



LUND UNIVERSITY

Välfärdsstaten (o)analyserad

Snickars, Pelle

Published in:
Digitalisering av kulturarvssamlingar

2025

Document Version:
Förlagets slutgiltiga version

[Link to publication](#)

Citation for published version (APA):
Snickars, P. (2025). Välfärdsstaten (o)analyserad. I A. Näslund, & P. Snickars (Red.), *Digitalisering av kulturarvssamlingar: Om forskningsprogrammet Digarv* (s. 227-265). Makadam förlag.

Total number of authors:
1

Creative Commons License:
CC BY-NC-ND

General rights

Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Välfärdsstaten (o)analyserad

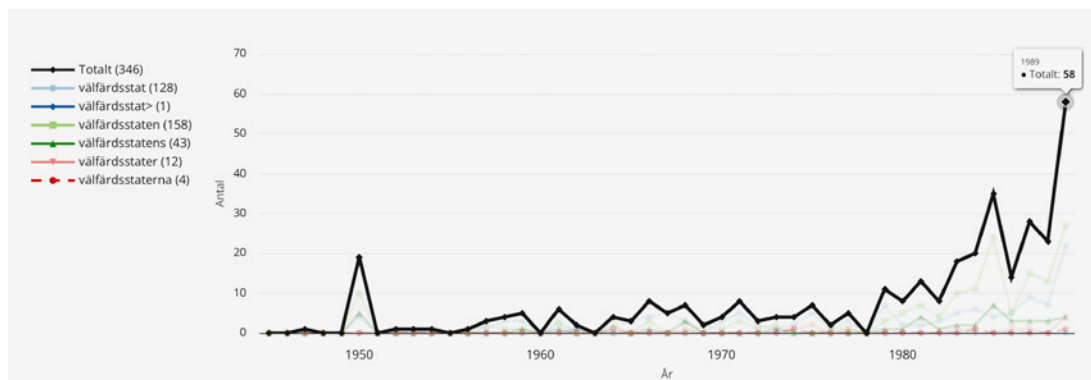
Pelle Snickars

Våren 1950 ägde en av de första större debatterna om välfärdsstatens idé rum i riksdagens andra kammare. Diskussionen för-
anleddes av ett längre utrikespolitiskt inlägg från Östen Undén,
socialdemokratisk utrikesminister sedan andra världskrigets slut.
Även om fem år förflutit sedan det fasansfulla kriget upphörde var
det ännu inte riktigt fred i Europa, tillstod Undén: "Uppdelningen
av Tyskland i en väst-zon och en öst-zon synes för närvarande kon-
solideras genom ockupationsländernas politik." Sverige skulle fort-
satt vara neutralt i det kalla krig som höll på att ta form, men
också visa på ett slags tredje väg, menade utrikesministern, mellan
det kapitalistiska väst och det kommunistiska öst. Det handlade
om att tankemässigt realisera men också praktiskt genomföra idén
om den socialdemokratiska välfärdsstaten. Den skulle förstås
främst gagna landets medborgare, påpekade Undén, men samtidigt
lyfta fram Sverige som ett internationellt föredöme. Ministern var
anspråksfull, och fick snabbt svar på tal. I ett polemiskt genmäle
menade liberalen Bertil Ohlin att det var naivt att tro "att den
svenska s.k. välfärdsstaten skulle påverka uppfattningen i [kommun-
nistiska] länder". Att ett svenskt reformarbete skulle kunna göra
intryck på nationer som numera styrdes av stalinistiska trossatser
var blögt. Högerpartiets ordförande Jarl Hjalmarson kunde inte
annat än instämma: "Den stora meningsbrytningen i världen går
inte mellan kapitalism och socialism utan mellan demokrati och
diktatur." Människor i den demokratiska världen fruktar inte var-
andra, fortsatte Hjalmarson: "Vad de bäva för är risken att tvingas
in under en regim, som berövar dem alla de värden som ge livet
mening för en västerlänning." Med emfas underströk högerledaren

att det därför fanns all ”anledning att varna för förhoppningen att Sverige genom att realisera ’välfärdsstatens idé’ skulle kunna påverka uppfattningen i länder med ’proletariatets diktatur’”.¹

Riksdagsprotokollet från denna debatt, AK 1950:11, finns sedan något decennium tillbaka tillgängligt i digital form på riksdagen.se – både som PDF-dokument och html-länk. Det är också möjligt att göra fritextsökningar på riksdagen.se i alla sådana digitaliserade protokoll, även om kvaliteten på *optical character recognition* (OCR) varierar. När Riksdagsförvaltningen först började digitalisera dokument från enkammarriksdagen (protokoll, motioner, utskottsutlåtanden, propositioner etcetera) så var OCR-tolkningen inte den bästa. I ett senare samarbete med Kungliga biblioteket (KB) digitaliserades också det mesta av det äldsta riksdagstrycket, både från ståndsriksdagen (1521–1866) och från tvåkammarriksdagen (1867–1970), ett omfattande textmaterial som gjordes öppet tillgängligt på webben 2018 – med avsevärt förbättrad OCR-tolkning. Att Undén, Ohlin och Hjalmarson debatterade utrikespolitik och den svenska välfärdsstatens idé våren 1950 har historiker alltså tämligen enkelt kunna söka fram sedan ett antal år tillbaka. Däremot har det varit betydligt svårare att jämföra och studera debatter i riksdagen om ett tema (såsom välfärdsstatens idé) över en längre tidsperiod. Men i den nya webbapplikationen riksdagsdebatter.se kan man göra just det. Söker man på termen *välfärdsstat* så klustrar sig träffarna kring 1927 – men då hade begreppet en annorlunda innebörd: ”den polisstyrda sociala välfärdsstaten” var enligt högermannen Otto Järte en styggelse.² Först i vårdebatten 1950 diskuterades konturerna av det som senare skulle karakterisera detta begrepp, nämligen en nationalstat (med socialdemokratiska förtecken) där det offentliga ansvaret för medborgarnas välfärd var synnerligen omfattande.

Tanken med applikationen riksdagsdebatter.se har varit att göra det lättare att utforska, läsa och ladda ner anföranden av svenska riksdagsledamöter som talat i riksdagen sedan tvåkammarriksdagens införande. Applikationen baserar sig på cirka artontusen digitaliserade riksdagsprotokoll från 1867 till idag. Eftersom ett protokoll av en riksdagsdebatt inte sällan kunde sträcka sig över mer än hundra sidor så handlar det om en mycket omfattande textkorpus. Applikationen har utvecklats vid det digitala humanioracentrat Humlab på Umeå universitet, där en utgångspunkt varit forskningsprojektet *Välfärdsstaten analyserad*. Arbetet handlade i en



• En sökning på termen välfärdsstat* i applikationen riksdagsdebatter.se mellan åren 1945 och 1989 genererar som grafen antyder en markant ökning efter 1980 – vilket möjligen kan förvåna. Applikationen innehåller alla riksdagsprotokoll från tvåkammarriksdagen 1867–1971, samt alla protokoll från den påföljande enkammarriksdagen (fram till idag).

första fas om att tillföra metadata till alla riksdagsprotokoll. Det innebar dels att länka samman vem som yttrade vad (och när) i kamrarnas debatter, dels att addera information om riksdagsledamöter såsom kön, levnadsår och partitillhörighet. Det kan förefalla som en rudimentär uppgift. Men eftersom det suttit tusentals ledamöter i riksdagen, som gärna talat ofta och länge i tiotusentals debatter, så har forskningsprojektet utfört ett ansenligt arbete med att strukturera dessa data. De initialt digitaliserade protokollen från Riksdagsförvaltningen och KB innehåller nämligen bara ostrukturerade data; i dessa kunde man till exempel inte söka fram alla yttranden om välfärdsstaten av Östen Undén, eller av någon annan socialdemokratisk ledamot för den delen. Strukturerade data innebär information som är organiserad på ett specifikt sätt, så att den enkelt kan bearbetas av annan mjukvara. Sådana data är vanligtvis lagrade i tabeller, där varje dataelement följer ett förutbestämt format eller struktur, såsom rader och kolumner i en csv-fil med kommaseparerade värden (i exempelvis Microsoft Excel). I korthet har strukturerade data inneburit att Undéns, Ohlins och Hjalmarsons anföranden i debatten 1950 separerades från varandra så att det tydligt gick att urskilja vem som sagt vad (och i vilken ordning), när yttrandet ägde rum, när det började och slutade, samt vem talaren var.

Det här kapitlet handlar om vårt arbete inom forskningsprojektet *Välfärdsstaten analyserad: Textanalys och modellering av svensk*

politik, media och kultur, 1945–1989. Fokus ligger på samarbetet mellan forskare och kulturarvssektor, mellan akademiker och utvecklare, samt inte minst mellan vetenskapliga ideal och krass forskningspraktik. Tanken är att beskriva vad projektet ägnat sig åt och resonera om de erfarenheter och utmaningar som det gav upphov till, både forsknings- och digitaliseringsmässigt. Som kapitelrubriken antyder har det varit ett både lyckat och delvis misslyckat forskningsprojekt. Och apropå kapitlets anslag var det också ett digitalhistoriskt forskningsprojekt, som ägnat mycket tid och kraft åt att strukturera data, framför allt riksdagsprotokoll. *Välfärdsstaten analyserad* kan bäst beskrivas som ett projekt i skärningspunkten mellan mediehistoria och digital humaniora – med ett tredelat syfte: att digitalisera äldre textmaterial, att kurerat sådan text och omforma den till strukturerade dataset, samt att bedriva forskning på dessa dataset med hjälp av olika slags digitala metoder.

Metodmässigt har projektet arbetat med en rad typer av algoritmisk textanalys, med utgångspunkt i storskalig empiri från politikens sfär (statliga offentliga utredningar och riksdagstryck), dagspress och skönlitteratur. I det senare fallet har alla svenska romaner utgivna mellan åren 1940 till 1965 digitaliserats – cirka tretusen stycken – liksom alla nummer av tidskriften *Bonniers litterära magasin*. All digitalisering har genomförts tillsammans med KB, och projektet har också återkommande samarbetat med nationalbibliotekets datalaboratorium KB-labb. Svensk efterkrigstid är en väl utforskad period, men genom att applicera digitala metoder på kurerade dataset var projektets grundtanke att granska politikens, nyhetsmediernas och kulturens sfärer på nytt. *Välfärdsstaten analyserad* har varit ett av de större projekten inom Digarv-programmet; budgeten uppgick till nästan tjugotre miljoner kronor. Projektet har involverat ett tjugotal medarbetare vid en rad olika lärosäten. Forskarna inom projektet har främst varit medie- och litteraturvetare, men också statistiker med expertis inom maskininlärning och probabilistiska modeller för textanalys. Utvecklare vid Humlab har därtill varit centrala för att ta fram verktyg och applikationer för storskalig textanalys.³

Projektbakgrund

– om två tidigare forskningsprojekt och en rapport

Upprinnelsen till projektet *Välfärdsstaten analyserad* var, som ofta i akademiska sammanhang, andra forskningsprojekt. På ett generellt plan har Digarv-programmet därtill byggt vidare på tidigare utlysningar av projektsamarbeten mellan akademien och kulturarvssektorn. I vårt fall rörde det sig om två tidigare projekt vid Umeå universitet och Humlab, varför det i det här sammanhanget är värt att dröja något vid dessa. En annan utgångspunkt var det arbete som jag själv gjorde för Kungliga bibliotekets räkning kring inrättandet av det som senare kom att kallas för KB-labb. Två tidigare forskningsprojekt och en förstudie av digital humaniora-karaktär bildade därför på olika sätt upptakten till projektansökan för *Välfärdsstaten analyserad*.

I efterhand framstår forskningsprojektet *Digitala lägg* – finansierat av Torsten Söderbergs stiftelse (2014–16) – som ett slags ouverty till *Välfärdsstaten analyserad*, ett mindre förprojekt med fokus på KB:s digitaliserade tidningsarkiv och portalen Svenska tidningar. Ambitionen var då att utforska digitaliserade 1800-talstidningar med hjälp av olika slags digitala metoder för textanalys. *Digitala lägg* intresserade sig framför allt för hur äldre medietekniker behandlades i *Aftonbladet* vid mitten av 1800-talet, där beskrivningen av den elektriska telegrafen utgjorde en fallstudie. En textkorpus sammanställdes med samtliga årgångar av *Aftonbladet* från åren 1830–1862, ett dataset på knappt tvåhundra miljoner ord. Genom samförekomstanalys, det vill säga en metod som urskiljer ord som förekommer tillsammans med varandra, undersökte projektet hur elektrisk telegrafi introducerades i *Aftonbladet*. Det blev emellertid snart tydligt att det inskannade tidningsmaterialet inte höll den kvalitet som de digitala metoderna krävde. För att kunna genomföra åtminstone en del av de textanalyser *Digitala lägg* föresatt sig, tvingades forskarna manuellt städa och försöka mäta andelen fel och brus i de textfiler som digitaliseringen genererat. KB digitaliserade då som nu sitt äldre tidningsbestånd på Riksarkivets mediekonverteringscentrum (MKC) i Fränsta (i Medelpad), men en bristande OCR-teknik gjorde projektarbetet minst sagt tungrovt. Bara variationsrikedomen av hur OCR-motorn tolkat stavningen av den elektriska telegrafen i de gamla tidningsläggen var betydande: tele{raf, tele«craf, telbgral, j7elegraf, ulegraf, fesegraf – och hundratal andra kombinationer. Projektet identifierade inte mindre än

två tusen digitalt feltolkade varianter av ordet telegraf, och sexhundra varianter av ordet elektrisk.⁴

En slutsats av projektet *Digitala lägg* var att (medie)historiker som intresserar sig för äldre dagspress i digital form alltid måste uppmärksamma hur läggen digitaliserats, liksom på vilka sätt som de därefter bearbetats genom exempelvis OCR och textsegmentering. Feltolkade ord var legio i de digitaliserade numren av *Aftonbladet*, vilket ibland berodde på att originalsidor var trasiga eller hade veckat sig. Men ett större problem var att artiklar inte segmenterats på rätt sätt; inte sällan fortsatte strängen med tolkade ord från en tidningsartikel horisontellt tvärs över till nästa artikel, vilket skapade en väldig semantisk oreda. Faktorer som tryckteknik, layout och bristande papperskvalitet försämrade också korrektheten hos den OCR-producerade texten. De tidningsfiler från 1800-talets mitt som gjordes tillgängliga i digital form var helt enkelt inte samma tidningsblad som en gång trycktes, även om Svenska tidningars gränssnitt kunde ge sken av att så var fallet – eller som historikern Tim Hitchcock påpekat: ”the problem is that while we think we are searching newspapers, we are actually searching markedly inaccurate representations of text”.⁵

Ett mer praktiskt resultat av *Digitala lägg* var att äldre svenska dagstidningar i digital form när allt kom omkring inte var speciellt lämpade för storskalig textanalys; de innehöll alltför mycket feltolkade ord och felsegmenterad text. Detta bistra faktum utgjorde sedermera en viktig anledning till varför *Välfärdsstaten analyserad* tog sig an en helt annan tidsperiod. Kungliga bibliotekets digitalisering av 1800-talets press genererade visserligen utmärkta bildfiler för visning i gränssnittet Svenska tidningar, men för dåvarande OCR-program var språkförbistringen betydande – tolkning av tätt tryckt text och alltför många spalter i *Aftonbladet* genererade både inkoherens och ordsallad.⁶

Om *Digitala lägg* utgjorde det ena bakgrundsprojektet som *Välfärdsstaten analyserad* vilade på, så var forskningsprojektet *Digitala modeller* det andra. För att möjliggöra storskalig textanalys övergav min forskargrupp på Humlab äldre dagspress som källmaterial. Tidningar ersattes av bokmediet – närmare bestämt av Tekniska museets årsbok, *Daedalus*. Det var en publikation med klart och tydligt svart tryck på vita boksidor i en modernistisk design, och därför lämpad för optimal bildfångst. *Digitala modeller* var ett teknik- och mediehistoriskt projekt, som i korthet handlade om att

undersöka digitaliseringens möjligheter och effekter för både akademisk forskning och mer praktiskt arbete vid en svensk minnesinstitution (Tekniska museet). Projektet genomförde därför tre olika slags digitaliseringsinsatser: traditionell digitalisering av material ur en bergshistorisk samling på museet, massdigitalisering av årsboken *Daedalus*, och 3D-digitalisering av ett trettiotal gamla trämodeller.⁷ I det här sammanhanget var digitaliseringen av *Daedalus* av störst intresse; den gjorde det möjligt att strukturera all text till ett dataset om cirka tre miljoner ord, och all OCR-tolkning var föredömlig. Till skillnad från *Aftonbladet* var det med denna källa möjligt att genomföra olika typer av storskalig textanalys, bland annat historiografisk temamodellering av hela tidsskriften.⁸ Eftersom *Digitala modeller* var ett samarbete mellan ett universitet i Västerbotten och ett museum i huvudstaden var det också av vikt att ledningen på Tekniska museet gav klartecken till att öppet publicera alla årtiotvå årgångar av *Daedalus* på webben (där de fortfarande finns kvar idag).⁹ På så vis blev det möjligt för forskarna på Humlab att (på distans) både kvantitativt modellera alla textdata och kvalitativt läsa i de inskannade årsböckerna.

En erfarenhet från *Digitala modeller* var att forskningen blev som bäst när förutsättningarna fanns för att kombinera analys av större mönster i ett dataset med alla enskilda utsagor, det vill säga möjligheten att zooma in i och ut ur data, men att likväl hålla sig inom samma dataset. Det ger också en antydning om att nästan all forskning som ägnar sig åt att undersöka storskaliga textuella dataset som historisk empiri också närläser dessa textdata på mer traditionellt historievetenskapligt manér. Men det är samtidigt en forskningspraktik som förutsätter att den enskilda empirin – i det här fallet årsböckerna av *Daedalus* – är digitalt tillgänglig på ett enkelt sätt. Denna infrastrukturella forskningspraktik (i det lilla) var en viktig förutsättning för *Välfärdsstaten analyserad*, där det inte enbart handlade om att analysera ett dataset, utan nästan lika mycket om att undersöka enskilda utsagor.

En fråga som *Digitala modeller* intresserade sig för var hur olika volymer av *Daedalus* förhöll sig till och beskrev teknikens historia, en annan hur mer övergripande teknikhistoriska mönster framträdde i årsboken som sammansatt dataset. Samtidigt blev det uppenbart för oss forskare att det var svårt att säga något mer generellt om svensk teknikhistoria; det var egentligen bara möjligt att uttala sig om synen på denna såsom den framträdde i årsboken.

Om beslutet att arbeta med mer modernt textmaterial (som *Daedalus*) utgjorde en förutsättning och bakgrund för tillkomsten av *Välfärdsstaten analyserad*, så var en annan att omfattningen av det textmaterial som skulle undersökas behövde vara mycket större än en museiårsbok på sammanlagt femtontusen sidor. Den textkorpus som *Daedalus* genererade var behändig att arbeta med, men det var också ett historiskt dataset som begränsade forskningen, främst i bemärkelsen att det var svårt att dra mer generella slutsatser – vårt enda dataset som empiri var helt enkelt för småskaligt. Ibland kan sådana enskilda historiska dataset naturligtvis vara så omfattande att de har mer generell relevans – till exempel *the Hansard*, med transkriptioner av alla debatter i det brittiska över- och underhuset från 1803 till idag, ett dataset på flera hundra miljoner ord.¹⁰ Likväl: en forskningserfarenhet från *Digitala modeller* var att när dataset hade ett alltför litet omfång, ja då blev utfall och resultat av körningar ofta kringskurna.

Om forskningsprojekten *Digitala lägg* och *Digitala modeller* på flera sätt bildade upptakten till *Välfärdsstaten analyserad*, var en tredje utgångspunkt den tankemässiga förändring som delar av bibliotekssektorn i Sverige genomgick vid mitten av 2010-talet. KB har alltid haft till uppgift att tillhandahålla en effektiv forskningsinfrastruktur, men nu började denna ställvis ändra karaktär. "Forskningen efterfrågar idag allt oftare tillgång till stora datamängder, snarare än enskilda digitala objekt", kunde biträdande riksbibliotekarie Lars Ilshammar hävda. "Det innebär att vi bevittnar ett paradigmskifte i synen på vad den digitala utmaningen egentligen består i."¹¹ Trenden var internationell och delvis en effekt av framväxten av forskningsfältet kring digital humaniora. KB hade vid mitten av 2010-talet viss erfarenhet av ett sådant nymornat forskarintresse; när digitaliseringen av fler än sextusen statliga offentliga utredningar var klar 2015 anordnades till exempel ett SOU-hack på nationalbiblioteket, där kod och dataset därefter tillgängliggjordes på utvecklarsajten Github. På Humlab i Umeå började vi därför också att intressera oss för detta utredningsmaterial med statliga betänkanden – som framgent även skulle ingå som empiriskt underlag i *Välfärdsstaten analyserad*.¹²

Det var under den tid då jag själv arbetade som forskningschef på KB (åren kring 2010) som alla statliga offentliga utredningar mellan 1922 och 1999 började digitaliseras. Jag kan tyvärr inte ta åt mig äran av detta arbete; idé och initiativ kom från min

dåvarande kollega Agneta Holmenmark. Det skulle bli en av nationalbibliotekets allra mest lyckade digitaliseringsinsatser, inte minst eftersom SOU:erna – liksom riksdagstrycket – till sin natur var offentliga och därför inte inskränktes av några upphovsrättsliga hinder. Precis som i fallet med *Daedalus* var det därför möjligt att undersöka alla betänkanden som både enskilda utsagor och sammansatt dataset. Samtidigt innebar ett sådant textuellt sammanställande av alla utredningar till en enda korpus att det empiriska underlaget ändrade karaktär. All form av tillgängliggörande av storskaliga kulturella eller textuella dataset förändrar alltid empirin; själva analysobjektet blir till i forskningsprocessen. Ur ett enskilt forskningsperspektiv var sådana transformationer hanterbara, värre var det för ett nationalbibliotek. Var det verkligen KB:s uppgift att tillhandahålla större datamängder av digitaliserat kulturarv?

I efterhand framstår det som framsynt att biblioteksledningen på KB ansåg att så faktiskt var fallet. En bidragande orsak var att flera andra europeiska nationalbibliotek, liksom en rad amerikanska universitetsbibliotek och Library of Congress, då hade börjat inrätta bibliotekslaboratorier för att understödja forskning på storskaliga dataset – ”collections as data”, med historikern Ryan Cordells (senare) terminologi.¹³ Med hjälp av ett anslag från Riksbankens Jubileumsfond hösten 2017 gjorde jag året därpå en utredning för KB:s räkning kring möjligheten att etablera ett sådant bibliotekslabb. Arbetet var tämligen omfattande och inbegrep en rad studieresor till snarlika datalabb vid British Library i London och vid det nederländska nationalbiblioteket i Haag. I den rapport som sedermera presenterades för ledningen på KB gjorde jag gällande att det nya datalabbet borde ha fem övergripande mål: 1) att stödja forskning inom digital humaniora genom nya tillämpningar och metoder; 2) att kurera och sammanställa dataset; 3) att aktivt delta i forskningsansökningar med datalabbet som partner; 4) att stödja KB:s interna digitala omställning med datalabbet som innovationsnav; 5) att stödja digital kunskapsförsörjning bland nationalbibliotekets medarbetare.¹⁴ Eftersom KB:s digitala samlingar blev alltmer efterfrågade inom datadriven forskning i humaniora och samhällsvetenskap inrättades KB-labb tämligen omgående året därpå. Syftet var till en början (från 2019) att understödja forskning på stora datamängder genom metodstöd, beräkningsresurser och tillgång till KB:s samlingar i form av dataset – det vill säga ungefär vad jag föreslagit i min rapport. En föreståndare rekryterades (Love

Börjesson), och 2021, efter att externt ha utvärderats av docent Mats Fridlund vid Centrum för digital humaniora vid Göteborgs universitet, blev KB-labb en permanent del av nationalbibliotekets verksamhet. ”Utvärderingen bekräftar att KB-labb har blivit en framgång och svarat mot ett ökat behov av forskning på stora datamängder. Vi har också sett att labbet bidrar till samhällsnytta inom både offentlig och privat sektor”, påpekade riksbibliotekarie Karin Grönvall nöjt i ett pressmeddelande.¹⁵

Digitalisering och reglering – på KB

Välfärdsstaten analyserad har på flera sätt varit ett slags distribuerat forskningsprojekt. Till en början utgjorde Umeå universitet huvudnod, men gradvis (när projektmedlemmar fick nya tjänster) fördelades projektet på allt fler lärosäten från Uppsala, till Lund och Malmö. Dessutom var delar av projektet hela tiden förlagt på KB i Stockholm. Om forskningsverksamheten utgick från fyra olika universitet inkluderade projektet också ytterligare tre distinkta verksamhetsnoder: 1) utvecklare på Humlab; 2) digitalisering på KB; 3) datakurering på KB-labb. Den här geografiska spridningen och uppdelningen på olika noder innebar under projekttiden en hel del utmaningar, inte minst beträffande forskningsmaterial som av upphovsrättsliga skäl inte fick distribueras utanför KB. Samtidigt drog *Välfärdsstaten analyserad* nytta av att KB-labb under projekttiden parallellt startade upp sin verksamhet. Eftersom jag själv varit inblandad i etableringen av KB-labb föll det sig naturligt att det projekt jag hade hand om (tillsammans med ett annat) uttryckligen prioriterades av KB under datalaboratoriets uppstartsfas. Inom *Välfärdsstaten analyserad* var KB därför en central aktör, dels genom den digitalisering som organiserades där – med en projektfinansierad koordinator – dels genom det kureringsarbete som utfördes på KB-labb. I det senare fallet hade projektet även en fördel av att ett annat snarlikt forskningsprojekt också prioriterades av KB: *Mining for meaning: Den offentliga migrationsdiskursens dynamik*, ett projekt inom analytisk sociologi vid Linköpings universitet. I det projektet ingick docent Måns Magnusson som statistiker och maskininlärningsexpert, och han hade i princip samma roll i *Välfärdsstaten analyserad*. Magnusson blev därför en nyckelperson för det forskningsarbete som utfördes på KB-labb. De två forskningsprojekten hade också en annan gemensam nämnare eftersom

de bågge intresserade sig för storskalig analys av svensk dagspress under efterkrigstiden.¹⁶

Den huvudsakliga anledningen till att vårt projekt valde att fokusera på svensk efterkrigstid och rekordårens decennier mellan 1945 och 1989 hade med det empiriska materialets beskaffenhet att göra. Det kan förefalla underligt att låta empirin styra forskningen, men ovanstående redogörelse förklarar ungefär hur vi tänkte. Upplägget var heller inte märkligare än den historievetenskapliga tradition som sedan länge handlat om att upparbeta ett speciellt arkiv, eller en specifik samling av texter (eller annan avgränsad empiri). När jag själv till exempel var postdok i Berlin under 2004 handlade min forskningsuppgift om att gå igenom huvudstadens dagspress kring 1910 för att få korn på hur tidningsmediet då behandlade tidig film. Även den empirin var tydligt avgränsad.

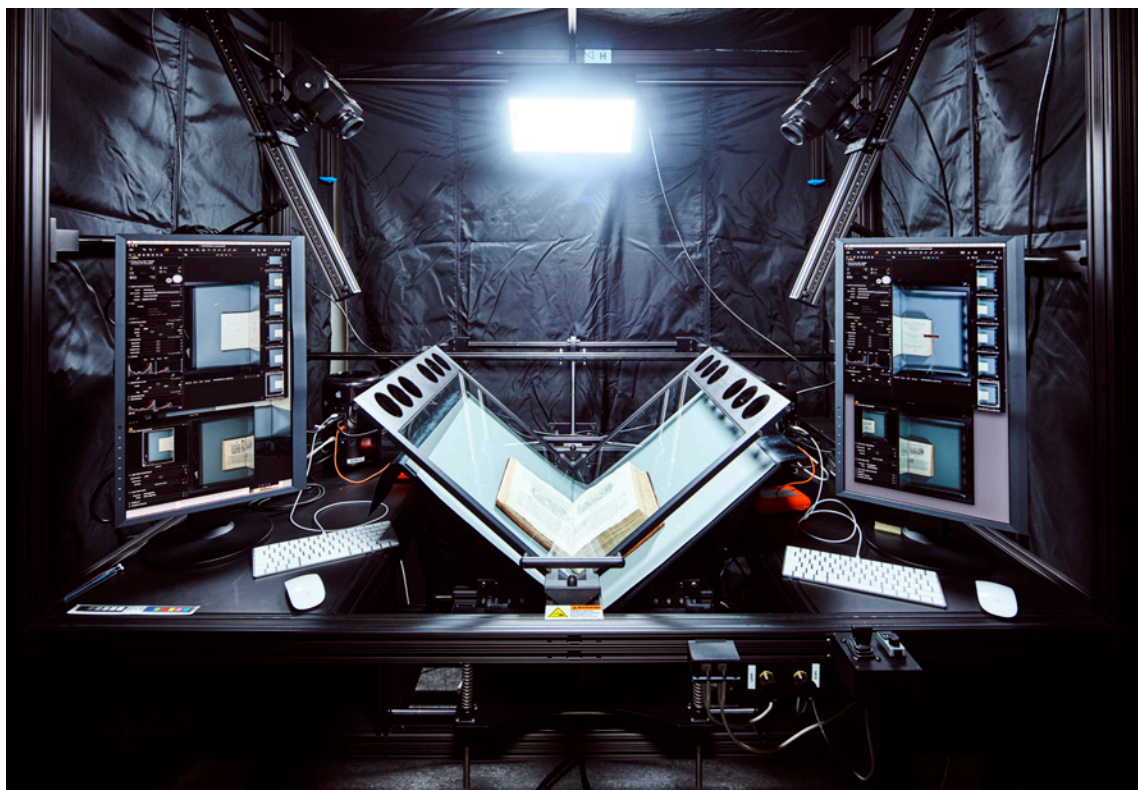
Men att *Välfärdsstaten analyserad* inte skulle arbeta med äldre dagspress var vi på det klara med; tidigare forskningserfarenheter avskräckte. Vår tanke var att ett modernare pressmaterial inte skulle innehålla lika mycket OCR-fel, textbrus och andra defekter orsakade av själva digitaliseringen. KB hade också skannat ett betydande bestånd av större svenska tidningar från projektets tänkta tidsperiod, som *Expressen*, *Dagens Nyheter*, *Svenska Dagbladet* och *Aftonbladet*. En annan viktig anledning till valet av tidsspänn hade att göra med de statliga offentliga utredningar som min forskargrupp tidigare framgångsrikt bearbetat på Humlab med olika slags digitala metoder. Detta textmaterial var redan sammanställt till ett större dataset, och det hade också visat sig väl lämpat för storskalig analys. Det svenska utredningsväsendet (i denna form) påbörjades dock först 1922. Det visade sig emellertid att antalet utredningar 1945–1989 var mycket stort, fler än tretusenettihundra betänkan- den publicerades då. När vi på Humlab sammanställde alla betänkan- den under denna period till ett dataset omfattade det nästan nittio miljoner ord.

Om de svenska välfärdsåren utgjorde en period av förändring, med utbyggda institutioner, nya levnadsmönster och förändrade relationer till omvärlden, så var grundtanken med vårt projekt att utifrån ett omfattande textmaterial utforska denna period med metoder som tillät analys av diskursiva förändringar på makronivå. De textsamlingar som projektet valde ut representerade tre centra- la sfärer i det svenska välfärdssamhället: nyhetsmedier, politik och kultur. I projektets mediesfär ingick dagspress, och i den politiska

sfären offentliga utredningar. Till den senare kategorin föll det sig naturligt att också addera annat offentligt tryck, närmare bestämt riksdagstryck – inte minst eftersom Riksdagsförvaltningen ungefär vid samma tidpunkt, när vi drog upp riktlinjerna för vårt projekt, just slutfört digitaliseringen av (nästan allt) riksdagstryck, som KB sedan också tillgängliggjorde. Vår forskningsansökan drog med andra ord nytta av – och byggde vidare på – tidigare digitaliseringsprojekt inom arkiv, bibliotek och museer.

När det gällde den tredje så kallade kultursfären var vi till en början osäkra på vilket textmaterial den skulle innehålla. Men eftersom både dagspress och riksdagstryck redan var tillgängliga i digital form föll valet på att digitalisera en helt ny materialkategori. För att representera kultursfären tog *Välfärdsstaten analyserad* sig därför an att digitalisera alla svenska romaner utgivna under projektets tidsspänn, samt periodens mest emblematiske kulturtidskrift, *Bonniers litterära magasin*. En poäng var att projektet på så vis kom att innefatta både kurering och strukturering av tidigare digitaliserat kulturarv, samt massdigitalisering av ett helt nytt textmaterial. Naturligtvis var vi medvetna om att Litteraturbanken vid Göteborgs universitet redan då digitaliserat flera tusen skönlitterära verk, bland annat i samarbete med KB – och då nästan enbart äldre publikationer inom den kulturella allmänningen, öppet tillgängliga för alla. Vi kände förstås också till att merparten av de romaner vi skulle digitalisera – liksom stora delar av BLM – fortfarande hade upphovsrättsliga begränsningar. En ambition från min sida var samtidigt att låta *Välfärdsstaten analyserad* utgöra startskott för massdigitalisering av tryck på Kungliga biblioteket, en fråga jag återkommer till i slutet av detta kapitel.

Sammantaget var en utgångspunkt för vårt projekt att forskning, digitalisering och kurering av data skulle löpa parallellt – och gärna i samspel med varandra. I projektet digitaliserades och kurerades texter i syfte att möta forskningskrav, samtidigt som projektets empiriska forskningsdel gav kontinuerlig återkoppling tillbaka till de instanser som iordningställde texterna som dataset. Så var tanken i teorin; i praktiken blev förfarandet ofta annorlunda. Beträffande de dataset som vi tidigare arbetat med på Humlab följde vi vår projektplan; själv publicerade jag ett flertal artiklar baserade på det ovan beskrivna datasetet av tretusen betänkanden.¹⁷ Samtidigt stod det snart klart att den tidsperiod vi valt ut inte riktig lämpade sig som avgränsning utifrån ett biblioteks- och samlingsperspektiv.



- I forskningsprojektet *Välfärdsstaten analyserad* digitaliserades 2 981 svenska romaner från 1940 till 1965, samt alla utgåvor av kulturtidskriften *Bonniers litterära magasin*. Mellan åren 1932 och 2004 handlade det om ialles 566 nummer på sammanlagt 48 287 sidor.

BLM publicerades till exempel mellan 1932 och 2004, och det framstod som märkligt att vi inte skulle inkludera alla nummer av tidskriften när den nu äntligen digitaliserades.

Även för uppgiften att kurera och strukturera riksdagsprotokollen – snart i ett initialt samarbete med Riksdagsbiblioteket – framstod projektets tidsperiod som udda. I en första fas valde vi därför att arbeta med riksdagstryck under en hundraårsperiod (1920–2020), för att sedermera förskjuta denna bakåt i tiden till tvåkammarriksdagens införande. Även beträffande de romaner som projektet digitaliserade blev tidsperioden förändrad, men i det fallet handlade det främst om finansiering. Under förarbetet med projektansökan räknade vi med att kunna använda destruktiv skanning och skära upp den så kallade andraupplagan av äldre skönlitteratur på KB, och på så vis pressa digitaliseringspriserna på MKC Fränsta genom destruktiv skanning. Men det visade sig inte vara

praktiskt genomförbart. Dessutom var det svårt att genom Libris och KB:s egen bibliotekskatalog Regina få exakt information om hur många svenska romaner som egentligen publicerats; som påpekats rörde det sig om cirka tretusen för hela perioden 1940 till 1965. Men under 1940- och 1950-talen publicerades det något färre romaner, ungefär trettonhundra per decennium. Om tidsramen försköts en aning bakåt skulle medlen som budgeterats för digitalisering av romaner kunna täcka ett kvarts århundrade. Så blev också fallet – alla svenska romaner mellan 1940 och 1965 blev slutligen digitaliserade.

Budgeten för *Välfärdsstaten analyserad* inkluderade två utgiftsposter på cirka fem miljoner kronor som skulle täcka digitalisering av romaner och BLM, samt en projektkoordinator på KB. Digitaliseringsarbetet kan i efterhand bäst beskrivas som ett slags separat delpjekt, där KB och koordinator Stefanie Topelius på egen hand fattade de flesta beslut. Projektet kunde dra stor nytta av de erfarenheter som KB skaffat sig under mer än ett decennium av att digitalisera dagstidningar i Fränsta, eftersom BLM som tidskrift kunde digitaliseras på ett snarlikt sätt. Precis som med många andra tidningar visade det sig dessutom att KB hade dubblettexemplar av BLM i sin depå i Bålsta (norr om Stockholm). Dubbletterna var i gott skick och behövde inte några bevarandeåtgärder; de kunde därför användas för destruktiv skanning (med gynnsammare pris). För att registrera all metadata om BLM användes databasen Signe som KB nyttjar för att hantera och registrera de dagstidningar som skickas till MKC för digitalisering. Men eftersom BLM var en tidskrift behövde Signe uppdateras med viss insats från utvecklare. Noterbart är därtill att i de fall då dubletter för BLM saknades var KB tvunget att använda sig av det så kallade nationalexemplaret, vilket av hävd kräver ett mer varsamt hanterande.

Som tidskrift var BLM en materialkategori som KB tidigare endast digitaliserat sparsamt. *Välfärdsstaten analyserad* kom därför att fungera som testbädd för kommande tidskriftsdigitalisering på nationalbiblioteket; en tid senare digitaliserades till exempel tidskriften *Moderna tider* (1990–2002) med utgångspunkt i erfarenheterna från arbetet med BLM. Dessa tidskrifter gjorde också att portalen med svensk presshistoria bytte namn, från Svenska dagstidningar till Svenska tidningar. I en delrapport från KB hösten 2020 påpekades att vårt projekt etablerat en produktionslinje för digitalisering av tidskrifter med extern leverantör (MKC), inklusive

framtagning av metadataspecifikation samt övriga komponenter för beställning och utförande. Produktionslinan innefattade därtill leverans av filer från MKC till KB:s digitala arkiv Mimer, samt tillhandahållande av BLM i portalen Svenska tidningar (tidningar.kb.se) samt på KB-labb för projektets räkning.¹⁸

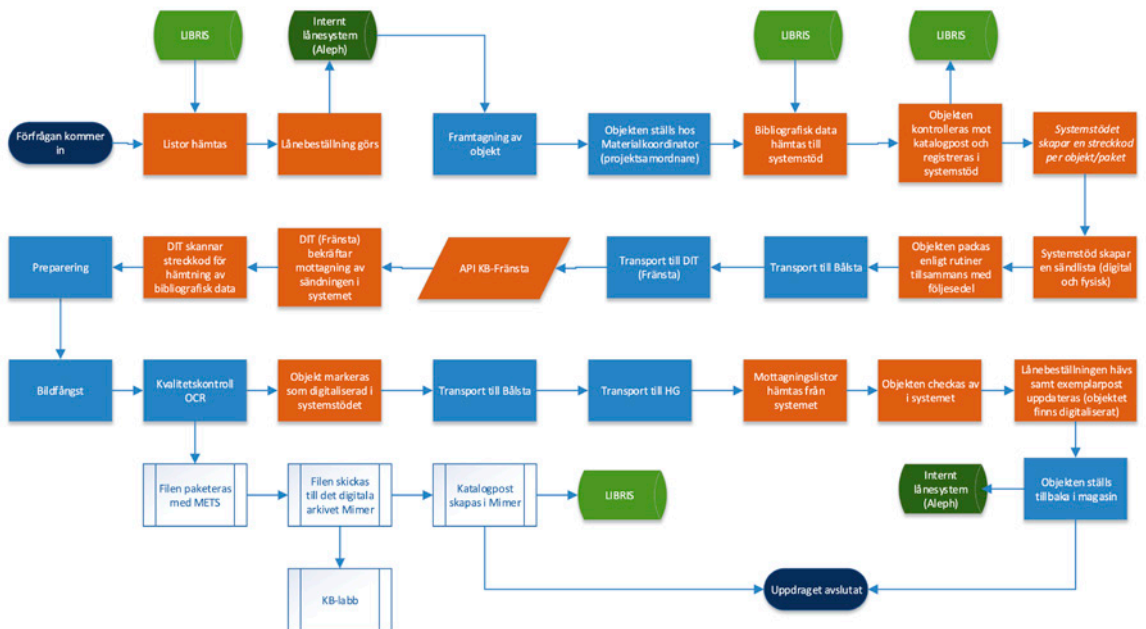
När det gällde digitaliseringen av de svenska romanerna behövde KB inte på samma sätt ta fram nya riktlinjer och metadataspecifikationer; nationalbiblioteket hade under en längre tid skannat skönlitterära böcker i samarbete med Litteraturbanken. Den sistnämnda har dock främst ägnat sig åt äldre litterärt tryck, med fokus på kanoniserade författare. *Välfärdsstaten analyserad* däremot inbegrep all publicerad skönlitteratur – där nationalexemplar i regel användes för massdigitalisering. En hel del av de böcker som skannades var för övrigt aldrig någonsin frambeställda; arken fick sprättas innan böckerna digitaliserades. Den allra bökigaste frågan var emellertid att katalogmässigt fastställa exakt vilken skönlitteratur – främst romaner men också några novellsamlingar – som egentligen publicerats och pliktlevererats till KB under projektets tidsperiod. Även om översatt skönlitteratur var exkluderad, liksom barn- och ungdomslitteratur, saknade KB erfarenhet av att ta fram korrekta listor över det egna beståndet. Man kan tycka att ett nationalbibliotek borde veta precis vad som ingår i samlingarna, men det var alltså inte riktigt fallet. I så måtto utgjorde bokdigitaliseringen inom *Välfärdsstaten analyserad* en sorts generalprov för KB, med bäring bortom projektramen, där själva urvalsprocessen för vad som skulle digitaliseras krävde en betydande insats. Det är i detta sammanhang värt att citera Stefanie Topelius slutrapport från 2022 *in extenso* eftersom den antyder både de problem KB stod inför och de betydande resurser som nationalbiblioteket ansamlade för att få till stånd digitalisering av tretusen romaner.

Kungliga bibliotekets samlingar har under tidsperioder katalogiserats på olika sätt. Det finns i dagsläget [två] databaser som man kan utgå från för att få fram en lista över vilka objekt som finns i samlingarna under en definierad tidsperiod. Det är dels Libris XL (som är en nationell samkatalog för ett stort antal svenska bibliotek), samt Regina, vilken är KB:s egen beståndsdatabas. Dessa två speglar inte varandra fullt ut då ett objekt kan finnas beskrivet i en av dem men inte i den andra. [Utgår man från] en lista hämtad från Libris XL så är det inte säkert att det fysiska objektet faktiskt finns i KB:s samlingar då katalogiseringen kan ha skett hos en annan institution. Objektet kan också

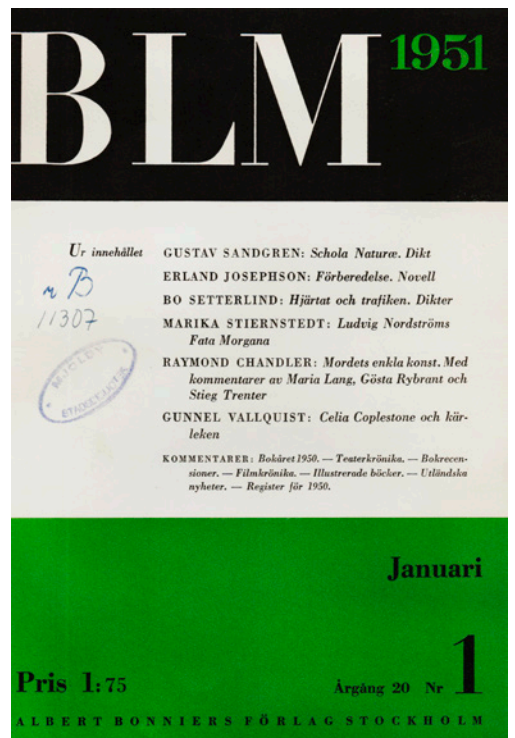
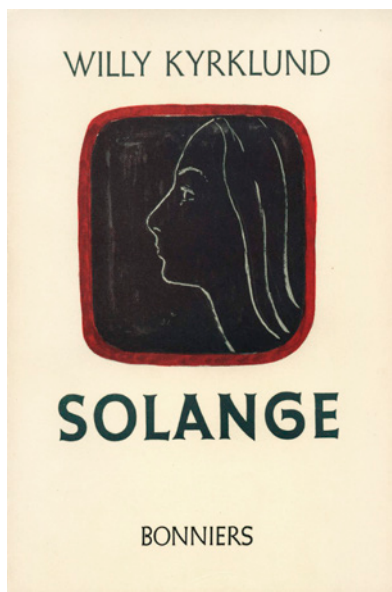
ha förkommit eller inte alls blivit levererat som pliktexemplar – i Libris XL finns inte heller uppgifter om var i KB:s samlingar som objektet finns placerat. I Regina däremot finns KB:s samlingar beskrivna med ett signum som berättar var objektet [är] uppställt i våra magasin. Men även här kan det hända att objekt saknas (förkommit eller blivit felplacerad), objekten har också under vissa perioder katalogiserats på olika sätt [...] Allt detta sammantaget innebar att när vi väl beslutat oss för att utgå från listor hämtade i Libris XL, så krävdes en fortsatt insats att rensa dessa från dubletter, felaktiga katalogposter, översättningar och/eller barn- och ungdomslitteratur som inte blivit katalogiserat som detta. Listorna konverterades till Excelfiler. Nästa steg blev att stämma av vad som faktiskt fanns i våra samlingar. När objektet lokaliserats och tagits fram så kontrollerades det ännu en gång mot katalogposten och placerades på en specifik för projektet avsatt plats i våra magasin. Där tog [KB:s enhet] för bevarande över och kontrollerade samt dokumenterade status för varje volym i bevarandedatabasen Fenix. Här ingick även ett omfattande arbete med att sprätta de volymer som inte tidigare lånats ut samt att mjuka upp ryggar. Därefter packades materialet i för projektet inköpta vagnar och med objekten numrerade efter de Excelfiler som man utgått från. Vagnarnas hyllplan fotades efter packning som en del av transportdokumentationen [och böckerna skickades därefter till MKC].¹⁹

Som projektledare för *Välfärdsstaten analyserad* var jag imponerad av den kraftsamling som KB:s personal gjorde för att digitalisera BLM och alla svenska romaner – från katalogisatörers metadata-specifikationer till praktiska insatser av bokbindare med de fysiska böckerna och utvecklare med skript för filleverans.²⁰ De medel som projektet bidrog med för detta digitaliseringsarbete täckte inte på långa vägar den egeninsats som KB bidrog med. Det var förstås något som nationalbiblioteket också insåg. Faktum är att projektet ledde till en översyn av hur KB skulle beräkna intern resurs- och tidsåtgång för digitaliseringsprojekt med extern finansiering. Numera finns det flera sådana mallar framtagna på KB, till exempel för nationalbibliotekets egeninsats vid digitalisering av monografier, i tolv distinkta steg: förteckning, framtagning, bevarandeåtgärder, katalogkontroll, registrering i databas, packning och mottagning av böcker, följt av projektkoordinering, administration, IT-stöd, digitalt bevarande, samt juridik och avtal.

Om KB var synnerligen tillmötesgående när det gällde insatsen att digitalisera det tryck som projektet efterfrågade, var nationalbiblioteket mindre välvilligt inställt till utökad forskningsaccess på



- Att digitalisera BLM eller svenska romaner kan framstå som enkelt, men grafiken för det interna systemstödet på KB för digitalisering av svenskt tryck antyder att så inte är fallet. Illustrationen återfinns i slutrapporten KB–WeStAc: En rapport om digitisering på KB (2022).



distans. Eftersom *Välfärdsstaten analyserad* ägnade sig åt en tidsperiod i svensk nutidshistoria begränsades tillgänglighet till BLM och de svenska romanerna av existerande upphovsrätt (sjuttio år efter författarens död). De senare kunde studeras i digital form på KB-labb, och BLM även genom portalen Svenska tidningar, men då bara via de sjutton universitetsbibliotek och de fjorton stadsbibliotek som (mot en licensavgift) erbjuder full åtkomst till databasen via en specifik terminal. Då *Välfärdsstaten analyserad* var ett nationellt distribuerat forskningsprojekt spritt över stora delar av landet innebar detta avsevärda problem. I mitt arbete med den rapport som låg till grund för etableringen av KB-labb hade jag fört ingående diskussioner med det datalabb som inrättats i Haag (som lätt förvirrande också heter KB LAB). För forskningsprojekt som arbetade med ett specifikt dataset iordningställde nederländska KB LAB alltid ett kontrakt som gav forskare access till alla textdata, både digitaliserade textfiler och som sammansatt dataset, under själva projektperioden. Min förhoppning som projektledare var därför att samma slags tillgång till data skulle kunna gälla för *Välfärdsstaten analyserad* – det vill säga, inte nödvändigtvis access till BLM och alla romaner som digitaliserade PDF-filer i original, utan mer som extraherade txt-filer (endast text utan layout) som vi inom projektet därefter på Humlab skulle iordningställa till olika dataset. Om ett sådant förfarande gick att ordna i Nederländerna, resonerade jag, så borde det också vara möjligt i ett snarligt EU-land som Sverige. Att KB internt prioriterade vårt forskningsprojekt borgade också för möjligheten med en sådan lösning.

Till saken hörde att jag som koordinator för Digarv-programmet, via våra programmöten, kände till att ett forskningsprojekt (mot en avgift) av KB tilldelats utökad access till portalen Svenska tidningar via sina datorer på lärosätet. Upphovsrätten på tidningsområdet var heller inte fastlagd, utan snarare en sorts pågående förhandlingsfråga. På grund av implementering av olika EU-direktiv och tolkningen av dessa nya bestämmelser var svensk upphovsrätt helt enkelt föränderlig; den hade förskjutits ett antal gånger, och är ännu inte helt klarlagd. KB hade följaktligen under det senaste decenniet steg för steg så sakteliga gjort allt fler äldre tidningslägg tillgängliga – från en initial spärr satt till 1864, därpå förskjuten framåt till 1901, till den hundraårsgräns som numera råder (år 2025). Det var förstås utmärkt från ett forskningsperspektiv. Samtidigt tillstod de jurister jag konsulterade att KB:s tolkning

av upphovsrätten ofta varit konservativ; juridik är ju sällan svart eller vit. Det senaste policybeslutet som KB fattat, att i tidningsgränssnittet öppet tillhandahålla tidningar ”fram till 1924 enligt en rullande modell där material successivt blir tillgängligt”, hade alltså ingen egentlig upphovsrättslig grund.²¹ Det var snarare en sorts rådande kompromiss – ett legalt gungfly om man så vill, öppet för successiva förändringar och förhandlingar. Dessutom tillkom vid denna tidpunkt, åren kring 2020, de undantag i upphovsrätten, med möjlighet att utföra text- och datautvinning på upphovsrättsligt skyddat material, som implementeringen av EU:s allra senaste upphovsrättsdirektiv skapade.

Ovanstående juridiska omständigheter gjorde att vi forskare inom *Välfärdsstaten analyserad* var förhoppningsfulla när det gällde utökad access till det textmaterial som projektet digitaliserat (och bekostat). Men dessvärre stötte vi på patrull. Trots ett om inte osäkert, så åtminstone omdiskuterat rättsläge, valde KB:s ledning att avslå vår förfrågan om utökad access, och detta alltså i ett forskningsprojekt som till en inte ringa del omfattade nationalbibliotekets egen verksamhet och bekostade delar av densamma. Efter många turer och förfrågningar var biträdande riksbibliotekarie Lars Ilshammar tvungen (får man förmoda) att meddela att vårt projekt inte på distans kunde få tillgång till de filer som digitaliseringen på MKC givit upphov till, varken i original som PDF-dokument eller som txt-filer. Det spelade ingen roll att forskare och utvecklare vid Umeå universitet befann sig mer än sexhundra kilometer från huvudstaden. Det var heller inget tal om att skriva något kontrakt; fastän projektet skulle lägga tre miljoner kronor på att digitalisera tretusen romaner så kunde dessa endast studeras på KB-labb på Karlavägen i Stockholm. Till BLM gick det visserligen att få access via respektive universitetsbibliotek, men däremot inte via Humlab – trots att terminalen med dagspress på Umeå universitetsbibliotek låg nästan vägg i vägg.

Här ska inte raljeras över denna juridiska aktsamhet och defensiva hållning; KB har trots allt i andra upphovsrättsliga sammanhang och förhandlingar med organisationen Bonus Copyright Access gjort betydande insatser för utökad forskningsaccess, framför allt när det gäller tillgång till audiovisuellt material på distans. Men det måste också tillstå att bristen på tillgång till materialkategorier inom den kultursfär som *Välfärdsstaten analyserad* vilade på, gjorde att projektets grundidé delvis sköts i sank. Denna begränsning

gjorde helt enkelt att vi tappade styrfart inom de delar av projektet där forskare skulle ta sig an svenska romaner och BLM. Eftersom vår litteraturvetare i projektet, Alexandra Borg, inte fick tillgång vare sig till de högkvalitativt digitaliserade filerna av BLM eller tidskriften i txt-format – ja, då valde hon istället att privat köpa alla årgångar i fysiskt format.

Forskning och kurering – på Humlab och KB-labb

I oktober 2019 anordnade Vetenskapsrådet (VR), Riksbankens Jubileumsfond (RJ) och Vitterhetsakademien en workshop apropå den kommande (andra) Digarv-utlysningen av forskningsmedel, även den med utgångspunkt i regeringens uppdrag att stödja data-driven forskning inom området digitalisering och tillgängliggörande av kulturarvssamlingar för forskningsändamål. Det var på ett sätt den första sammankomsten för Digarv-programmet, och det första mötet i den serie med externa föreläsningar och interna programmöten som vi koordinatörer framgent skulle organisera, som vi påpekat i denna boks inledning. Det var också en workshop där de första resultaten från *Välfärdsstaten analyserad* presenterades. Inom projektet hade vi då påbörjat ett arbete med storskalig textanalys av dagspress mellan 1945 och 1989. Med utgångspunkt i *Dagens Nyheter* och *Aftonbladet* intresserade sig några av projektets forskare för vad som blev politiskt under tidsperioden ifråga. Metoden handlade bland annat om att ur båda tidningarna extrahera textblock där termerna politisk, politiskt eller politiska förekom – via olika långa ordfönster – och analysera dessa både kvantitativt och kvalitativt. Förekomsten av så kallade *bigrams* – såsom ”politisk medvetenhet” etcetera – uppvisade en minst sagt brokig variation av vad som ansågs politiskt i dags- och kvällspress under svensk efterkrigstid och välfärdsår. Det handlade dels om förväntade samförekomster med begrepp såsom politisk film, politisk teater eller politisk indoktrinering, dels om mer oväntade resultat i tidningsläggen: politisk fotbollsmatch, politisk gudstjänst, eller politisk ärtsoppa. Det kunde vidare konstateras att antalet sådana politiska samförekomster steg markant i både *Dagens Nyheter* och *Aftonbladet* från och med slutet av 1960-talet, för att sedan tydligt klinga av åren kring 1980.

I arbetet med att extrahera textblock ur tidningarna på KB-labb hade vårt projekt stor hjälp av doktorand Miriam Hurtado Bodell,

verksam i systerprojektet *Mining for meaning* (som också finansierades av VR). Tillsammans med KB-labb togs fyrahundratusen sådana textblock fram. Det gjordes också en initial överenskommelse om att skicka vissa resultat från KB-labb till Humlab, för vidare textanalys. Denna uppgörelse i det lilla föregicks emellertid av ett tämligen omständligt dividerande kring huruvida det var möjligt att från (eller ur) sådana ordfrekvenser – av olika slags extraherad text ur dagspress – återskapa tidningsartiklar i originalform (för vilka det förelåg upphovsrätt). En lösning var att textblocken gjordes om till alfabetiska ordlistor istället för orden i originalföljd. Här var KB skeptiska eftersom de trodde att ordlistorna skulle gå att återställa till ordföljden i textblocken (som då ånyo skulle likna originalartiklar med upphovsrätt). En annan variant var att skicka dessa textblock som en sorts ordpåsar i bokstavsordning, så att utvecklarna på Humlab kunde skapa temamodeller av dem. Men även här förhöll sig KB avvaktande eftersom det då, åtminstone teoretiskt, skulle vara möjligt att sätta ihop dessa ordpåsar till textblock igen (vilka potentiellt närmade sig artiklar i originalform med upphovsrätt). Sammanfattningsvis önskade sig KB både hängslen och livrem – en juridisk hållning som gjorde forskningsarbetet tungrovt (milt talat). Vid denna tidpunkt hade KB-labb samtidigt etablerat ett arbetsrum för externa forskare i nationalbibliotekets lokaler på Karlavägen i Stockholm. Det var till gagn för både *Välfärdsstaten analyserad* och projektet *Mining for meaning*. Men det stadfäste samtidigt att allt storskaligt forskningsarbete och all kvantitativ utforskning av KB:s digitala samlingar måste utföras på plats i huvudstaden. I fallet med *Välfärdsstaten analyserad* hade projektet i så måtto att göra med två fysiska datalaboratorier – KB-labb och Humlab – där det initialt inte alltid var helt friktionsfritt att få utvecklare att samarbeta med varandra.

Det största hindret var dock tillgång på distans till det digitaliserade mediematerialet. Eftersom *Välfärdsstaten analyserad* också var ett mediehistoriskt projekt, var vi som humanister förstas intresserade av att även närläsa tidningsläggen för att närmare kontextualisera hur dagspressen exempelvis resonerade kring ”politisk ärtsoppa” – vilket, visade det sig, inte var en vänsterföreteelse, utan istället förekom i en tidningsannons för Folkpartiet 1967, liksom i pressinbjudningar tjugo år senare från Tåbymoderaterna (norr om huvudstaden). Via portalen Svenska tidningar gick det att få fram dessa uppgifter, men bara genom att uppsöka den dataterminal

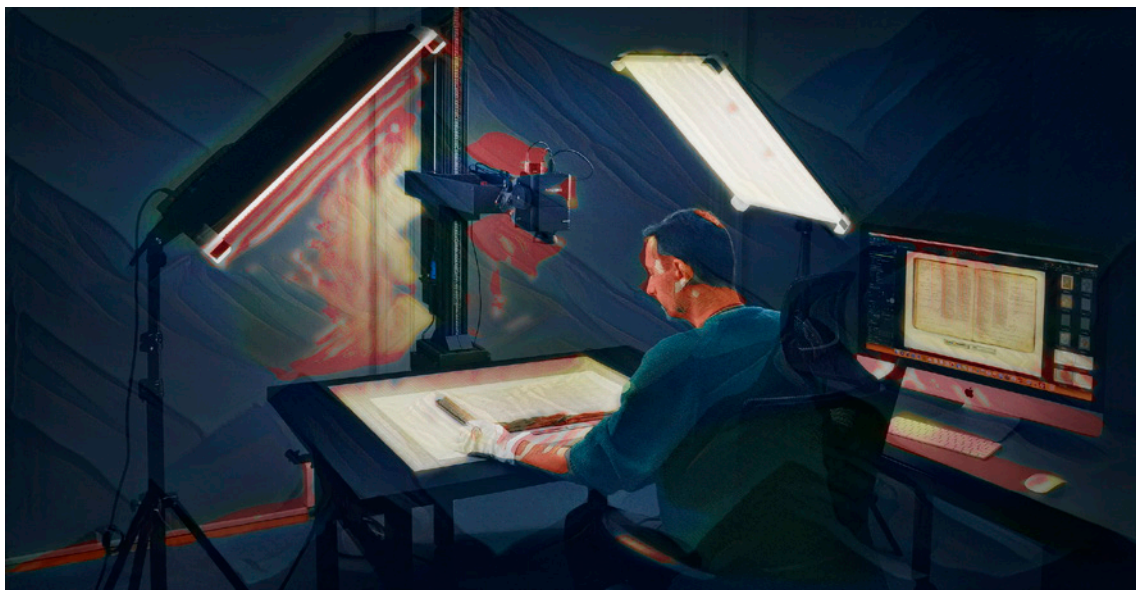
som stod placerad i källarplanet på Umeå universitetsbibliotek. För att få tillgång till *Dagens Nyheters* tidningsarkiv kunde vi runda denna omständighet genom att använda oss av en privatfinansierad prenumeration på DN (med access till tidningens historiska pressarkiv från 1864 och framåt). I övrigt fick forskarna Johan Jarlbrink och Fredrik Norén (som utförde det mesta av detta forskningsarbete) masa sig från sina kontor i Humanisthuset på Umeå universitet, gå ut i kylan, passera Lindellhallen och ta trapporna ned till dagspressterminalen i bibliotekets undervåning – i akt och mening att kontrollera och få fram originalartiklar ur de digitaliserade tidningsläggen.²²

Det var ett vacklande infrastrukturellt ramverk som få andra vetenskapsområden skulle acceptera. Det omständliga förfarandet och bökret med att skicka ordfrekvenser, samt omöjligheten att på den egna datorskärmen kunna växla mellan storskalig analys av textblock och de enskilda utsagorna i tidningsarkivet – där de senare alltså bokstavligen måste inhämtas i utprintat format efter en promenad – gjorde forskningen mödosam. Covidpandemin bidrog också med sitt. Men vad som i efterhand framstår som mest bekymmersamt var att vi mediehistoriker i projektet egentligen allra mest hade önskat att arbeta med just dagspress, om nu inte forskningslogistiken blivit så svårbemästrad. Och om vi i det tidigare projektet på Humlab, *Digitala lägg*, hade tillgång till alla gamla nummer av *Aftonbladet*, men då inte kunde genomföra storskalig analys som önskat på grund av semantiska felaktigheter efter digitalisering – ja, då var situationen nu den omvända: den datadrivna analysen gick utmärkt att få till stånd, men forskningsarbetet haltade likafullt eftersom vi på distans inte fick access till de enskilda tidningsläggen och de artiklar som omnämndes. Med lite god vilja från KB – eller mot en avgift som i ett annat Digarv-projekt²³ – hade denna besvärande situation kanske kunnat lösas, men så blev det inte.

Vid denna tidpunkt under projektets första verksamhetsår tillkom ytterligare en försvårande omständighet: den datakurerare som i *Välfärdsstaten analyserad* var placerad (och skulle anställas) på KB-labb kom inte på plats. Tjänsten som datakurerare på halvtid under fem projektår på KB-labb var tillsammans med budgetposten för en utvecklare på Humlab (på samma procentsats under lika lång tid) det allra mest påkostade i hela projektet. Det rörde sig om nästan två miljoner kronor var som budgeterats för dessa bägge tjänster. Efter förhandling och samråd med KB-labbs föreståndare,

Love Börjesson, beslutades dock att dessa medel skulle föras tillbaka till Umeå universitet; KB lämnade alltså ifrån sig en betydande summa pengar. För projektets vidkommande innebar det att vi förlorade en medarbetare som skulle ha anställts internt på KB-labb och på så vis förankrat *Välfärdsstaten analyserad* på nationalbiblioteket. Tanken var förstås att en sådan medarbetare i KB-labbs arbetsmiljö skulle förenkla framtagandet av dataset ur de textuella sfärer som projektet intresserade sig för. När detta inte fungerade fick vi tänka om. Att anställa en datakurerare på Umeå universitet som skulle arbeta halvtid i Stockholm var naturligtvis omöjligt. Istället omkonterades dessa medel till den statistiska institutionen på Uppsala universitet, där Måns Magnusson förtjänstfullt tog på sig ansvaret för denna arbetsuppgift – och framgent delegerade den till två doktorander och fyra masterstudenter vilka periodvis anställdes som amanuenser.

När projektansökan för *Välfärdsstaten analyserad* skrevs hade KB-labb ännu inte påbörjat sin verksamhet. Men i den rapport jag publicerade 2018, *Datalabb på KB*, var just framtagning och kurering av dataset en av de mest centrala uppgifterna för en sådan kommande verksamhet. Det föll sig därför naturligt att inkludera en snarlik tjänstebeskrivning i vår projektansökan. Funktionen var viktig – men ingen kan se in i framtiden. För vad som inträffade under KB-labbs uppstartsfas var att verksamheten snabbt ändrade karaktär. Föreståndare Börjesson tydde tecknen i tiden, och lade om kursen för sitt labb. Om jag i min rapport 2018 hävdade att kurering av kulturella och textuella dataset var av vikt, så inträffade nästan samtidigt det nära nog blixtsnabba tekniksprång vi alla nyss genomlevt – numera betecknat som AI-boomen. Inom denna blev KB-labb en viktig nationell aktör genom att publicera en rad språkmodeller. Sådana modeller hade långt större verkansgrad i samhället än kurerade dataset. Modellerna för språkförståelse, såsom KB-BERT, utgår från omfattande artificiella neurala nätverk vilka tränats på stora mängder text ur nationalbibliotekets samlingar. Istället för att iordningställa små dataset med BLM och romaner kom KB-labb att ägna sig åt att publicera flera språkmodeller på Github och Hugging Face. Där hade de snart laddats ned fler än två miljoner gånger, med betydande inverkan på tillämpningen av AI i offentlig sektor i Sverige. Börjesson utsågs till finalist i tävlingen årets AI-svensk 2021, och KB-labb letade sig rentav in i regeringens forsknings- och innovationsproposition 2024:



• Åren kring 2020, då KB-labb och projektet *Välfärdsstaten analyserad* startades upp, blev maskininläring, språkmodeller och generativ AI till vardagliga företeelser. Det var en snabb teknisk samhällsomvandling som förändrade arbetet på KB-labb, från kureri av specifika dataset till utveckling av allmänna språkmodeller. Fotografiet är hämtat från den danska firman Phaseone som arbetar med digitalisering av kulturarv; bilden har modifierats med ett enda klick på ett av Photoshops neurala filter.

”Regeringen ser behov av att ytterligare kartlägga och analysera möjligheter och utmaningar med Kungl. bibliotekets arbete med AI-modeller för forskning samt hur verksamheten kan bidra till att möta behov av AI från även andra relevanta verksamheter i offentlig sektor.”²⁴

Även om verksamheten på KB-labb blev annorlunda än vi förespeglat, lyckades vi inom *Välfärdsstaten analyserad* att framgent etablera ett gott samarbete. Projektet fick exempelvis hjälp med att förbättra OCR-kvaliteten på en hel del av vårt empiriska underlag (såsom vissa riksdagsprotokoll och några enskilda nummer av BLM). Därtill kunde vi återkommande organisera ett flertal workshops på KB-labb. Som ersättning för den uteblivna tjänsten som datakurerare kom projektet som sagt att anställa statistikdoktorander och masterstudenter vilka arbetade någon dag i veckan på KB-labb, för att understödja projektets övriga forskare. Arbetet med dagspressen sinkades dock; dels var det fortsatt besvärligt att

```
[2]: load_gui = ntm.LoadGUI(* ntm.LoadGUI(data_folder=data_folder, state=current_state(), slim=True).setup()
display(load_gui.layout())
```

[illegible][illegible][illegible][illegible]

- på distans få tillgång till enskilda artiklar, dels visade det sig problematiskt att strukturera vissa tidningsdata. Bland annat intresserade vi oss för att arbeta storskaligt med endast ledartexter, men det visade sig vara ett alltför tidsödande arbete att särskilja sådana artiklar från övrig tidningstext. Och eftersom digitaliseringen på MKC av BLM och alla romaner tog mer tid i anspråk än beräknat – digitalisering pågick under hela tre år, från 2019 till 2022 – kunde projektet inte heller riktigt få styrfart på analysen av detta material. En bidragande orsak var dessutom att de fyra masterstudenterna i statistik som hjälpte till med att strukturera och kurerat data råkade besitta just den typ av kompetens som marknaden då alltmest efterfrågade. Följden blev att de efter några månaders projektanställning gick vidare till långt mer välbetalda arbeten som data scientists på

bland annat Statistiska centralbyrån och AI-bolaget CourtCorrect i London.

Genom doktorand Hannes Waldetofts arbete på KB-labb lyckades projektet emellertid strukturera vissa metadata kring de digitaliserade romanerna. Via en överenskommelse med KB kunde därtill ordfrekvenser exporteras till Humlab, vilket möjliggjorde ett initialt storskaligt forskningsarbete också på detta textmaterial. Fastän jag själv var forskare i vårt projekt måste jag tillstå att den arbetsinsats som KB lade ner på att digitalisera dessa romaner vida översteg det tämligen ringa forskarintresse vi ägnade resultatet – även om det också kom att utgöra grund för ytterligare en forskningsansökan (som jag återkommer till). Under alla förhållanden sammanställdes en temamodell med tvåhundra teman med utgångspunkt i korpusen av de tretusen romanerna. I det webbaserade gränssnittet Jupyter Lab var det möjligt att studera bland annat hur sådana teman relaterade till varandra, vilka ord ur romanerna som förekom i olika teman och hur dessa teman varierade över tid mellan 1940 och 1965.

Tyvärr var det av upphovsrättsliga skäl inte möjligt att inkludera de skönlitterära texterna som txt-filer i detta gränssnitt, trots att Jupyter Lab hade den tekniska kapaciteten för detta. En ytterligare besvärande omständighet var att ingen i vårt projekt hade tillgång till de digitaliserade romanerna, detta till skillnad från BLM som återfanns i portalen Svenska tidningar. De tretusen svenska romanerna existerade bara i filformat på en server på KB-labb, eftersom nationalbiblioteket inte hade något fungerande gränssnitt för svenskt tryck (och inte heller för upphovsrättsskyddat material). Inte ens jag som projektledare hade någon som helst access till dessa filer. När Litteraturbanken till exempel hörde av sig med en förfrågan om att inkludera de författare som vi digitaliserat, där upphovsrätten upphört, så kunde vi därför inte vara till någon hjälp. Sammanfattningsvis genererade dessa juridiska begränsningar ett slags akademisk letargi kring det empiriska material i *Välfärdsstaten analyserad* som forskarna inte kom åt. Men eftersom vi också arbetade med annan empiri flyttades fokus dit istället. Ett praktiskt resultat av projektet var därför både pragmatiskt och fatalt: en sorts upphovsrättslig tvångsstyrning drev obönhörligen all forskning mot de materialkategorier som var öppet tillgängliga och inte hade några rättsliga begränsningar.

Riksdagstryck för alle

I december 2021 anordnades den fjärde nationella konferensen om digital historia vid Umeå universitet. Huvudtalare var historikern Jo Guldi, som i flera sammanhang arbetat med *the Hansard*, den ovan nämnda stora textkorpusen med alla debatter i det brittiska över- och underhuset under mer än tvåhundra år. Jag gav konferensens andra keynote, med fokus på arbetet inom *Välfärdsstaten analyserad*. Vid denna tidpunkt var digitaliseringen av BLM och de svenska romanerna ännu inte klar, därtill hade vi forskare fastnat i arbetet med dagspress på grund av de upphovsrättsliga begränsningar som kringskar analysarbetet. Om Guldi framstod som en fullfjädrad *parliamentary historian* gällde nästan detsamma för flera av medlemmarna i min forskargrupp. De flesta av oss var egentligen mediehistoriker. Men svårigheten att arbeta med de digitaliserade tidningsläggen hade gradvis stöpt om oss till en sorts parlamentshistoriker – på mer eller mindre frivillig basis. Min egen presentation fokuserade därför mest på arbetet med statliga utredningar 1945–1989, där vi använt temamodellering som metod för storskalig textanalys.

För forskningsarbetet med statliga betänkanden hade utvecklare på Humlab skrivit kod och laddat upp temamodeller i utvecklingsmiljön Jupyter Lab, ett arbete som senare också låg till grund för den ovan beskrivna digitaliseringen av romanerna. I det web-baserade gränssnittet fanns kodceller där man som forskare med knappar och manuella reglage kunde styra parametersättningen av koden. Den kunde också enkelt modifieras och visa valfritt antal ord i varje tema, lista dem över tid, eller jämföra teman med varandra. Resultat i Jupyter Lab förevisades som tabeller, diagram eller enklare nätverk. Avgörande för detta forskningsarbete med större dataset var emellertid att alla enskilda betänkanden också enkelt kunde sökas fram genom att använda KB:s gränssnitt Digitaliserade SOU:er 1922–1999, där alla utredningar snabbt kunde laddas ned i PDF-format. Det var precis den typen av forskningspraktik som fattades när det gällde projektets arbete med dagspress.

Eftersom forskningsarbetet med de statliga utredningarna fungerade väl, tog vi i forskargruppen beslut om att expandera denna materialkategori och inkludera riksdagstryck. Det var visserligen något som vi aviserat i vår ansökan, men också varit tveksamma över givet materialets omfång. Nu blev den politiska sfären inom *Välfärdsstaten analyserad* snart mycket större än den mediala (dags-



1981/82



1984/85



1985/86



1987/88





- Om *Välfärdsstaten analyserad* hade problem med vissa materialkategorier på grund av rådande upphovsrätt, så gällde det inte riksdagstrycket – ett text-material som av demokratiska skäl är öppet tillgängligt för alla medborgare. Via systerprojektet SWERIK kom även Riksdagsbiblioteket att ingå i arbetet med att strukturera riksdagsprotokoll till analyserbara data.

press), för att inte tala om den kulturella (romaner och BLM). Här ska inte hymlas med att detta förfarande skapade en sorts obalans, åtminstone i jämförelse med vad vi i ansökan sagt att projektet skulle ägna sig åt. Samtidigt visade sig arbetet med riksdagstrycket vara produktivt. Med Riksdagsförvaltningens och KB:s digitaliseringsarbete som grund lyckades vi på ganska kort tid skapa ett strukturerat dataset av alla riksdagsprotokoll från första och andra kammaren, samt från enkammarriksdagen, mellan 1920 och 2020. Det rörde sig om ungefär elvatusen protokoll, fördelade på sjuhundrafemtiotusen anföranden. Parallellt med detta arbete tog Fredrik Norén och Måns Magnusson initiativ till att skriva en ny

forskningsansökan till RJ om infrastrukturellt stöd för att skapa ett ekosystem av länkade öppna data av *allt* riksdagstryck från tvåkammarriksdagens införande till idag. I regel brukar forskningsprojekt avlösa varandra, men i fallet med *Välfärdsstaten analyserad* gav projektet alltså redan halvvägs upphov till ett nytt forskningsprojekt; hösten 2022 meddelade RJ att denna infrastrukturella ansökan beviljats.

En annan effekt av det intensifierade arbetet med riksdagstrycket var att vårt projekt blev alltmer internationaliserat. *Välfärdsstaten analyserad* var i huvudsak tänkt som ett nationellt forskningsprojekt, men när arbetet kom att fokusera på dokument från riksdagen insåg vi snabbt att det fanns många historiker i andra länder som arbetade med samma typ av digitaliserat material – av den enkla anledningen att det inte förelåg några upphovsrättsliga restriktioner. Inom den europeiska forskningsinfrastrukturen för språkteknologi, CLARIN, fanns exempelvis ett trettio-tal länders dataset av riksdagstryck öppet tillgängliga. Samma öppenhet hade också inneburit att forskare byggt olika slags applikationer för att analysera sådana dataset. Det gällde bland andra *the Hansard* – som kunde analyseras via applikationer som The Hansard Viewer eller Hansard at Huddersfield – eller den tyska Open Discourse som gav tillgång till alla debatter i den tyska förbundsdagen, *Plenarprotokolle des deutschen Bundestages*, från 1949 till idag. På Humlab tog Fredrik Norén och jag därför initiativ till en serie internationella workshops online – Text Mining Parliamentary Democracy – där vi under ett par års tid organiserade flera presentationer av olika slags forskningsarbeten med digitaliserat riksdagstryck i olika länder.

Norén var också projektledare för det nya RJ-projektet SWERIK, som från och med 2023 löpte parallellt med *Välfärdsstaten analyserad* och där även Riksdagsbiblioteket snart blev till en central aktör. SWERIK byggde på flera sätt vidare på vårt tidigare arbete med det digitaliserade riksdagstrycket; i en rad arbetspaket handlade det om att annotera och bygga en databas med metadata för alla riksdagsledamöter, att skapa en infrastruktur där ledamöter i databasen länkades till relaterade riksdagsdokument, samt att anpassa denna annoterade korpus till ParlaMint, en europeisk XML-standard (*extensible markup language*) som utvecklats just för annotering av riksdagsdata. Utan att gå in alltför djupt på tekniska detaljer innebar detta arbete att anföranden i alla riksdagsprotokoll kom-

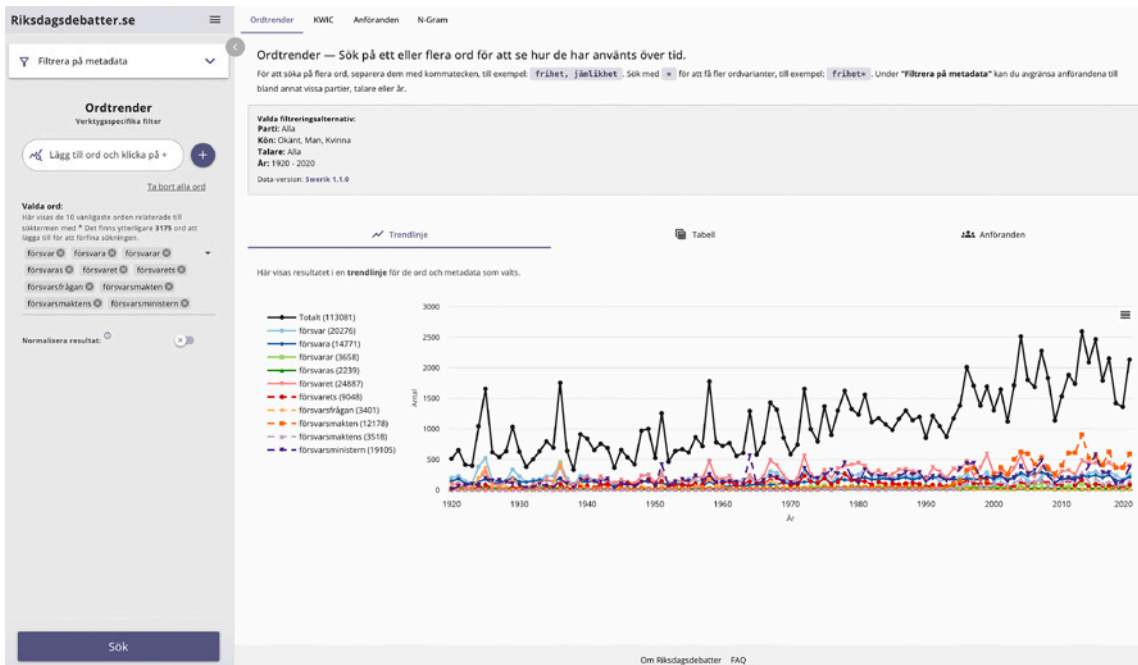
pletterades med metadata för namn på talare, geografisk hemvist, partitillhörighet och kön – ett arbete som kan beskrivas med en mening, men som var mycket tidskrävande. Eftersom detta dataset blev minst sagt omfattande var det i princip omöjligt att strukturera och städa det på manuell väg. En rad skript i programspråket Python kom därför att användas för kurering, med versionskontroll på Github (av den annoterade korpusen). Andra skript användes för att förbättra metadata; de tränades på mindre dataset för att göra datadrivna förutsägelser, varefter de kontrollerades manuellt. För de statistiker och maskinlärningsexperter som arbetade i SWERIK och *Välfärdsstaten analyserad* var detta iordningställande av data ett forskningsarbete i egen rätt; Måns Magnusson, doktorand Väinö Yrjänäinen och forskningsingenjör Robert Borges utförde härvidlag ett imponerande arbete som också resulterade i ett flertal publikationer.²⁵

Tilläggas bör att när det gällde namn på riksdagsledamöter byggde våra två projekt om alla metadata från grunden genom ett informellt samarbete med Wikidata. Där används en auktoritetsdatabas med identifierare, där alla individer (såsom riksdagsledamöter) tilldelas ett unikt identifieringsnummer – socialdemokraten Östen Undén har exempelvis nummer Q298317. Dels innebar detta samarbete en produktiv dialog med den nätgemenskap som fanns kring Wikidata, dels gjorde identifieringsnummer det möjligt att särskilja ledamöter från varandra (om de hade samma namn), förse dem med tidsangivelser när de satt i riksdagen (ibland i omgångar) och ändra deras roll i riksdagen över tid (ledamot, minister, talman etcetera), liksom (mer sällan) förändrad partitillhörighet.

Precis som med de statliga utredningarna var det möjligt att öppet publicera hela vårt gemensamma dataset med riksdagsprotokoll och göra det tillgängligt i Jupyter Lab. Via ett användarvänligt gränssnitt kunde man där som forskare köra (och ändra på) Pythonkod direkt i webbläsaren (på ett snarlikt sätt som beskrivits ovan). Om *SOU*:er tidigare fick sökas fram separat via *KB*:s sajt, så var det nu till och med möjligt att direkt länka samman alla anföranden i riksdagen i Jupyter Lab. Eftersom alla tal av samtliga riksdagsledamöter i elvatusen protokoll (mellan 1920 och 2020) fanns tillgängliga i *XML*-format på Github, gick det att per automatik länka in dem i Jupyter Lab-miljön. Det gjorde det möjligt att både närläsa och distansläsa, att zooma in i och ut ur riksdagsdata, från övergripande diskurser till enskilda yttranden. I ett

temanummer om digital historia i den idéhistoriska årsboken *Lychnos* publicerade jag därför en artikel om riksdagsdata och mediepolitik – vilket förstås bara var en aspekt av hur denna infrastruktur kunde användas.²⁶ Givet alla diskussioner och debatter i Sveriges riksdag så säger det sig självt att otaliga andra frågor än de som handlade om medier kunde studeras ungefär på samma sätt. För de frågor som politiker av tradition ägnat mest tid åt – arbete, skatter, sjukvård, utbildning, försvar, miljö – fanns det naturligtvis en ännu rikare empiri av riksdagsanföranden.

Det öppna arbetet med riksdagsprotokollen i Jupyter Lab resulterade också i idéer om att utvidga denna access till fler intresserade forskare, journalister och allmänheten. För vid denna tidpunkt hade framför allt SWERIK blivit alltmer uppmärksammat av svenska statsvetare som insåg att det annoterade riksdagstrycket nu blivit långt mer forskarvänligt än tidigare. Att bygga ett mer publikt gränssnitt ingick emellertid inte i projektbeskrivningen för vare sig SWERIK eller *Välfärdsstaten analyserad*. Fredrik Norén ansökte därför om ytterligare medel från Umeå universitet för att utveckla en webbaserad applikation för alla upparbetade riksdagsdata. Det var bakgrunden till applikationen riksdagsdebatter.se – som främst varit kopplad till RJ-projektet SWERIK, men som också utgör en sorts sentida utlöpare av *Välfärdsstaten analyserad*. Som jag påpekade i kapitlets inledning har tanken med riksdagsdebatter.se varit att göra det lättare att utforska, läsa och ladda ner alla anföranden av riksdagsledamöter som talat i riksdagen alltsedan 1867. I *Välfärdsstaten analyserad* arbetade vi med en korpus på cirka elvatusen riksdagsprotokoll under en hundraårsperiod, men nu har mängden anföranden via SWERIK ökat till artotusen digitaliserade protokoll. I applikationen finns möjligheter att söka på ordtrender eller göra så kallade KWIC-sökningar, *keywords in context*, som tillåter undersökningar av hur ord eller fraser använts i olika sammanhang via olika slags stora (och reglerbara) ordfönster. Det går också att göra sökningar på diverse *n-grams*, det vill säga diagram över hur ofta ord förekommer. Eftersom alla riksdagsdata strukturerats är det också möjligt att göra sökningar på partitillhörighet, även om ett problem varit att regelrätta partier inte fanns i början av tvåkammarriksdagen. Den stora fördelen med riksdagsdebatter.se är emellertid att det går att göra mer övergripande sökningar på ordtrender – såsom välfärdsstat* – och därefter på ett enkelt sätt få fram vilka debatter som ägnades åt detta. I gränssnittet går det



- På Karl Sandels fotografi från andra kammarens försvarsdebatt 16/3 1938 har ledamöterna grupperat sig i den så kallade gropen – där hörde man nämligen anföranden från talarstolen allra bäst. Några mikrofoner eller högtalare fanns då ännu inte i riksdagen. I applikationen riksdagsdebatter.se framgår just att försvarsfrågan figurerade synnerligen frekvent i riksdagen; bara under 1938 talade riksdagsledamöter om försvaret 327 gånger (vilket framgår av grafen).



också att i txt-format ladda ned anföranden, precis så som vi i *Välfärdsstaten analyserad* önskat varit fallet med BLM, svenska romaner och dagspress. Beträffande den debatt jag inledde detta kapitel med, om välfärdsstatens idé i riksdagens andra kammare våren 1950, kan man som forskare faktiskt med ett enda knapptryck få tillgång till alla tiotalet debattörer som då diskuterade – från utrikesminister Undén till partiledarna Ohlin och Hjalmarson, samt en rad andra ledamöter.

Avslutning: något av ett misslyckande

Att *Välfärdsstaten analyserad* delvis blev ett misslyckat forskningsprojekt går inte riktigt att bortförklara. Det aviserade arbetet om att på makronivå jämföra olika sfärer av text från politikens, mediernas och kulturens områden under svensk efterkrigstid och välfärdsår resulterade inte i något komparativt praktiskt forskningsarbete, och följaktligen inte heller i några publikationer som drog åt det hållet. När det gällde kulturens sfär var det största problemet att det tog för lång tid att digitalisera alla svenska romaner, även om tidsperioden blev kortare än tänkt (1940–1965). Här ska KB emellertid inte lastas; tvärtom gjorde nationalbiblioteket och dess personal en god insats med detta digitaliseringsarbete. Att det hela drog ut på tiden berodde framför allt på att uppgiften visade sig vara mer komplicerad än vad vi föreställt oss den när vi skrev ansökan – som då naturligtvis gjordes i dialog med KB. Det var helt enkelt komplicerat att få fram uppgifter om och exemplar av *alla* romaner under en viss tidsperiod. Beträffande BLM gick det snabbare att få klart all digitalisering, även om det tog tid innan tidskriften publicerades i portalen Svenska tidningar. Precis som med dagspressen fick vi dock problem när vi skulle kurera tidskriftens innehåll till strukturerad data, så att det exempelvis blev möjligt att sortera ut alla bok- eller filmrecensioner. Om det för dagspress visat sig vara bökigt att avskilja ledartexter från kultursidesartiklar i större skala, gällde detsamma för BLM i mindre skala. Trots att våra maskinlärningsexperter tagit fram ett flertal skript för att på automatisk väg strukturera riksdagsprotokollen, fick vi inte dessa att riktigt fungera på de femtiotusen sidorna av BLM.

Men framför allt blev forskningen kring dagspress, romaner och BLM i *Välfärdsstaten analyserad* kringskuren av upphovsrättsliga skäl. Även om vi under projektets uppstartsfas arbetade med flera

digitaliserade tidningar mellan 1945 och 1989 som större dataset, presenterade vår forskning om dem på konferenser och därtill publicerade vissa resultat – så var det som om luften pös ur medie- och kultursfären på grund av svårigheten att växla mellan mikro och makro. Flera av oss forskare i projektet betraktade sig som digitala humanister, men vi var samtidigt också traditionella mediehistoriker. Dagspressmaterialet förlorade därför successivt sin lockelse när det visade sig vara omöjligt att zooma in i och ut ur text (inom samma dataset), så som vi tidigare arbetat i forskningsprojektet *Digitala modeller*. I princip förhöll det sig på samma sätt med romanerna och BLM, även om den försenade digitaliseringen av dem orsakade mest problem. Vän av ordning kan här invända att projektet, för att svara upp mot de makroorienterade frågeställningar vi aviserat i ansökan, inte behövde närläsa dessa texter eftersom det handlade om komparativ storskalig analys. Det stämmer visserligen, men samtidigt tog det lång tid för KB, dess ledning och verksjurister, att ge klartecken för att exportera till exempel ordfrekvenser ur dagspress (och senare romaner) till våra utvecklare på Humlab. Även detta juridiska förfarande sinkade forskningsarbetet.

Det är alltid enkelt att vara efterklok; tveklöst hade det storskaliga textarbetet underlättats om en datakurerare med utveck-larkompetens varit stationerad på KB-labb. Då hade med all sannolikhet åtminstone det mer storskaliga jämförande arbetet med olika textsfärer kommit till stånd – som ju egentligen var oberoende av diverse hermeneutiska närläsningar av enskilda tidningsartiklar, BLM-nummer eller romaner. Samtidigt var det ju precis så som *Välfärdsstaten analyserad* var designat. För att underlätta sammanställande, kurering och strukturering av dataset samt export till utvecklarna på Humlab hade som sagt nästan två miljoner kronor avsatts för en särskild tjänst – vilken KB-labb alltså av-sade sig. Som jag tidigare beskrivit fanns det emellertid goda skäl för detta beslut. Generellt fungerade samarbetet med KB väl under projektperioden, framför allt beträffande digitaliseringsarbetet med BLM och romanerna (även om det tog lång tid). Dessutom hade nationalbiblioteket uttryckligen prioriterat vårt forskningsprojekt, något jag själv tog till intäkt för att frågan om fjärraccess till vårt digitaliserade forskningsmaterial nog skulle lösa sig. Den sammantagna budgetposten för KB inom projektet var trots allt nästan sju miljoner kronor. Men detta gick alltså inte vägen; data

med upphovsrätt kunde bara i begränsad form lämna KB – trots att nationalbiblioteket faktiskt ingick i vårt projekt.

I sammanhanget bör det framhållas att verksamheten på KB-labb naturligtvis hade att följa svensk lagstiftning. Hur denna egentligen skulle tolkas var dock en annan sak; verksamheten där fick under projektiden också kritik från annat forskarhåll på grund av svårigheter att få loss data och textmaterial.²⁷ Välfärdsstaten förblev på ett plan alltså oanalyzerad i vårt projekt, men vi var desto mer framgångsrika i andra sammanhang. Vårt projekt inbegrep trots allt både forskning, kurering och digitalisering. Om vissa forskningsresultat var mindre omfattande, var projektet mer lyckosamt när det gällde att digitalisera två större textsamlingar (romaner och BLM). Dessutom gjordes flera framsteg när det gällde att automatiskt kurerat och strukturera större textsamlingar, skript som gjordes allmänt tillgängliga på utvecklarsajten Github. Några av oss forskare i projektet håller också i skrivande stund på att färdigställa en monografi om riksdagens mediehistoria, där vi med utgångspunkt i det storskaliga textarbetet inom *Välfärdsstaten analyserad* kombinerar kvalitativ forskning med kvantitativ analys.²⁸

Till sist: som i många akademiska sammanhang har *Välfärdsstaten analyserad* givit upphov till en ny forskningsansökan, delvis baserad på arbetet med att digitalisera de svenska romanerna på KB. I skrivande stund (våren 2025) har jag färdigställt en större ansökan till VR för att etablera en nationell infrastruktur för svenskt tryck (Sweprint). Syftet är att stödja all forskning som inbegriper analys av svenskt tryck genom att bygga en digital plattform, massdigitalisera svenskt tryck och aggregera tidigare digitaliserat textmaterial. Infrastrukturen kommer också – i samarbete med KB-labb – att erbjuda stöd för avancerade digitala analysmetoder. Sweprint baseras på ett konsortium bestående av universitetsbiblioteken i Lund, Umeå, Uppsala, Stockholm och Göteborg, samt KB och Svenska Akademien. Och för att inte ånyo hamna i bryderier om upphovsrätt och fjärraccess inbegriper den faktiskt en hel modul som enbart ska ägna sig åt rättsliga frågor, som det står i projektansökan: "[the module] is tasked with investigating the exact legal consequences of digitizing Swedish print, including how to best cover such materials with as little research restrictions as possible".

NOTER

1. Riksdagens protokoll AK 1950:11, 11, 24, 26.
2. Riksdagens protokoll AK 1927:35, 22.
3. Forskningsprojektet *Välfärdsstaten analyserad* finansierades av VR under sex års tid, mellan 2019 och 2024. Professor Pelle Snickars, författare till denna artikel, har varit projektledare. *Välfärdsstaten analyserad* var först förlagt på Umeå universitet, därefter (från 2022) var Lunds universitet medelsförvaltare. Projektet har utöver projektledaren involverat fem forskare: docent Måns Magnusson – till en början på Aalto-universitetet i Finland, och därefter på Uppsala universitet – docent Johan Jarlbrink och FD Erik Edoff, bägge vid Umeå universitet; FD Fredrik Mohammadi Norén – först på Umeå universitet, sedan på Malmö universitet – samt docent Alexandra Borg, Uppsala universitet. Projektet har även delfinansierat doktoranderna Väinö Yrjänäinen, Uppsala universitet, Johan Malmstedt, Umeå universitet, samt Hannes Waldetoft, Uppsala universitet. Inom projektet har också två programmerare vid Humlab på Umeå universitet varit verksamma: chefsutvecklare Roger Mähler och FD Tomas Skotare. På Uppsala universitet har forskningsingenjör FD Robert Borges hjälpt till med att skriva kod och strukturera data. På Kungliga biblioteket har projektledare Stefanie Topelius varit ansvarig för alla digitaliseringsarbeten inom projektet – och på KB har även personal på KB-labb varit med på ett hörn av projektet. Till sist har *Välfärdsstaten analyserad* haft fyra masterstudenter i statistik vid Uppsala universitet anställda som amanuenser (i kortare perioder), vilka i huvudsak varit sysselsatta med att kurerat data: Robin Nasser, Liam Tabibzadeh, Emil Lanzén och Markus Ridhagen.
4. Forskningsprojektet *Digitala lägg* vid Umeå universitet involverade Pelle Snickars (projektledare), utvecklare Christian Colliander och Roger Mähler, samt inte minst Johan Jarlbrink, som utförde det mesta av det praktiska forskningsarbetet. För en vidare diskussion, se Johan Jarlbrink och Pelle Snickars, "Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive", *Journal of Documentation* nr 6, 2017, samt Johan Jarlbrink, "Telegrafens på distans: Ett digitalt metodexperiment", *Scandia* nr 1, 2018.
5. Tim Hitchcock, "Confronting the digital, or how academic history writing lost the plot", *Cultural and Social History* nr 1, 2013.
6. Ett resultat av forskningsprojektet *Digitala lägg* var också att personal på Kungliga biblioteket började intressera sig för hur den egna tidningsdigitaliseringen kunde förbättras. 2018 fick man t.ex. ett anslag från RJ för att utvärdera och anpassa en förbättrad OCR-process vid massdigitalisering av dagspress. Ambitionen var att uppdatera produktionen av maskinläsbar text och utveckla en modul för OCR-bearbetning, där centrala parametrar kunde justeras för att matcha olika karakteristika hos källmaterialet.
7. Forskningsprojektet *Digitala modeller* var ett samarbete mellan Humlab på Umeå universitet och Tekniska museet och finansierades av RJ och Kungliga Vitterhetsakademien 2016–19. För mer information om projektet, se Jenny Attemark-Gillgren och Pelle Snickars (red.), *Digitala modeller: Teknikhistoria och digitaliserings specificitet* (Lund: Mediehistoriskt arkiv, 2019).

8. Temamodellering (*topic modeling*) är ett samlingsnamn på olika algoritmer som kan klassificera stora korpusar i tematiska strukturer baserat på ord-frekvens. För en vidare diskussion, se Johan Jarlbrink, "Från masugn till mikro-vågsugn: Historiografisk topic-modellering av årsboken *Daedalus* (1931–2017)" i *Digitala modeller* (2019).

9. Alla årsböcker återfinns på projektsajten digitalamodeller.se, som Tekniska museet tog ansvar för efter projektslut 2020. Eftersom *Daedalus* alltsedan 1931 publicerats av Tekniska museet eget förlag, i samarbete med föreningen Tekniska museets vänner, menade museidirektör Peter Skogh och (dåvarande) avdelningschef Åsa Marnell att museet hade rätten att publicera alla årsböcker fritt på webben. En författare i årsboken menade emellertid att detta förfarande utgjorde ett upphovsrättsintrång, trots att vederbörande under årens lopp fått betydande hjälp av museet. Parterna förlikades sedermera – all heder till museiledningen för de forskarvänliga beslut som då togs.

10. Historikern Jo Guldi har bland annat i en storskalig textanalys av politiska diskussioner om brittisk infrastruktur under 1800-talet och hur dessa förändrade sig över tid, empiriskt baserad på *the Hansard*, påpekat att sådan undersökningar gör det möjligt att spåra "the invisible categories that structure mind, language, priorities, or prejudice in a given corpus". Se Jo Guldi, "Parliament's debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change", *Technology and Culture* nr 1, 2019.

11. Lars Ilshammar, "Förord", *Datalabb på KB* (Kungliga biblioteket, 2018), <https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-339> (senast kontrollerad 1/5 2025).

12. Min dåvarande doktorand Fredrik Norén tog sig bland annat an alla statliga utredningar som ett dataset för att synliggöra den statliga informationspolitikens förändring från 1960-talet till 1980-talet. Tillsammans publicerade vi också en annan artikel baserad på samma dataset med sou:er där vi distansläste svensk filmpolitik via utredningsväsendet. Se Fredrik Norén, "Information som lösning, information som problem: En digital läsning av tusentals statliga utredningar", *Nordicom-Information* nr 3, 2016, samt Fredrik Norén och Pelle Snickars, "Distant reading the history of Swedish film politics in 4500 governmental sou reports", *Journal of Scandinavian Cinema* nr 2, 2017.

13. Ryan Cordell, *Machine learning + libraries* (LC Labs 2020), <https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf> (senast kontrollerad 1/5 2025).

14. Pelle Snickars, *Datalabb på KB* (Kungliga biblioteket, 2018), <https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-339> (senast kontrollerad 1/5 2025).

15. "KB fortsätter satsningen på datadriven forskning", pressmeddelande 10/9 2021, <https://www.kb.se/om-oss/nyheter/nyhetsarkiv/2021-09-10-kb-fortsatter-satsningen-pa-datadriven-forskning.html> (senast kontrollerad 1/5 2025).

16. Det sociologiska forskningsprojektet *Mining for meaning* finansierades av VR 2019–2024 och leddes av biträdande professor Marc Keuschnigg vid Linköpings universitet. En av doktoranderna i projektet, Miriam Hurtado Bodell, genomförde ett omfattande arbete på KB-labb med pressmaterial från

1945 till 2020 – vilket även var till stor nytta för vårt projekt. Hon disputerade 2024 på en avhandling som just använde storskalig, datadriven analys för att studera tidningars roll i olika diskurser kring immigration. För en vidare diskussion, se Miriam Hurtado Bodell, Måns Magnusson och Marc Keuschnigg, "Seeded topic models in digital archives: Analyzing interpretations of immigration in Swedish newspapers, 1945–2019", *Sociological Methods & Research* 22/8 2024, <https://doi.org/10.1177/00491241241268453> (senast kontrollerad 1/5 2025).

17. Se t.ex. Pelle Snickars, "Från chiffer till klartext? Temamodellering av statliga offentliga utredningar 1945–1989", *Scandia* nr 1, 2022; Pelle Snickars, "Tema kulturarv: Temamodeller om ABM i svenska statliga offentliga utredningar 1945–89", *Nordisk Museologi* nr 2, 2022; Pelle Snickars, "Modeling media history: On topic models of Swedish media politics 1945–89", *Media History* nr 3, 2022.

18. Stefanie Topelius, "Delrapport för projektet KB-WeStAc", 17/9 2020. Opulicerad, intern rapport från KB.

19. Stefanie Topelius, *KB-WeStAc: En rapport om digitisering på KB* (Kungliga biblioteket, 2022), <https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-742> (senast kontrollerad 1/5 2025).

20. Vid sidan av Stefanie Topelius ingick ytterligare sju personer från Kungliga biblioteket i den projektgrupp som hanterade digitaliseringen av BLM och de svenska romanerna: Jonas Ahlberg, Mårten Asp, Stina Degerfors, Märta Forsberg, Heidi Rosen, Ylva Sommerland och Julia Wahlund.

21. Se Svenska tidningar, "Användarregler och upphovsrätt" 17/10 2024, <https://feedback.blogg.kb.se/forums/topic/upphovsratt/> (senast kontrollerad 1/5 2025).

22. Arbetet resulterade sedermera i denna text: Fredrik Norén, Johan Jarlbrink, Alexandra Borg, Erik Edoff och Måns Magnusson, "The transformation of 'the political' in post-war Sweden", *Digitised newspapers: A new eldorado for historians? Reflections on tools, methods and epistemology* (red.) Estelle Bunout, Maud Ehrmann och Frédéric Clavert (Berlin: De Gruyter, 2022).

23. Digarv-projektet *Expansion och mångfald: Digital kartläggning och analys av den utominstitutionella scenkonsten i Göteborg 1965–2000* vid Göteborgs universitet betalade under en period en extra licensavgift till KB för att få access till portalen Svenska tidningar.

24. *Forskning och innovation för framtid, nyfikenhet och nytta* (Prop. 2024/25:60), 67. För en diskussion om utvecklingen av KB-labb, se Love Börjesson et al., "Transfiguring the library as digital research infrastructure: Making KBLab at the National Library of Sweden", *College, Research & Libraries* nr 4, 2024.

25. Se t.ex. Väinö Yrjänäinen et al., "The Swedish Parliament Corpus 1867–2022", *ELRA Language Resource Association* (2024), <https://aclanthology.org/2024.lrec-main.1400.pdf> (senast kontrollerad 1/5 2025).

26. Pelle Snickars, "Randanteckningar om svensk mediepolitik 1920–1970: Ur temamodellering av 11 000 riksdagsprotokoll", *Lychnos: Årsbok för idé- och lärdomshistoria* (Uppsala: Lärdomshistoriska samfundet, 2022).

27. Se t.ex. Martin Röshammar, "kb-labb: Infrastruktur för AI-utveckling med förhinder", *Biblioteksbladet* 16/6 2023, <https://www.biblioteksbladet.se/nyheter/kb-labb-infrastruktur-for-ai-utveckling-med-forhinder/> (senast kontrollerad 1/5 2025).

28. Fredrik Norén, Johan Jarlbrink och Pelle Snickars, *Riksdagens mediehistoria* (Lund: Mediehistoriskt arkiv, kommande 2026).