med de enskilda utsagor som ingår, det vill säga att zooma in och ut ur datan (men att likväl hålla sig inom samma dataset). Det ger också en antydan om att nästan all forskning som ägnar sig åt att undersöka storskaliga textuella dataset som historisk empiri också närläser den på mer traditionellt historievetenskapligt manér. Men det är samtidigt en forskningspraktik som förutsätter att den enskilda empirin (årsböckerna, utredningarna eller protokollen) är digitalt tillgängliga på ett enkelt sätt, det vill säga att den infrastruktur som nästan all DH-forskning är beroende av inte bara inkluderar dataset som ska undersökas utan också de enskilda utsagorna som de består av.

Att ett nog så omfattande empiriskt forskningsmaterial jämväl sätter upp gränser för vad som kan undersökas, samt reglerar och villkorar de slutsatser som kan dras, är nu inte konstigt. Källmaterial har ju en tendens att styra den forskning som bedrivs, det gäller även digital historieforskning, och att den senare intresserar sig för empirins beskaffenhet (som dataset) är knappast förvånande. Ändå var en central utgångspunkt för vårt nuvarande forskningsprojekt, ”Välfärdsstaten analyserad”, att vi inte enbart skulle arbeta med ett dataset – utan flera. När projektet formulerades var tidigare forskningserfarenheter från Humlab därför centrala: vi skulle fokusera på historiska dataset som var omfattande (långt mer ordrika än alla volymer av en årsbok), vi skulle digitalisera material eller återanvända existerande dataset där OCR och segmentering inte ställde till för mycket problem, och vi skulle bedriva forskning på dataset från olika samhällsområden (under samma period) för att undersöka relationer mellan dem. Att projektet tidsmässigt kom att fokusera på svensk efterkrigstid och välfärdsåren handlade därför till inte så ringa del om att textuellt källmaterial från denna tid lämpar sig väl för att transformeras till dataset, dessutom var en del empiri redan digitaliserad (dagspress, utredningar och riksdagsdata). Projektet handlade i så måtto om forskning liksom kurering av existerande dataset samt digitalisering av annan empiri (romaner och tidskriften BLM). En grundläggande tanke för projektet var också att bedriva digital forskning, kurering och digitalisering parallellt – i samspel med varandra –

detta eftersom massdigitalisering tidigare ofta utförts utan insikter och krav från forskare som arbetar med dataset och digitala metoder. I projektet digitaliseras och kureras därför texter i syfte att möta forskningskrav, samtidigt som projektets forskningsdel ger kontinuerlig feedback tillbaka till de instanser som digitaliserar och iordningställer texterna som dataset. Projektet har ett tätt samarbete med Kungliga bibliotekets datalabb.

I jämförelse med internationell digital humaniora och forskning kring storskalig textanalys så utmärker sig vårt projekt genom att vi valt att sammanställa flera olika dataset för storskalig komparation. Upplägget är ambitiöst och noterbart är att själva utforskandet är starkt beroende av arbetsinsatser som ibland ligger bortom projektledarens kontroll. I skrivande stund håller Kungliga biblioteket exempelvis på att digitalisera all svensk skönlitteratur från 1950-talet, vi tror att det rör sig om 1 246 romaner. Vad handlade egentligen denna skönlitteratur om? Innehöll den kanhända underliggande teman kring emancipation, individualisering eller globalisering – vilka återkom eller speglades i dagspressen – och kanhända plockades upp i snarlika diskurser i riksdagsdebatter eller i utredningar? Det är sådana, mer övergripande frågeställningar på makronivå som vi är intresserade av att undersöka med hjälp av datorstödda metoder. Samma typ av temamodellering som vi utfört på dataset av de statliga utredningarna tänker vi alltså använda även på skönlitteratur och dagspress.

"I read books", har litteraturprofessorn Moretti påpekat, och man får väl tro honom – "but when I work in the Literary Lab they're not the basis of my work. Corpora are [...] When we work on 200,000 novels instead of 200, we are not doing the same thing, 1,000 times bigger; we are doing a different thing. The new scale changes our relationship to our object, and in fact *it changes the object itself*".[32] Inom projektet "Välfärdsstaten analyserad" finns medel för en så kallad datakurerare på halvtid under hela projekttiden – en omfattande budgetpost, och ett resultat av tidigare forskningserfarenheter. Det är vederbörandes uppgift att (i analogi till Morettis uttalande) skapa och kurera de dataset som vår forskning ska bedrivas på. En konstitutiv utgångspunkt för historievetenskapligt forskningsarbete kring storskaliga dataset är att dessa (ofta) är ett resul-

32. Franco Moretti, "Patterns and Interpretation", *Pamphlets of the Stanford Literary Lab* 15 (2017), <https://litlab.stanford.edu/LiteraryLabPamphlet15.pdf> (1/4 2022).

tat av forskningsprocessen i sig. Att en större skala, som Moretti påpekar, förändrar forskarens relation till empiri är en sak, en annan är hur empiri omvandlas till data – eller förhåller det sig tvärtom? I vissa fall är det utifrån en specifik empiri som ett dataset konstrueras, exempelvis hundratals nummer av BLM (vilka vi sammanfört till ett enda dataset). I andra fall existerar redan empirin som läsbar data – det gäller exempelvis alla statliga utredningar under de senaste 20 åren liksom alla former av webbarkiv – och i åter andra fall, till exempel beträffande empiri från data.riksdagen.se så finns redan dataset som jämväl (ofta) måste iordningställas ytterligare för att möjliggöra analys.

Om termer som data och information ibland används utan åtskillnad, så är frågan vad som utgör empiri och vad som är data heller inte entydig. Ibland talas det om empiriska data eller om empirisk datainsamling. Historievetenskapen är av tradition empiriskt orienterad, men hur empiri (som dataset) hanteras och omsätts i digital-historiska analyser är något diffust. För vilket är egentligen det empiriska underlaget för historisk textanalys i en Jupyter Lab-miljö? Ett dataset (med miljoner ord) som består av iordningställd, inskannad textuell empiri, men som kan modelleras på olika sätt beroende på vilken algoritm eller metod som används. Den temamodellering som beskrivits ovan har exempelvis av vissa digitala historiker betecknats som en sorts kodbaserad, teknologisk förmedlingsinstans mellan forskaren och de dataset som undersöks, ”an intermediary between the researcher and archival thruth”.[33] Om empirisk kunskap enligt ordboken bygger på erfarenhet, fakta eller experiment — och inte bara på teoretiskt tänkande – så är den i digital analyskontext beroende av mjukvara.

Under alla förhållanden bör det understrykas att sammanställande av storskaliga dataset förändrar ett empiriskt underlag – själva analysobjektet blir till i forskningsprocessen. Att iordningställa ett dataset innebär å den ena sidan att snarlik empiri (utredningar, protokoll, tidskriftsnummer) sammanförs till en helhet (som inte tidigare existerat), å den andra sidan att somligt textuellt innehåll elimineras (för ökad analyserbarhet). Kurering av textuell data inbegriper dessutom ofta att metadata tillförs: ordklasser taggas och OCR förbättras genom automatiseringsskript, där exempelvis alla ”örn” i riksdagsprotokollen görs till ”om” (även om det

33. Jo Guldi, ”Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora”, *Journal of Cultural Analytics* 1:2 (2018).

kan leda till problem när det gäller diskussioner om rovfåglar). Segmentering (att dela upp en text i ord eller ordliknande enheter) är också en vanlig kureringsuppgift. I riksdagsprotokollen har det exempelvis varit möjligt att använda frasen, ”Herr Talman” (som nästan alltid inleder anföranden) som ett sätt att dela upp textmassan i protokollen. Som påpekats har denna datakurering förlitat sig på maskininlärning för att på automatisk väg förbättra datasetets kvalité (som kan beräknas). Förbättrad analyserbarhet förutsätter med andra ord en acceptabel empirisk datakvalitet.

Men all datakurering handlar också om en sorts grovstädning av text. Att lemmatisera termer och begrepp (sammanföring av olika böjningsformer av ord), samt att helt sonika radera somligt innehåll. I fallet med det dataset av *Daedalus* vi sammanställde togs exempelvis all reklam ur årsboken bort, liksom medlemsförteckningar, listor på museets bidragsgivare, interna rapporter samt referenslistor. Stoppord filtrerades också, och i den temamodellering som genomfördes på årsboken som dataset ingick enbart substantiv och namn, vilket är en vanlig strategi för just denna digitala metod. Den historievetenskapliga empiri som ligger till grund för datadriven digital historieforskning är med andra ord långt ifrån identisk med den som återfinns i arkiven.

Med en metaforik hämtad från miljöområdet handlar iordningställande av storskaliga textuella dataset som historisk empiri om ett slags kretsloppstänkande. All algoritmisk maskininlärning baserar sig på *feedback* och rekursion, men även lagring och distribution är dynamiska. På Github – en plattform för lagring av mjukvara – kan dataset sparas i olika versioner. Det dataset som vi sammanställt av riksdagsprotokollen lagras där i olika versioner, ett versionshanteringsverktyg håller reda på ändringar och textuella förbättringar över tid. Det finns med andra ord inte ett dataset av en viss empiri – utan många. För poängen med att lansera och öppet erbjuda ett dataset är i regel att andra utvecklare kan bidra till förbättringar. I den här typen av forskningssammanhang är distribution av dataset därför ett viktigt forskningsresultat – så även i vårt fall. På projektsidan för ”Välfärdstaten analyserad” på Github kan man därför läsa att vårt dataset av riksdagens protokoll finns tillgängligt i ett visst format (Parla-clarin), att det innehåller alla riksdagsledamöter (det rör sig om tusentals), samt att all kurering finns dokumenterad, ”traceable logs of all curation and segmentation as a git

history [...] If you find any errors, it is possible to submit corrections to them".[34] En fråga som vi brottas med inom vårt projekt är hur en sådan förbättrad empiri (som dataset) ska komma andra historiker och kulturarsvinstitutioner till del. I det senare fallet är ingen lösning i sikte. ABM-sektorn befattar sig ju sällan med denna typ av forskningsdata. På KB-labb finns exempelvis ännu inte någon uttalad praxis för hur forskningsbearbetade dataset (baserade på material i nationalbibliotekets samlingar) ska erbjudas andra forskare. Frågan är under beredning, och under tiden är en lösning att använda Github, en annan att nyttja Svensk nationell datatjänst för att dela historiska forskningsdata.

Avslutning

Att historievetenskaplig empiri måste iordningställas till dataset för att möjliggöra storskalig textanalys är knappast förvånande, inte heller att ett sådant arbete kan vara mödosamt och beroende av omfattande programmeringsinsatser. Vad som möjligen kan överraska är hur komplicerat förhållandet mellan data och empiri ofta är i en digital-historisk analyskontext, liksom att iordningställande av dataset i regel är en iterativ procedur som förändrar källmaterial i grunden. I den här typen av forskningssammanhang kommer ett empiriskt underlag i regel att resultera i flera olika dataset – som för det första är beroende av hur effektiv tillgänglig mjukvara är att kurera och rätta till felaktigheter, där uppgraderad programvara förstås ständigt tillkommer. För det andra beror iordningställande av dataset på vilka specifika forskningsfrågor som ska besvaras – givet att datamodellering kan utföras på väldigt många olika sätt. Den teoretiska relationen mellan empiri och data (som kurerade dataset) är därför synnerligen komplex, och i allra högsta grad beroende av såväl mjukvara som forskningspraktik.

En annan invändning – ja, somliga skulle kalla det för en besvärande omständighet – är att vissa empiriska materialkategorier inte alls är avsedda för att betraktas och sammanföras till ett dataset. Att transformera textuell empiri till storskaliga dataset är därför en process som omdanar somliga materialkategorier mer än andra. Riksdagsprotokoll är med förlov sagt en mer enhetlig kategori än samtliga romaner som publicerades under 1950-talet. För att få korn på vad socialdemokratiska

34. Github, Welfare State Analytics, "Swedish parliamentary proceedings", <https://github.com/welfare-state-analytics/riksdagen-corpus-old> (1/4 2022).

kvinnor ansåg om en viss fråga i riksdagsdebatter under efterkrigstiden så kan första och andra kammarens drygt 12 000 protokoll som dataset vara en lämplig historisk källa. Riksdagsdata är naturligtvis en ytterst heterogen empiri, men den kan likafullt betraktas som ett slags homogent källmaterial: där behandlas vad som yttrades i riksdagens bägge kamrar. Detsamma gäller i viss mån statliga utredningar – en sorts statens röst om man så vill – och kanske i någon mån dagspress, åtminstone om det exempelvis gäller förekomsten av politiska *bigrams* i *Aftonbladet* och *Dagens Nyheter* under en viss period.

Men för att återknyta till Moretti så utgör skönlitteratur en annan datakategori. Ingen har ju någonsin haft för avsikt att läsa alla romaner under en viss tidsperiod som en enda gemensam text. Att sätta samman de knappt 1 300 svenska romaner från 1950-talet som vi digitaliserat till ett dataset innebär därför att skapa ett för forskningsändamålet konstruerat dataset som aldrig tidigare existerat. Vår empiri framstår här som en sorts datalogisk effekt av postulerade forskningsfrågor. Måhända är det därför illustrativt att mängder av de skönlitterära nationalexemplar vi hämtat upp från Kungliga bibliotekets underjordiska magasin i Humlegården aldrig lånats ut – före digitalisering har de behövt sprättas. Om en term som empiri inom historievetenskapen motsvaras av det som kan observeras i källorna, så handlar det i detta fall om ett bokstavligen uppdiktat dataset.

I den här artikeln har jag resonerat om de möjligheter och tillkortakommanden som digitala forskningspraktiker innebär vid studiet av storskaliga textuella dataset som historisk empiri. Sådana dataset, på ibland flera hundra miljoner ord, utgör en ny sorts empiri som ställer historievetenskaperna inför betydande utmaningar. Det har bland annat framkommit att iordningställande av dataset är ett minst sagt resurskrävande arbete där algoritmisk maskininlärning samsas med manuell kurering av data. Inte sällan omstöper dessa processer empiri till dataset – i olika versioner. Att sammanställa alla svenska romaner eller alla statliga utredningar 1945–89 till en enda text kan måhända förefalla överilat, ja rentav ett slags ahistoriskt tillvägagångssätt för att bedriva forskning. Men utifrån ett mediehistoriskt perspektiv är så inte fallet. Arkivens, bibliotekens och museernas samlingar är alltid först mediala – därefter diskursiva. Nästan all empiri har sin egen mediehistoria vilken i regel ger besked om en rad mediespecifika omständigheter: från dokumenta-

tions- och katalogiseringskriterier, mikrofilmad dagspress och fotografi – som i grunden förändrade hur framför allt museer förtecknade sina samlingar – till dagens digitala dataset. I den forskning jag bedrivit under mer än ett decennium, som bland annat handlat om att historisera dagens digitaliseringsverksamhet, har jag intresserat mig för de diskursiva systemens medialitet inom ABM-sektorn. Jag har återkommande fokuserat på dåtidens mediala villkor, samt hur dessa präglade vad som samlades in och dokumenterades.[35] Digital access utgör här den senaste utvecklingen, och den föränderliga medietransfer som dagens digitalisering ger upphov till är därför inte är någon nyhet. Det finns snarare goda skäl att påminna om att nya mediebruk – som analysverktyg i Jupyter Lab-miljö – gör att synen på det förflutna ständigt förändras. I vilken medieteknisk form som historien kommer oss till mötes spelar roll, det gäller även för storskaliga textuella dataset.

A hundred million words: Reflections on historical research with large-scale textual datasets as empirical evidence

The research project Welfare State Analytics: Text Mining and Modelling Swedish Politics, Media & Culture, 1945–1989 uses probabilistic methods and text-mining models to study three massive textual datasets from Swedish politics, news media, and literary culture. By topic modelling and distant reading a dataset from some 3,100 Swedish Government Official Reports, findings have been made which previous historical scholarship has neglected – or rather, cannot detect because of the limitations of traditional, small-scale examinations of only a few such reports. This article presents some of the project's findings, but concentrates on the practical issues of curating large-scale textual datasets, and thus the possibilities – and shortcomings – of digital history research practices.

Large-scale textual datasets, often containing hundreds of millions of words, are a new type of empirical material that presents the historian with fresh challenges. The preparation of datasets is usually a resource-intensive

35. Pelle Snickars, *Kulturarvets mediehistoria: Dokumentation och representation 1750–1950* (Lund 2020).

task, where algorithmic machine learning is combined with the manual curation of data, a process that compiles the empirical material into datasets (in different versions).

Plainly, historical empirical material must be compiled into datasets to enable large-scale analyses, and such work can be laborious, as it depends on extensive programming efforts; what may come as a surprise is how complicated the relationship between data and empirical material can be in a digital-historical context, and the fact that preparing datasets is usually an iterative procedure that fundamentally changes the historical sources. In this type of research, compiled empirical material will usually result in several datasets, depending not only on how effective the available software is to curate and correct errors but also the specific research questions – given that data can be modelled in many ways. The relationship between empirical material and curated datasets is therefore complex, and highly dependent on both software and research practices.

Keywords: data curation, machine learning, textual datasets, digital history