KILDE OG DATA

OVERVEJELSER OM HISTORIEFAGET OG DE DIGITALE METODER

JOHAN HEINSEN

Det er næppe længere en kontroversiel påstand, at det digitale og computationelle ændrer historiefaget. Tilsvarende opfattelser gør sig gældende på tværs af humaniora og samfundsvidenskab. Om vi kan lide det eller ej, ændrer ikke på sagen. I historiefaget har nye tilgange affødt endnu en faglig kategori: digital historie. De fleste synes enige om, at dette ikke er et nyt felt i sig selv – og altså ikke defineret ved sin genstand – men derimod ved nye praksisser, nye måder at gøre historie på.[1] Disse nye måder fletter vores fag sammen med andre fagligheder, der ikke traditionelt har været vores nærmeste disciplinære slægtninge. Dette korte indlæg præsenterer nogle overvejelser over, hvad denne gøren betyder for nogle af de fundamentale metodiske og videnskabsteoretiske antagelser, vores fag hviler på. Jeg vil pege på nogle af de potentialer, denne udvikling bærer. Men jeg vil også tillade mig at rejse et par kritiske spørgsmål. Med den digitale histories import af teknikker, følger nemlig også andre måder at tænke empirien og dens værdi på. Det gør en forskel, når vi taler om "data" frem for "kilder".

Vi kan starte med det helt simple spørgsmål: Hvad er det nye ved et digitalt historiefag?

1 I en oversigtsartikel formulerer Romein m.fl. denne pointe: "In this sense, we can understand digital history not as a distinct discipline or field, but as a community of practice of researchers from different backgrounds who look across institutional and disciplinary boundaries to engage in historical practices with the methodological and epistemological concept of other disciplines." (Romein m.fl. "State of the Field", 293).

HVILKEN FORSKEL GØR DET DIGITALE?

En grundlæggende og helt konkret forskel er, at flere og flere af os underviser i metoder og værktøjer, der involverer computere og materiale, computeren kan afkode. I Aalborg og Aarhus er denne proces nået længst. Her har man i flere år haft deciderede kurser i digitale og computationelle metoder på bachelor- og kandidatuddannelserne i historie, ofte tæt knyttet til mere velkendte kompetencer i brugen og skabelsen af arkiver. Begge steder undervises der på nuværende tidspunkt i analysestrategier, der gør brug af programmerings- og visualiseringsværktøjer. Og begge steder er disse undervisningsforløb skrevet ind i studieordninger, hvorved de digitale kompetencer tænkes som en integreret del af det at være historiker, uagtet at det ofte er langt fra hele underviserstaben, der selv besidder disse kompetencer. På andre af landets universiteter er det digitale ikke lige så integreret i undervisningsporteføljen, men det har alligevel sneget sig ind her og der – knyttet til mere tematisk eller periodeorienterede undervisningsforløb.

At der undervises i noget afspejler, at der forskes i noget eller i hvert fald tales meget om at forske i noget. Skal man være lidt grov, tales der for nuværende mere, end der egentlig gøres, men hvor der er røg, er der som regel også ild i et eller andet.

Hvordan har det digitale påvirket, hvad vi gør som historikere, når vi forsker? Herhjemme synes det at have gjort udslag i hvert fald en håndfuld praksisser. Jeg vil præsentere de mest udbredte først:

1. Vi er blevet brugere af digitale arkiver. Det er også i denne funktion, at vi har haft den mest robuste tænkning af det digitales betydning. Helle Strandgaard Jensens arbejder har sat værdifulde ord på hvilke kompetencer dette kræver, og hvordan vi kan tænkte vores eksisterende historiefaglighed med ind i denne kontekst.[2] Fra et mere arkivfagligt perspektiv har Marianne Sletten Paasch blandt meget andet diskuteret, hvordan digitale bevaringsstrategier former fremtidens brugsscenarier i arkiverne.[3] Udenlands har historikere ligeledes diskuteret de digitale arkiver blandt andet med et tilbagevendende fokus på, hvad det gør ved vores forståelse af fortidige materialer, at de bliver søgbare.[4]
2. Mange af os er i processen blevet skabere af vores egne arkiver. Der er som sådan ikke noget nyt i, at historikere indsamler materialer, og at disse ender som deres egen samling. Forskellen handler om skala. Da mere og mere materiale er blevet gjort stadigt lettere tilgængeligt, er indsamlingen og omarbejdningen af materiale blevet desto mere omfattende. Går man til konferen-

2 Jensen: "Digital Archival Literacy for (all) Historians"; Schriver og Jensen: "Arkivets digitalisering".

3 Paasch: *Gemt eller glemt?*

4 Putnam: "The Transnational and the Text-Searchable".

cer, især de internationale, hører man da også flere og flere historikere omtale deres materiale som datasæt. Jeg har gjort det selv, uden at tænke over det. Lytter man efter på universitetsgangene eller læser kollegers fondsansøgninger, møder man også datasættet. Som et grundformat for vores empiri, sniger datasættet sig således ind i vores akademiske dagligsprog. Reelt er der ret stor spændvidde i, hvad der menes med et datasæt. Nogle gange betyder det blot, at man har en samling af transskriberede tekster. Andre gange, at materialet er tabuleret og kategoriseret. Alligevel antyder sprogbrugen måske noget grundlæggende om karakteren af bearbejdningen af materialet (det vil jeg komme ind på senere). Det spiller også sammen med en bredere forskningspolitisk dagsorden om, at forskningsempiri generelt gerne skal tilgængeliggøres (så vidt det lader sig gøre i henseende til ophavsret, GDPR og almen etik). At sige, at man har et datasæt, er således på mange måder det samme som at sige, at man har skabt en form for arkiv, der kan leve videre, uanset om man er god til f.eks. at skabe metadata eller ej.

3. For nogle har denne dataskabelses- og tilgængeliggørelsestrang taget en mere drastisk drejning. Her tænker jeg især på de historikere, der har kastet sig over at bruge *deep learning*-værktøjer til at skabe store korpusser af materiale. Transkribus er det mest kendte og udbredte af disse værktøjer, men flere alternativer eksisterer eller er på vej. Teknologien er kommet milevidt fra de OCR-baserede tekstgenkendelsesløsninger, der har eksisteret i årtier og som virkede fint på trykt tekst i høj opløsning, men ellers var ganske ubrugelige. De første danske modeller handlede om latinske og trykte bogstaver, men i de seneste par år har blandt andre Nina Koefoed i Aarhus og jeg selv i Aalborg skabt modeller, der læser gotisk fra 17- og 1800-tallet med en præcision, der varierer kraftigt efter materialets karakter, men nogle gange producerer tekst, der uden problemer kan forstås af mennesker – og computere. For ti år siden omtalte historikeren Tim Hitchcock vores fag som halvvejs igennem en digitaliseringsrevolution. Han tænkte på de store digitaliseringsprojekter af vestens trykte kanon, der på daværende tidspunkt var undervejs.[5] Kunstig intelligens synes at have bragt os et skridt videre i denne proces. For tiden trænes der rundtomkring i Danmark modeller, der kan læse 15- og 1600-tals tekst, men også modeller, der kan læse f.eks. Mediestreams gnidrede avissider med en præcision, der simpelthen ikke var mulig for bare få år siden. I projektet Retro-Digitalisering.dk, ledet af Aarhus Stadsarkiv, deltager desuden en lang række mindre institutionelle aktører i udviklingen af disse værktøjer, der således har en bemærkelsesværdigt bred forankring allerede.[6] Blandt de store fonde har især Carlsberg kastet penge efter projekter, der skal afføde Transkribus-mo-

5 Hitchcock: "Confronting the Digital".

6 https://www.retrodigitalisering.dk/ (08.05.2023).

deller.[7] Om få år vil der findes modeller til langt det meste af det materiale vi danske historikere bruger, og situationen vil være den samme i de fleste lande, hvor forskningsinfrastrukturen tillader det. Det vil betyde, at alt dansksproget materiale reelt vil kunne gøres søgbart – eller omformes til data for andre algoritmer, f.eks. såkaldt *named entity recognition*, der kan identificere personer og steder – eller store sprogmodeller, der kan skrive resumeer af tekster. At transskribering gennem Transkribus koster penge betyder imidlertid også, at vi nu kan se endnu en igle fæstne sig på allerede blege forskningsbudgetter. At forskningsmidler reelt har været brugt og stadig bruges til at skabe disse modeller, der herefter bliver til et kommercielt produkt for en tredjepart, er åbenlyst ikke uproblematisk.

4. Det, der kommer ud af disse digitaliseringsprojekter, er typisk rå tekstdata. Ser man på det institutionelle landskab, skyder det op med forskningsenheder og projekter, der benytter teknikker udviklet under den brede paraply af *digital humanities* til at studere fortidig tekst. Det er ikke ensbetydende med, at der faktisk er historikere ombord, men nogle gange er der. De teknikker, der bruges, har typisk ikke ophav i et enkelt fag, men i tværvidenskabelige miljøer med klar inspiration fra f.eks. computer science og forskning i *natural language processing*. I denne kontekst er det afgørende, at man forstår tekst som data, der kan behandles computationelt og kan kvantificeres. Konkrete analyser behandler ofte en enkelt type tekst, f.eks. litterære romaner, avisannoncer eller Twitter-opslag. Metoderne roterer derfor meget ofte om at finde mønstre i et serielt materiale uden at læse materialet med øjnene.[8] Meget *digital humanities* praktiseres uden brug af programmeringsværktøjer, og der findes lettilgængelig software til alt fra tekstanalyse til netværksvisualisering. Det er således ikke så meget værktøjerne, der fremstår fremmede for historikeren, men for mange nok snarere det kontinuerlige fokus på en enkelt type tekst. I udgangspunktet har vi jo netop været interesserede i konkrete processer og alle de levn, disse har affødt – ikke på en udvalgt type af materiale i sig selv.
5. Parallelt (og ikke fuldstændigt uafhængigt) er opstået flere og flere miljøer med rod i *social data science* – den samfundsvidenskabelige pendant til *digital humanities*. Også her er man interesserede i tekst og andre former for data, der med computationelle teknikker kan gøres meningsfulde for en computer. Det giver imidlertid mening at tænke denne trend som forskellig fra *digital humanities*, da man i *social data science* langt tydeligere bygger oven på traditioner for statistisk analyse. Man arbejder således med data for at afsøge mønstre og

7 Se https://www.carlsbergfondet.dk/da/Bevillingshaver/Formidling/Bevillingsoversigt/CF20_0228_Nina-Javette-Koefoed (08.05.2023).; https://www.carlsbergfondet.dk/da/Forskningsaktiviteter/Bevillingsstatistik/Bevillingsoversigt/CF22_1426_Louise-Nyholm-Kallestrup (08.05.2023).

8 Et dansk eksempel er Jensen m.fl.: "Scalable Reading of Structured Data".

kausalitet uden nødvendigvis at have en faglig relation til kvalitative metoder til brug af tilsvarende materiale. Meget hårdt sat op kan vi måske tænke forskellen på *digital humanities* og *social data science* som forskellen på, hvad der sker, når man omfavner computeren fra henholdsvis et idiografisk og nomotetisk videnskabsideal. I virkeligheden er der selvfølgelig et væld af gråzoner, men det forklarer alligevel, hvorfor man i *social data science* altid taler om modeller. Her kan være tale om mere klassiske statistiske modeller, f.eks. den for mange velkendte regression, men selvfølgelig også stadig mere avancerede former for *machine learning*. Modellering er i denne forstand tænkt ret traditionelt, som en proces, der gør forskeren i stand til at forudsige det næste tilfælde af x. Dermed forudsættes det groft sagt, at alle tilfælde af x er ens nok til, at f.eks. historicitet ikke forstyrrer modellens spådomsevner. Herhjemme er det begrænset med overlappet imellem miljøerne i *social data science* og historiefaget, men andre steder ser man stadigt hyppigere, at historikere bruger sådanne modelleringsteknikker og de programmeringsværktøjer (især Python), der har udviklet sig til standard i feltet. Det synes især at gøre sig gældende i de grene af vores fag, der har tænkt sig mest entydigt som samfundsvidenskabeligt forankret, f.eks. dele af socialhistorien og den økonomiske historie.

Vi kan måske tænke dette som fem trin, der på mere og mere gennemgribende vis ændrer på, hvad vi gør som historikere. På første trin er det digitale et spørgsmål om, at det er blevet lettere at bruge arkivet, fordi det har ændret brugerflade. På de næste trin ændrer arkivet også karakter i kraft af ændrede vilkår for tilblivelse og cirkulation af materiale. På tredje trin gøres arkivet desuden til data for en algoritme, der synes at have potentiale til at accelerere udviklingerne på de første trin yderligere. På fjerde og femte trin er arkivet i sig selv forstået som data, ofte også som *big data.* På sidste trin har vi på alle måder forladt humanvidenskabens traditionelle domæne.

Disse trin kan også forstås som en (ikke nødvendigvis lineær eller irreversibel) vej imod et databegreb.[9] Og dette databegreb erstatter et ældre vokabular, hvor vi historikere talte om vores materiale som kilder. Denne transformation er ikke kun et spørgsmål om at gøre sig genkendelig (som når vi skriver om vores data i vores forskningsansøgninger, fordi vi ved, at de læses af tværvidenskabelige paneler) eller at fremstå tidssvarende (som når vi på konferencer taler om vores datasæt for at vise, at vi også er med på noderne). Som en der har begået sig på alle trin i ovenstående, vil jeg vove den påstand, at det gør en kvalitativ forskel, hvis vi i udgangspunktet tænker om vores materiale som data og ikke som kilder.

9 I en nylig artikel i *History and Theory*, har Stephen Robertson forsøgt at tegne konturerne af digital histories "egenskaber". Her er det helt centrale element, at digital historie roterer om "data" og transformationen af materialet fra "sources" til "data" (Robertson: "The Properties of Digital History").

Og jeg mener, at det er vigtigt, at vi italesætter den forskel – også for de studerende, der med al hast introduceres for fagre nye verdener gennem en efterhånden betragtelig hob af digitale værktøjer født andetsteds.

DATA ELLER KILDE?

Der har i en del år været en både intensiv og bred debat om databegrebet, og hvad det implicerer i moderne akademisk liv. For næsten 10 år siden skrev geografen Rob Kitchin eksempelvis om det fremstormende databegreb, at det implicerede, at data blandt andet var "huge in volume, (…) high in velocity (being created in near real-time); diverse in variety, being structured and unstructured in nature; exhaustive in scope, striving to capture entire populations or system".[10] Den nye virkelighed var derfor en dataoverflod, hvor empirien ofte ikke var blevet til som resultat af et forskningsdesign, men var et produkt af igangværende processer – ofte online. Dette stod i kontrast til tidligere, hvor forskningsdata typisk var nøjsomt kurateret eller samplet og dermed målrettet skabt. Hvor store dele af de kvantitative teknikker, der prægede menneskevidenskaber tidligere, netop søgte at håndtere dette benspænd (at data var et lille sample), er *big data* netop store, fordi de forestilles at indeholde alt det digitale materiale, der er affødt af en given proces og eksempelvis er indsamlet ved at *scrape* det fra nettet – eller i vores tilfælde fra arkivet. Databegrebet implicerer således ofte en ide om fuldstændighed.

I denne nye situation synes de videnskabelige konventioner at ændre sig. Hvor både kvalitativt og kvantitativt orienterede forskningsprojekter tidligere handlede om at udnytte en begrænset mængde af ressourcer til størst mulig effekt ved at designe det bedst mulige udsnit af empiri for derved at kunne sige noget så vægtigt som muligt, er dette simpelthen ikke længere den alment anerkendte præmis. Hvis forskeren ikke længere arbejder med et nøje designet sample, men derimod med en totalitet eller fuldstændig population, ændres det, man kan sige med data. De teorier, der tidligere skulle guide definition af problemstilling, indsamling af empiri og analysestrategi, mister værdi. Enkelte har med profetisk sikkerhed ligefrem annonceret "the end of theory".[11] Selvom de færreste akademikere abonnerer på dette synspunkt, har flere påpeget, hvordan datafikseringen går hånd i hånd med en omformulering af vidensidealer.[12] Fra data-entusiasmens position bliver spillet et, hvor man med en nærmest uendelig datamængde kan benytte et væld af teknikker (hjulpet af stadig mere tilgængelig regnekraft) til at identificere mønstre og bygge modeller, uden nøje designede foruddefinerede forskningsspørgsmål. Validitet handler derfor om størrelse.

Databegrebet er således ofte associeret med en ny empirisme, vægtning af induktive metoder og en understregning af, at fakta etableres kvantitativt og står

10 Kitchin: "Big Data, new epistemologies and paradigm shifts".
11 Anderson: "The End of Theory".
12 boyd og Crawford: "Critical Questions for Big Data".

i kontrast til kvalitativ fortolkning. Denne forestilling har mange (nogle gange modstridende) udtryk. I vores eget fag synes denne omformulering ofte sammenfiltret med makrohistoriske dagsordener, der har genintroduceret ideen om en historieskrivning, der kan syntetisere menneskehedens udviklingshistorie i sin helhed.[13] Historikeren Walter Scheidel har udtrykt dataoptimismen og den opfattede kontrast til ældre erkendelsesinteresser således: "While too many humanities scholars remain committed to 'irreducible complexity,' fetishization of cultural idiosyncracies and condemnation of social evolutionism, big data projects keep advancing our understanding of how the world got to be the way it is".[14]

Kitchin identificerede bevægelsen imod denne videnskabsforestilling med strømninger indenfor især *business science*. Her formuleredes en dagsorden, hvor dataene blev set som bærende iboende mønstre, der blot skulle afdækkes for at kunne skabe værdi. Teori og domæneviden var mindre afgørende, hvorimod teknik, legemliggjort i den fremstormende *data science* (en videnskab uden foruddefineret genstand udover "data"), fik en hastigt voksende betydning. Samtidig fik data en ny værdisættelse, og især er datas størrelse i stadig større grad blevet artikuleret som det, der skaber muligheden for indsigt. Et for alle velkendt eksempel er de allestedsnærværende algoritmer bygget til at anbefale forbrugere nye produkter. Disse er konceptuelt uhyre simple: hvis du kunne lide film x, kan du nok også lide film y, for andre forbrugere, der kunne lide x syntes også om y. Der skal ikke nogen stor hypotese til at formulere dette regnestykke. Men det kræver meget data at få et nogenlunde indblik i, hvordan forbrugeres præferencer korrelerer eller divergerer. Den grundlæggende antagelse er åbenlyst nomotetisk: Forstået som forbrugere er vi alle af samme type, blot med præferencer, der skal kortlægges ved at mine de dataspor vi efterlader. Jo mere data, desto mere profitabel model.

Vægtningen af induktive metoder (måske hyppigt introduceret til historikere igennem teknikker såsom topic-modellering, hvor computeren selv finder emnerne i et tekstkorpus) afspejler, at databegrebet ofte kommer med positivistisk arvegods. Vi kan se dette afspejlet i lingoen. Åbner vi populære lærebøger i data science lærer vi, at data er "messy" eller "clean" alt efter hvor meget støj de indeholder.[15] Data kan "mines" eller "udvindes".[16] Data er et råmateriale, hvis bare vi kan "filtrere" og dermed rense det for alt, der ikke modelleres eller korreleres på en måde, der skaber værdi. Homogenisering eller dimensionalitetsreduktion er lig med en forøgelse af dataenes kvalitet, fordi det muliggør skabelsen af bedre modeller – også selvom de både i matematisk og konceptuel forstand medfører et

13 Se eks. globalhistorikeren Patrick Mannings diskussion af hvad han opfatter som vores fags data-drevne forpost. Manning: *Methods for Human History,* 165.

14 https://twitter.com/WalterScheidel/status/1645423395377479683?s=20 (08.05.2023).

15 McKinney: *Python for Data Analysis*, 12; VanderPlas: *Python Data Science Handbook,* 440.

16 Silge og Robinson: *Text Mining with R.*

tab af kompleksitet. Det er det rensede, atomiserede og kvantitative datapunkt, der kan aggregeres og derved muliggøre en pålidelig forudsigelse. Eller sagt på en anden måde: Det faktuelle opfattes som kontekstløst.

Som underviser i historisk metode, hører jeg ofte ekko af det materielle kildebegreb i dette sprog om data. Som med de fleste spøgelser er det selvfølgelig min egen projektion, men det er måske alligevel sigende. Med databegrebet følger forestillingen om at eliminere tendenser og lade dataene tale for sig selv. De bærer selv de sandheder, der defineres som værdifulde. Det undersøgende subjekt er ikke væsentligt. Forskellen er måske, at domæneviden for det positivistiske historiefag paradoksalt nok var afgørende for netop at identificere tendenserne. Det er det ikke for data science. Her er det derimod koden, der er ophøjet, fordi den gør processen reproducerbar og dermed særligt videnskabelig.

I denne afstøbning synes databegrebet at genintroducere (implicit i en teknikorienteret prosa) ideen om human- og samfundsvidenskab som skabelsen af et afkoblet videnskabeligt øje, der observerer; et *view from nowhere*. Kritikken imod disse positivistiske grundsten har været højlydt. Mange forskere har over de sidste 10-15 år påpeget, at data aldrig taler for sig selv; at de skabte data ikke afspejler en fuld virkelighed; at dette *view from nowhere* er dybt indlejret i partikulære kulturelle, økonomiske og materielle kontekster; at datavisualisering udjævner, synkroniserer og maskerer samtidig med, at det gør synligt.[17] Feministisk kritik har været særligt skarp og påpeget, hvordan datafikseret forskning har bygget på patriarkale, cis-maskuline totaliseringsfantasier: *Big Dick Data*.[18] På et helt elementært plan er der åbenlyst noget mærkeligt fallisk over forskeres stolthed over størrelsen på deres datasæt. Kritikere har også peget på, hvordan den postulerede induktion faktisk hviler på fejlagtige antagelser.[19] Når dataene ikke er skabt som forskningsdata, men derimod er det, der produceres af levende, *real-time* processer, er slutningen de facto ikke til en almengyldig regel, men til et sandsynliggjort udsagn om den konkrete proces. Det er med andre ord abduktion, ikke induktion. Andre har påpeget behovet for en genealogisk historisering af nutidens dataforståelse.[20] Fra et antropologisk udgangspunkt har forskere peget på, hvordan *machine learning*-algoritmerne er mest interessante, når de slår fejl – altså når de ikke skaber værdi ved at kunne forudsige, men derimod snubler i processen, fordi dataene indeholder noget, der ikke umiddelbart lader sig modellere.[21] Her finder den idiografiske erkendelsesinteresse åbenlyst noget ekstremt givtigt.

Der er således masser af tung akademisk modvægt til det positivistiske databegreb. Der er masser af alternative databegreber og situerede positioner, vi

17 Rettberg: "Ways of knowing with data visualizations".
18 D'Ignazio og Klein: "Numbers Don't Speak for Themselves".
19 Kitchin: "Big Data, new epistemologies and paradigm shifts".
20 Beer: "How should we do the history of Big Data?"
21 Munk, Olesen og Jacomy: "The Thick Machine".

som historikere kan spejle os i, når vi giver os i kast med at bruge computationelle tilgange. Det er ikke manglen på kvalificeret refleksion over databegrebet, der bekymrer mig. Problemet er, at disse positioner ikke slår bredt igennem. Den positivistiske *data science* har store penge i ryggen. Forlag som O'Reilly spytter velskrevne open access lærebøger ud, der finder veje til pensum alle steder. Techfirmaer som Google og Microsoft promoverer egne læringsværktøjer, både online og gennem samarbejder med arbejdspladser.[22] Og flere og flere algoritmer præger vores hverdag, som mennesker og forskere. Nyeste skud på stammen er pt. de meget omtalte store sprogmodeller, hvoraf ChatGPT er den klart bedst markedsførte. Her får vi dekontekstualiseret prosa præsenteret i et produkt, der med sit chat-interface og sin karikerede slaveagtige ærbødighed fremstår som designet til at skjule, at der ikke er tale om viden med proveniens, men om sandsynlighedsregning baseret på et ekstremt stort korpus af tekstdata. Ideen om data selv som det værdifulde fundament for viden (desto større, desto bedre) er således ikke kun akademisk: Det er også kommercielt. Enhver, der har prøvet at få ChatGPT til at fortælle noget om dansk historie er dog (forhåbentlig) blevet pinligt bevidst om, at modellen 'ved' meget mere om nogen ting end om andre.

Når vi som historikere taler om vores data og bruger dette ord med ukritisk selvfølgelighed, er det fordi data er et af tidens ord. Og det konnoterer en videnskabelighed, vi forestiller os, er attraktiv (profitabel, moderne, tværdisciplinær). Men det vi implicit siger er ofte ikke foreneligt med de epistemologiske modeller, vi som historikere har udviklet gennem langvarig og grundig refleksion. I en skarp artikel har arkivteoretikeren Devon Mordell eksempelvis peget på, hvordan der i arkivverdenen er et frembrusende *archives-as-big-data*-paradigme, der åbner døren på klem til den eller for længst opgivne ide om arkivet som neutralt og arkivaren som blot en kustode, fremfor en aktiv medskaber af viden.[23]

Helt afgørende er det måske, at vi slet ikke behøver dette databegreb – medmindre vi netop vil fortælle vores læsere eller studerende, at vi er i gang med at homogenisere og reducere materialets kompleksitet for at bygge en model. Vi har nemlig et andet begreb om empiri: kildebegrebet. Dette blev åbenlyst født ud af en tilsvarende positivistisk tænkning, men det er længe siden, at vi forstod det med positivismens grundantagelser. Det allestedsnærværende databegreb er materielt. Det er vores funktionelle kildebegreb ikke længere.

Kildebegrebet har i denne kontekst flere styrker:

1. Det peger i udgangspunktet på forskerens situerethed. En kilde er en kilde til noget, i kraft af, at vi stiller spørgsmål.

22 Se f.eks. Googles pralende retorik om deres rolle som aktør i uddannelsen af danskere https://grow.google/intl/dk (08.05.2023).

23 Mordell: "Critical Questions for Archives as (Big) Data", 150.

2. Denne rettethed spejles af kildens relation til den fortid eller samtid den fortæller om. Kilden har den samme "*live*" eller "*real-time*" karakter som datafetichisterne ofte synes er noget af det særlige og nye ved data i vores computationelle samtid. Selvfølgelig bliver materialet i de fleste tilfælde ikke til, imens vi sidder og tænker på dem (for os der studerer fortid i hvert fald). Men en kilde er affødt af virkelige processer, som den var en del af – og som præger den.
3. Kilder er derfor ikke født af vores erkendelsesinteresser, men det er vores aktive arbejde med kilderne, der får dem til at fortælle om de virkeligheder, der har frembragt dem. I denne kontekst kan det, som er "støj" for en model, være nøglen til indsigt en fortidig virkelighed. Støjen kan også levnsudnyttes. Nogle af de mest sofistikerede kildekritikker i historiefaget er netop dem, der har tænkt værdien af noget utroværdigt eller exceptionelt.
4. Kildebegrebet er tæt forbundet til en tænkning af arkivet, der har forladt tidligere forestillinger om arkivet som noget 'rent', der ventede på at åbenbare sine iboende sandheder. I stedet har vi udviklet en forståelse af arkivet som en historisk proces i sin egen ret: som præget af tendenser, som ufærdigt, som skævt og skævvridende. Kildebegrebet kommer dermed også med et fokus på tavsheder og udviskning.

Når databegrebet minder mig om den positivisme, vores disciplin har afsvoret, minder kildebegrebet mig således netop om, at det er muligt at have en levedygtig empirisk-funderet videnskab, der ikke forestiller sig materialet som et råmateriale, der skal mines. Kildebegrebet – i den version, der er modnet af utallige fagdisciplinære vendinger – er det bedre databegreb.

Det synes jeg ikke, vi skal glemme. Det digitale er kommet for at blive. Det rummer et væld af muligheder, vi skal gribe. Men blot fordi vi låner redskaber udefra, behøver vi ikke også at reproducere videnskabelige antagelser, der ikke er vores. Dem vi faktisk er trænede til at gøre os fejler ikke noget, bare fordi de er født analoge.

LITTERATUR

Anderson, Chris: "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" *Wired* 2008: https://www.wired.com/2008/06/pb-theory/ (08.05.2023).

Beer, David: "How should we do the history of Big Data", *Big Data & Society* 3 (1), 2016: https://doi.org/10.1177/2053951716646135

boyd, danah og Kate Crawford, "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society* 15 (5), 2012: 662–679. https://doi.org/10.1080/1369118X.2012.678878

Jensen, Helle Strandgaard: "Digital Archival Literacy for (all) Historians", *Media History* 27 (2), 2021 https://doi.org/10.1080/13688804.2020.1779047

Jensen, Max Odsbjerg, Josephine Møller Jensen, Victor Harbo Johnston, Alexander Ulrich Thygesen og Helle Strandgaard Jensen: "Scalable Reading of Structured Data", *Programming Historian*, 2022. https://doi.org/10.46430/phen0103

Hitchcock, Tim: "Confronting the Digital: Or How Academic History Writing Lost the Plot", *Cultural and Social History: The Journal of the Social History Society* 10 (1), 2013: 9–23 https://doi.org/10.2752/147800413X13515292098070

Kitchin, Rob: "Big Data, new epistemologies and paradigm shifts", *Big Data & Society* 1 (1), 2014.
Catherine D'Ignazio og Lauren Klein: "6. The Numbers Don't Speak for Themselves." 2020: https://data-feminism.mitpress.mit.edu/pub/czq9dfs5
Manning, Patrick: *Methods for Human History: Studying Social, Cultural, and Biological Evolution,* Palgrave: 2020.
McKinney, Wes: *Python for Data Analysis*, O'Reilly: 2013.
Mordell, Devon: "Critical Questions for Archives as (Big) Data", *Archivaria* 87, 2019.
Munk, Anders Kristian, Asger Gehrt Olesen og Mathieu Jacomy, "The Thick Machine: Anthropological AI between explanation and explication", *Big Data & Society* 9 (1), 2022. https://doi.org/10.1177/20539517211069891
Paasch, Marianne: *Gemt eller glemt?* Ph.d.-afhandling, Aalborg Universitet: 2018.
Putnam, Lara: "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast", *The American Historical Review* 121 (2), 2016, 337–402 https://doi.org/10.1093/ahr/121.2.377
Rettberg, Jill Walker: "Ways of knowing with data visualizations", i Martin Engebretsen og Helen Kennedy (red.): *Data Visualization in Society*, Amsterdam University Press: 2020, 35–48.
Robertson, Stephen: "The Properties of Digital History", *History and Theory* 61 (4), 2022, 86–106.
Romein, C. Annemieke m.fl.: "State of the Field: Digital History", *History: The Journal of the Historical Association* 105, 2020, 291–312 https://doi.org/10.1111/1468-229X.12969
Schriver, Astrid Ølgaard Christensen og Helle Strandgaard Jensen: "Arkivets digitalisering: en ny udfordring for historisk metode?" *Temp – Tidsskrift for Historie,* 13 (25), 2022, 5–27.
Silge, Julia og David Robinson, *Text Mining with R: A Tidy Approach,* O'Reilly: 2017.
VanderPlas, Jake: *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly: 2017.

JOHAN HEINSEN
LEKTOR I HISTORIE
INSTITUT FOR POLITIK OG SAMFUND, AALBORG UNIVERSITET
HEINSEN@DPS.AAU.DK