Big Data – et nyt humanvidenskabeligt forskningsfelt

af professor, dr.phil. Niels Ole Finnemann,
Det Informationsvidenskabelige Akademi, Københavns Universitet

Artiklen er en let bearbejdet udgave af forfatterens tiltrædelsesforelæsning som professor i Internetforskning, digitale medier og Digital Humanities ved Det Informationsvidenskabelige Akademi, Københavns Universitets Humanistiske Fakultet, den 21.11.2014.

Jeg vil i det følgende skitsere et forskningsprogram for studiet af digitale materialer. Det lyder lidt kedeligt. Det digitale hænger os, efterhånden som det breder sig alle vegne, ud af halsen. Lad os komme til sagen, indholdet, meningen i stedet for al den snak om mediet.

Det er helt efter bogen. Et godt medie er et medie, vi ikke lægger mærke til. Det må ikke forstyrre meningen. Sproget vil ikke ses, sagde den danske sprogforsker Louis Hjelmslev (1899-1965), hvad der nu ikke forhindrede ham i at studere det. Så kom den canadiske filosof Marshall McLuhan (1911-1980) og vendte det hele på hovedet. Det er mediet, der er meddelelsen. Den enkelte meddelelses indhold er uden betydning!

McLuhan har en pointe, men han trækker den helt ud af proportion. Det er rigtigt, at de medier, vi har til rådighed, betinger, hvad vi kan få at vide om verden. Med kikkerten kan vi få noget at vide, vi ikke kunne få at vide uden kikkerten, osv. Men viden er ikke slet og ret produktet af mediet. Viden er produktet af vores neurofysiologiske, mentale, kognitive og emotionelle kapaciteter, der også inkluderer evnen til at skabe eksternaliserede medier, der kan forstærke den ene eller anden menneskelige kapacitet. Det er der skrevet meget om, og jeg har også selv sat et par kommaer i den sammenhæng.

I det forskningsprogram, jeg præsenterer her, tager jeg afsæt i min tidligere forskning, hvor jeg blandt andet har beskæftiget mig med webmaterialer, arkiverede webmaterialer og andre typer digitale materialer med afsæt i en medieteoretisk forståelse af computeren. Jeg har i den forbindelse måttet lave lidt om på mediebegrebet (Finnemann 2014:300), og jeg arbejder i dag med at fusionere informations- og mediebegreberne, hvad jeg kommer lidt ind på sidst i denne artikel.

Det medieteoretiske perspektiv tjener dels til at placere computeren i den store mediehistorie, dels til at analysere hvorledes computerforståelsen udvikles historisk. Og lad mig starte med at resumere de tre vigtigste computerparadigmer med fokus på de egenskaber, der har direkte betydning for de vidensformer og kommunikationsmønstre, der manifesteres i digitale materialer.

1) Den klassiske forståelse af computeren som en regelstyret maskine, der håndterer velordnede data ved hjælp af et program. Fokus ligger på de processer, der foregår inde i maskinen, og maskinerne bliver primært programmeret af dataloger og andre IT-eksperter. Grundlaget for denne opfattelse blev lagt af den engelske matematiker Alan Turing (1912-1954), der i 1936 beskrev principperne for en universel computer. Denne opfattelse – med en række udbygninger – dominerer fra de tidligste år i 1930'rne til midten af 1980'erne og er helt dominerende i tolkningen af mainframe computere.
2) I kølvandet på fremkomsten af små bordcomputere fra sidst i 1970'erne udvikles en forståelse, hvor vægten forskydes fra det, der foregår inde i maskinen, til interaktionen mellem menneske og maskine, *Human Computer Interaction*. Der udvikles grafiske grænseflader, og der fokuseres på "brugere", der dog ofte er fageksperter på andre områder, og vi får en bred vifte af nye computerbaserede værktøjer med regneark, tegneprogrammer, tekstbehandlingsprogrammer osv. Det bliver muligt at bruge computerne til en stadigt voksende vifte af forskellige formål.

   Hvor hovedvægten tidligere lå på det regelbestemte, lægges den nu på hyper-tekstualitet, interaktivitet og multimodalitet. Man kan sige, at vægten glider fra Turings universelle computer til det, han i samme artikel i 1936 *en passant* kaldte valgmaskinen, hvor den videre proces beror på nye menneskelige input.
3) Med internettets gennembrud følger en tredje forståelse, der udvikles omkring netværksforbundne digitale medier. Fokus forskydes nu fra den enkelte computers indre systemlogik og brug over mod søgning, mønstergenkendelse, netværkskommunikation og netværksservicering af stadigt mere heterogene behov. IT kan ikke give mening i sig selv, og i takt med, at de digitale medier spredes ud i samfundet, bliver en hastigt voksende del af befolkningen aktive dataproducenter.

Vi kan lidt forenklet tale om en udvikling fra forståelsen af computerne som 'rule based' over 'tool based' til 'fool based'...! Det sidste paradigme tilfører, måske mere afgørende, også tre væsentlige, nye dimensioner i form af tre skalaer for trinløs variation, nemlig skalaen privat-offentlig, skalaen lokal-global, og skalaen hvem-med-hvem. De ældre forståelser forsvinder ikke, fordi der kommer nye, men de må vige pladsen som overordnede fortolkninger, idet de hver især kun fokuserer på nogle specifikke anvendelser. Der er derfor brug for en bredere, generel beskrivelse af, hvad der forstås ved digitale medier og materialer.

Digitale medier formidler digitale materialer: Mængder, bredder, dybder

Mit udgangspunkt er her (jfr. Finnemann 2014: 304ff.) følgende tre aspekter, der gør sig gældende i relation til alle digitale medier, uanset om det er *mainframes*, PC'er, tablets, mobiltelefoner eller mere specialiserede apparater:

- Digitale medier formidler altid digitale materialer.
- Digitale medier er altid søgemaskiner med et repertoire af søge- og navigationsmuligheder.

- Digitale medier medbringer altid et repertoire af metodiske muligheder for at kombinere, blande og præsentere digitale materialer, der selv er usynlige.

Disse tre dimensioner er altid til stede, men i skiftende blandinger. Vi ved fra mediehistorien, at der til hvert nyt medie også hører nye materialetyper, det være sig tekster, billeder, lydoptagelser, og i dag altså også digitale materialer. Digitale materialer omtales typisk som data, men termen data kan bruges i mange forskellige betydninger, der hver især er præget af en given faglig kontekst eller måske af den ene eller anden særlige forståelse af de digitale medier. Den slører med andre ord, at digitale materialer kan være meget forskelligartede.

I det følgende vil jeg derfor se nærmere på forskellige slags digitale materialer med det sigte at nå frem til nogle kriterier for beskrivelse af forskellige typer.

Der er mindst tre grunde til et sådant forskningsprogram: 1) Vi kan roligt regne med, at digitale materialer kommer til præge det 21. århundrede, 2) vi har behov for løbende at udvikle vores begreber om digitale materialer og om hvordan vi bruger dem overalt i samfundet og 3) jeg tror også, at vi af den vej kan skitsere nogle nye perspektiver for professionelle videns- og kulturformidlere i det 21. århundrede, hvad enten det er i regi af offentlige institutioner, der leverer public service til danskere, eller det er i regi af private virksomheder eller civile initiativer.

Jeg vil starte med at resumere det, jeg har kaldt en kernefortælling om det 21. århundrede:

1) En hastigt voksende del af samfundslivet vil komme til udtryk i stadig flere digitale genrer på digitale medieplatforme.
2) Vi får flere og mere forskelligartede digitale materialer, der spredes over stadig flere forskellige, mere skræddersyede digitale medier.
3) Vi får stadig flere forskellige søge- og præsentationsredskaber.

Et resultat af denne rimeligt sandsynlige prognose vil være en dramatisk vækst i informations-produktionen. Nogle kilder hævder, at mængden af information, der er produceret de sidste to-tre år, er større end den totale informationsmængde, der er blevet produceret i menneskehedens hidtidige historie. Denne vækstkurve vil efter alt at dømme fortsætte i de kommende årtier.

Det er os alle, eller i hvert fald et meget bredt – og voksende – udsnit af befolkningerne i mange lande, der producerer alle disse data. Det vidner om, at vi er i gang med et vidtgående kulturskifte, fremkaldt af de uendeligt mange forskellige, daglige mikroprocesser, vi hver især er involveret i. Vi er på en måde forståelsesmæssigt langt bagud for vores egen daglige praksis. Men denne praksis får vidtrækkende konsekvenser generelt og ikke mindst for de fag og institutioner, der står for vidensarkiverne og videns- og kulturformidlingen i det hele taget.

Det er ikke kun mængden, der er interessant. Det er også bredden og dybden. Lad os se på bredden. Digitale materialer rækker:

- Fra skanning af det ydre rum til kroppens indre;

- Fra regelstyret mainframe-computer via PC-redskaber til netværksforbundne digitale medier;
- Fra stedsbundne serviceinstitutioner til online services fra fjernoperationer i sundhedsvæsenet, til restrukturering af hele brancher som musik-, film- og bogbranchen og biblioteksvæsnet;
- Mobil opkobling overalt – fra intimsfæren over byrum til det yderste Thule;
- Nye genrer: nærsynkron dagligskrift over afstand (mail-chat-Facebook-Twitter);
- Fra velordnede datasæt til ikke-parametriske, vilde data (jf. nedenfor);
- Fra stedsbundne metoder til netværksbaserede analysemetoder;
- Tingenes internet;
- Fra 2D-print til 3D-print (ej end-of-print, men end-of-out-of-print);
- Nye forretningsmodeller og informationsmonopoldannelser;
- Fra en centralistisk medieoffentlighed med begrænset sæt af redaktionelle filtre til heteronomt redigerede offentligheder.

Dette felt strækker sig ret vidt, så langt øjet kan række og lidt længere. Fra det ydre rum til kroppens indre, til og med bevidsthedens realisering i vores neurofysiologiske system. Der imellem er der så også resten af samfundet og kulturen. Jeg tror, at dette billede er nok til at gøre det klart, at der er tale om meget forskelligartede typer af datamaterialer. Det betyder også, at vi ikke kan proppe alle disse data ned i en stor spand, som de kan passe ovre i datacentralen i fred og ro fra os andre. Men digitaliseringen går ikke kun i bredden, vi bygger samtidig hele tiden nye betydningslag ind i de digitale medier, og vi bygger løbende om på de digitale mediers funktionelle arkitektur. Det siger sig selv, at jeg ikke kan komme ind på alt dette i denne sammenhæng, eller rettere, det gør jeg indirekte ved at fokusere på det meget brogede billede af digitale materialer og også lidt på nogle af de analysemetoder, der er ved at udvikle sig specielt i relation til det, man kalder *Big Data*.

Det er langt fra alle disse datamængder, der er interessante, men nogle af dem er: I takt med at hele samfundet rykker ind på digitale medieplatforme, bliver for eksempel webmaterialer og andre digitalt fødte materialer et stadigt vigtigere, ofte unikt, historisk kildemateriale for det 21. århundredes politiske, kulturelle og sociale historie. Digitalisering af ældre materialer åbner samtidig nye muligheder for historisk orienteret forskning.

For virksomheder har nogle af disse mængder også stor betydning; det kan være data, de selv producerer, eller data, de får fra brugerne eller køber fra andre. Det er især her, man har talt om Big Data. De store datamængder kan ikke blot hjælpe eksisterende virksomheder i en række henseender; de kan også danne grundlag for nye virksomheder, måske baseret på nye forretningsmodeller. Selv om der er meget *hype*, der hurtigt forsvinder igen, gør datamængderne det ikke. De vokser.

Hertil kommer, at forskere i alle fag også producerer stærkt voksende datamængder. Både i EU og herhjemme er der fuld gang i opbygningen af forskningsinfrastrukturer omkring den slags kæmpestore databrønde. Men det foregår endnu uden noget reelt billede af, hvilke datamængder vi har, og hvad vi kan vente i de kommende år.

Endelig er nogle af disse datamængder også interessante for borgerne. Vi har krav og ønsker til online services, både som borgere og som privatpersoner.

Hvad enten man taler om Big Data eller ej, så bliver der stadig flere af dem, så spørgsmålet er, hvad skal vi stille op med dem? Og hvordan skal vi beskrive dem?

Big Data

Termen Big Data menes at være brugt første gang i 1997 af nogle forskere i NASA, der havde problemer med håndtering af store visuelle datamængder (G. Press, 2014). Det er dog først for nyligt, der er gjort forsøgt på at definere termen lidt mere. Lad os starte med en af autoriteterne i den slags spørgsmål, Oxford English Dictionary (OED). Ved Big Data forstås ifølge OED "data of a very large size, typically to the extent that its manipulation and management presents significant logistical challenges." OED anfører ældste kendte brug af "big data people" til 1980.

Den engelske Wikipedias definition er ikke meget anderledes: "an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications."

Det gennemgående punkt er "de store mængder", der overskrider eksisterende kapacitetsgrænser for håndtering. Det trækker tråde helt tilbage til de første computere, der havde en meget begrænset kapacitet, hvor data 'processeredes' serielt, en for en. Man talte i sin tid om 'von Neumanns flaskehals' efter arkitekturens ophavsmand. Men udover fokus på de store mængder, er de to definitioner mildt sagt noget vage i koderne. Hvis man tager dem bogstaveligt, bliver data til Big Data, når det bliver svært at styre håndteringen. Når dette problemet er løst, skal man så forstå det sådan, at det ikke længere Big Data?

Der findes ikke mere præcise definitioner end denne, men der findes forskellige tommelfingerregler, der bygges op omkring et udvalgt sæt af udbredte fællestræk. Nogle dataloger og IT-branchen taler meget om de tre, fire, fem eller seks V'er, der aggregerer en række hver især variable aspekter. Normandeau (2013) opsamler disse aspekter således:

Volume: man skal her især hæfte sig ved den eksponentielle vækst og ved det forhold, at mængderne allerede i dag er så store, at vi aldrig ville kunne analysere dem alene med traditionelle metoder.

Velocity: Der sigtes her især til enten realtid eller nær realtidsindsamling og brug, der dog kun gælder i nogle af de mange tilfælde, der typisk betragtes som Big Data.

Variety: Der sigtes her til det forhold, at en stor del af disse datamængder er "vilde" og ikke-parametriske, enten fordi de produceres på den måde, eller fordi man ønsker at kombinere forskellige datasæt, der er struktureret i ikke-kompatible parametre. Vi står med andre ord med alle mulige dataformater og blandinger (typisk webdata). Nogle kilder anslår, at 80 % af verdens data er "ustrukturerede" (eller ikke-parametriske/vilde).

Veracity. Her tænkes ikke mindst på problemer med urene datasæt og metoder til at kompensere for støj i kildematerialet. Hertil føjer man undertiden

Validity, der angår forholdet mellem datamaterialet og det formål, man har med at bruge det. Veracity og Validity synes at være ækvivalenter til samfundsvidenskabens begreber om reliabilitet og validitet.

Volatility, der refererer til datamaterialets tidsfølsomhed og mulige forældelse relativt i forhold til formålet.

Endelig anfører et syvende V om

Value: Der kan referere til økonomiske eller andre værdier, der udtrykkes i det givne formål.

Der findes ingen konsistent sammenkædning af disse kriterier og heller ikke klare grænser for hvert af dem. I bedste fald har man en art Wittgenstein'ske slægtsskabsrelationer, men alle disse relationer kan eksistere mellem alle former for digitale materialer. Når der f.eks. ikke er et klart kriterium for 'volume' (omfang) kan man ligeså godt spørge, om ikke ethvert omfang kan komme i betragtning? Hvis det ikke var Big Data i dag, så var det nok i går. Det er heller ikke alle projekter, der kombinerer 'variety' med 'velocity'. Tværtimod vil man ofte reducere varieteten af hensyn til hastigheden (og økonomien). Spørgsmålet om 'veracity' (pålidelighed) er heller ikke begrænset til at gælde Big Data, men derimod til specifikke analysemetoder, der også anvendes på små korpora og uafhængigt af de andre parametre.

Det, der imidlertid kan forsvares, er den påstand, jeg startede med, nemlig at vi står over for et spring i digitaliseringen, der indebærer en kolossal vækst i datamængder, der bliver indbyrdes mere forskellige, ligesom vi får et voksende register af nye metoder til at analysere og visualisere disse mængder. Til gengæld må vi regne med en kompetencemæssig specialisering omkring de enkelte metoder.

Det kan også forsvares at hævde, at Big Data-debatten rejser nogle videregående spørgsmål, der kan rettes mod alle typer data, store som små, homogene som heterogene, nemlig spørgsmålet om validitet, og hvad vi skal gøre af dem, og hvordan vi kan beskrive dem? Hvem skal bevare dem? Efter hvilke kriterier skal vi sortere og kassere? Hvem skal have adgang til dem? Hvordan skal der gives adgang, og hvilke søgeredskaber skal anvendes eller udvikles? Kan vi regne med, at Google passer på vore data? Eller Facebook? Eller de andre udbydere af såkaldte *cloud computing*-løsninger, og hvad med alle de mange materialer, der ikke er omfattet af Googles eller Facebooks datamonopoler? Hvad er rettidig omhu i denne sammenhæng? Og hvad med softwaren, vi bruger til at søge i og analysere med? Hvem tager sig af det? Alle disse spørgsmål står tilbage, når røgen efter Big Data-hypen har lagt sig, og verden kaster sig over den næste hype-bølge.

Jeg har hermed allerede også skitseret et relativt nyt, stort og voksende arbejdsfelt, der bare vil vokse og vokse i mange år. Men der henstår endnu et problem, hvis ikke vi kan karakterisere vore datamængder tilfredsstillende ud fra de nævnte V-parametre, hvordan kan vi så beskrive – og følgelig organisere – disse datamængder?

Det var ingen sag i forrige århundrede, hvor data slet og ret blev betragtet som et råmateriale, der blev ordnet i strukturerede databaser og håndteret med logiske eller computationelle procedurer. Det var dengang, data bare blev forstået som

et uspecificeret råstof, der kunne modelleres af forskeren: Det gælder stadigvæk i nogle sammenhænge, men det gælder ikke længere generelt.

De data, vi står med i dag, kan som oftest ikke beskrives på den måde. Og det giver anledning til nogle metodiske og epistemologiske overvejelser omkring 'big data'-tematikken, der ikke bliver voldsomt godt belyst i de definitioner, jeg lige har berørt. Derfor vil vi også lige se på V. Mayer-Schönberger & K. Cukiers beskrivelse (2013). De fremhæver fire punkter:

1) Mere – med tilføjelsen, at de store mængder træder i stedet for samples;
2) *Messy* – datasæt er uregelmæssige, støjfulde, heterogene;
3) Man finder korrelationer, men ikke forklaringer;
4) Dataficering – der i deres terminologi nærmest betyder *computation*, og som de adskiller fra digitalisering af tekst og billeder.

De betoner endvidere, at de store mængder kan analyseres på tværs og med andre formål end dem, der motiverede produktionen. Der indføres her en distinktion mellem de datasæt, der er 'messy', overfor datasæt, der er defineret parametrisk, som repræsentative samples (rensede for støj) eller afgrænsede enkeltcases. Ikke noget med stor og lille, men noget med støj og orden. Enkelthed versus kompleksitet. Tesen bag dette er, at det ikke gør noget med støj og uorden, fordi der er så mange flere data, der kan tages i betragtning, så man i virkeligheden får mange flere nuancer. Det er et argument, der også er fremført af forskerteamet bag Google Translate (Halevy, Nordvig, Pereira (2009)). Jeg kommer tilbage til den forskningsmæssige værdi nedenfor.

Det er dog ikke alle forskere, der følger det spor, Mayer-Schönberger (2013) og Halevy m.fl. (2009) plæderer for. Der er også mange, der stadig foretrækker at rense deres data for støj. Mens det i nogle tilfælde er transmissionstiden og den logistiske udfordring, der er støjkilden (= flaskehalsen), er det i andre tilfælde datainputtet – dvs. støjen i det digitaliserede råmateriale.

Det gælder for eksempel et dansk projekt, hvor man kortlægger undervandsrørledninger omkring olieboreplatforme. Her er kildematerialet fotooptagelser ned gennem et olieforurenet hav til en mudret havbund. For at se ledningerne, må man fjerne fotostøjen, der kan hidrøre fra en fiskestime, der passerer mellem kamera og havbund, eller fra andre ting, der er i vejen for identifikation af olierørene. Her bliver data med andre ord modelleret eller retoucheret – og pointen er, at det kan gøres med algoritmisk mekanik. Modellering af systemer er en klassisk operation, som jeg vender tilbage til, men her skal det blot bemærkes, at støjfjernelsen udskifter en usikkerhed med en anden, forhåbentlig mindre. Når reduktionen er sket, har det selvfølgelig begrænsende betydning for brugen af samme materiale til andre formål – hvilket er et af perspektiverne, der ellers knyttes til Big Data.

Når man renser for støj, definerer man selv, hvad der er støj relativt til det specifikke formål, som modellen skal repræsentere i et idealbillede. Det konkrete eksempel opfylder i øvrigt ikke kriteriet om realtidsdata og vil ikke kunne gøre det, fordi der kræves en efterbehandling af det benyttede fotomateriale.

Nu sidder der sikkert rundt om på humaniora en del kolleger og spørger, hvad det egentlig rager dem? Lad os tage et eksempel. Digitale kopier af ikke-digitale originaler må i dag være et område, hvor der kan bygges bro fra det 20. århundredes til det 21. århundredes humaniora. Der har i tidens løb været mange retrodigitaliseringsprojekter, og der er mange i gang. Pioneren på områder er Roberto Busa, en italiensk præst, der i 1946 begyndte at opbygge et digitalt index til søgning i Thomas Aquinas *Index Thomisticus* støttet af IBM. Det tog ca. 30 år, og frugten er i dag tilgængelig via www.

Lad os derefter kigge på et nyere retrodigitaliseringsprojekt, nemlig Googles grandiose projekt, der går ud på at lave digitale kopier af alle verdens trykte bøger.

Google Books

Google Books omfatter i dag måske 25 % af verdens trykte bøger. De stilles så vidt muligt – med respekt for rettighederne – til rådighed via *Google Library*. De er tilgængelige for enhver som enkeltværker, men de kan også danne grundlag for tværgående 'Big Data'-analyse bl.a. via en såkaldt "N-Gram Viewer", der crawler et givet korpus for søgetermer (fritekstsøgning) og præsenterer frekvenser af fundene på en tidslinje. Da materialet omfatter bøger fra 1500-tallet frem til i dag, er det blevet kaldt *long data*.

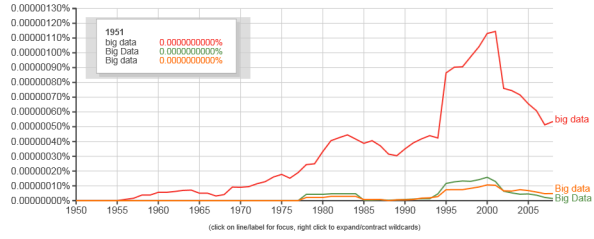
Et N-gram er en søgerutine, der søger på en given enhed i et korpus. N kan være et ord, et bogstav eller en anden enhed og have værdien 1 (ord), 2, 3 osv. Tanken har rod i Claude Shannons (1948) statistiske model for forudsigelse af sandsynligheden for en (sproglig) forekomst ud fra de foregående forekomster. Modellen kræver et enormt materiale af eksempler.

Det er ikke svært at se et enormt repertoire af ide-, begrebs- og begivenhedshistoriske søgninger, der kan foretages. Og det går lynhurtigt direkte fra lænestolen. Men der kan også laves mere grundige og detaljerede analyser hen over dette korpus. I bogen '*Uncharted*' (2013) præsenterer Erez Aiden og Jean Baptiste Michel (der har udviklet N-gram-vieweren sammen med Google) et N-gram-studie af de engelske stærke verber, der viser, at der en direkte korrelation mellem frekvens og henfald til svage former. Plausibelt, men hvis det stemmer med eksisterende forskning, er spørgsmålet, hvor meget mere vi har fået at vide? Og hvad nu, hvis dette resultat stred mod anden sprogforskning? Korrelationer er ikke nødvendigvis kausale. De kan være tilfældige og irrelevante. Men de kan også være fraværende.

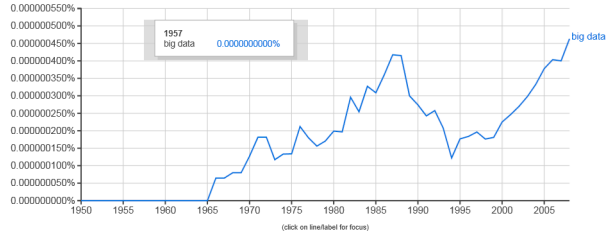
På modstående side ses et eksempel fra en lille øvelse, jeg lavede i foråret: De to kurver bevæger sig tydeligvist ude af *sync*. Det er svært at tro, at England halter ti år efter USA i den faglige interesse for Big Data, ligesom det er svært at forstå den amerikanske kurves fald op til 2008, da interessen for Big Data synes at eksplodere i netop de år. Her er altså et problem med fraværet af en korrelation, der burde være der.

Der er mange mulige forklaringer: Termen Big Data kan have været brugt med forskelig betydning enten i den givne periode eller i de to lande eller kombinationer heraf. Eller der har været andre termer brugt om samme fænomen i det ene korpus. Det kunne være 'massive data', der i øvrigt viser sig at være brugt langt

American English, 1950-2008



British English, 1950-2008



oftere og over længere tid i begge korpora. Det kan også tænkes, at de to korpora er uens sammensat, f.eks. i henseende til de fagområder og populærvidenskabelige udgivelser, hvor termen har været brugt. Og man kan endelig heller ikke afvise, at fortolkeren mangler baggrundsviden, det være sig om fænomenet, om sprogbrug, om brugen af bøgerne i de to samfund eller om andre kulturforskelle, der måtte kunne spille ind. At man på et tidspunkt måske har digitaliseret 'alle bøger' betyder ikke, at man har al den viden, disse samfund besad på det givne tidspunkt.

I en artikel i tidsskriftet *Big Data and Society* (19.3.2014) skriver Roger Burrows & Mike Savage, at der ifølge *Google Trends* i 2007 så godt som ikke var nogen, der søgte på termen "big data". Det sker først i 2010 og med en eksplosion i 2011. Alt efter, hvilket korpus vi vælger, når vi med andre ord forskellige konklusioner. Forestillingen om, at termen opstod i 1997, stemmer i øvrigt ikke med Google Books, men bygger måske på Google Trends, der ikke har mange søgeord før 1998.

Både fraværende og tilstedeværende korrelationer kan hjælpe til at rejse nogle spørgsmål om disse korpusspørgsmål, der viser, at det er vigtigt at have fagkyndig viden både om boghistorien og om det specifikke korpus, samt om terminologihistorien, ligesom viden om, hvorvidt og hvilke kulturforskelle, der måtte være mellem USA og England, er af relevans i *fortolkningen* af fundene.

Hovsa! Jeg sagde et af de forbudte ord; *fortolkning*. Det var ikke en lapsus. Tværtimod, de digitale datamaterialer kræver i den grad kritisk fortolkning. Man siger somme tider, at tallene taler, men hvis der overhovedet er tal, så taler de ikke deres eget sprog. De adlyder derimod Gödels teorem, der siger, at et formelt system ikke kan definere alle præmisser for systemet. Noget må være givet forud og udenfor. Her starter fortolkningen.

I en artikel fra 2012 peger Danah Boyd og Kate Crawford på en række fortolknings-momenter i relation til Big Data:

1) Alle forskere eller fagdiscipliner har et begreb om data, men vi har ikke det samme begreb; vi fortolker med andre ord gennem selektionen af de aspekter, vi lægger vægt på. Mange kender de fortløbende diskussioner af, hvad der er data, information og viden. Boyd og Crawford bygger her videre på et forarbejde til Gitelman (2013).
2) Fortolkning knytter sig også til den støjrensning, der sker i udvælgelsen af et korpus i relation til et eller andet undersøgelsesformål.

   I havbundseksemplet fra før: Fotomaterialet beror på en selektiv modellering af havbunden, hvor nogle træk defineres som støj (via optiske valg af f.eks. linse og lyssætning) og fravælges mere eller mindre ubevidst. I anden omgang renses også fotomaterialet. Begge dele sker relativt til et ønsket idealbillede. Det sker nogle gange i forhold til eksplicitte modeller, men det kan også være implicitte referencer f.eks. til en sprogopfattelse, som vi skal se i forhold til Google Translate.
3) Der er ligeledes fortolkning impliceret i afgørelsen af, hvad der skal registreres og blive til data, og i afgørelsen af, hvor stort datamaterialet skal være. Selv om man definerer "alle data" som 'alle bøger', er der tale om et *biased* korpus målt i forhold til f.eks. 'al menneskelig viden' eller i forhold til 'alle manuskripter' eller i forhold til bøgernes position i de forskellige landes kultur gennem tiden. At vurdere dette bias kræver viden om, hvilke bøger om hvilke emner, der blev trykt i de enkelte lande i de 500 år. Og om hvilken betydning bøger havde i datidens videns- og kulturformidling i de forskellige samfund?
4) Man bruger ofte modeller af noget, der repræsenteres, men forholdet mellem modellen og det, den skal repræsentere, er også baseret på en fortolkning af, hvad der skal registreres som data, og hvilke elementer, der indgå i modellens opbygning. At det er fortolkning, betyder ikke, at alle fortolkninger er lige meget værd, at vi er i den rene subjektivitet. Tværtimod, den mest overbevisende fortolkning er den bedste. Men her fanger bordet også i naturvidenskaben, for den mest overbevisende fortolkning kan, uanset hvor mange kriterier, man inddrager, aldrig blive mere end den bedste fortolkning på de kendte præmisser. Isaac Newtons naturlove er ikke længere, hvad de var, eller med Albert Einsteins ord: "As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality" (Einstein 1921).
5) Domspræmissen er med andre ord *hermeneutisk*, selv om der i nogle fagmiljøer måske stadig hersker en

forestilling om absolut neutralitet, objektivitet og facts, der ikke er fabrikeret.

6) Der er yderligere et punkt, hvor fortolkning er uomgængelig, nemlig om datamaterialet har et omfang eller en repræsentativitet, der giver mulighed for valide slutninger.

   Det er f.eks. ikke umuligt at regne baglæns fra et ønsket resultat til en model, der kan bruges som afsæt for at nå dette resultat.

   Statistiske analyser er korrelationsanalyser, der kan være mere eller mindre signifikante, men uanset den statistiske signifikans af en relation kan den ikke bruges til at etablere kausalitet og forklaring.

   På den anden side vil jeg også understrege, at man ikke bare kan tage den tyske filosof Hans-Georg Gadamers (1900-2002) hermeneutik og lægge ned over digitale materialer. De respekterer nemlig langt fra altid Gadamers eller andre af det 20. århundredes værkbegreber.

Trods de metodiske spørgsmål, der kan rejses omkring den statistiske analyse, tilbyder Google Books nye søge- og analysemetoder, der bør modtages med kyshånd af trængte humanister, selvom der nu bliver tale om softwarestøttede analysemetoder. Vi må i hvert fald konstatere, at humaniora her udfordres, fordi andre faglige tilgange begynder at gøre sig gældende for kildematerialer, vi før har haft for os selv. Det bør hilses velkommen, og hvorfor ikke betragte dem som humanvidenskabelige forskere, der beriger vores domæne?

Men det er og bliver samtidig kun *en ny bro til fortidens kilder*. Det vil sige et eksempel på digitalisering af ikke-digitale originaler. Lad os gå et skridt videre og se på et beslægtet, men alligevel anderledes eksempel. Igen fra Google.

Google Translate

Som Google Books er Google Translate baseret på retrodigitaliserede kopier af ikke-digitalt sprogmateriale, der analyseres, ligesom N-gram-vieweren kan bruges til at analysere et korpus.

Google Translate bygger grundlæggende på et korpus af allerede oversatte tekster, hvor teksten altså foreligger i to sprog. Hvis dette korpus er stort nok, kan man via analyser af statistiske sammenhænge mellem ordforekomster også generere 'oversættelser' af nye tekster med en relativt stor sandsynlighed – i hvert fald i teorien.

Ifølge Alon Halevy, Peter Nordvig og Fernando Pereira (2009) bygger man for en stor del på internationale oversættelseskorpora som f.eks. EU's dokumenter, der foreligger i mange forskellige sprogdragter. Der er altså et klart bias mod formelt, officielt sprog. Andre dele er basereret på tekster fra www, som det siges, og som Google jo høster i massiv skala. Man må formode, at der også er en del materiale fra de sociale medier. Meget tyder på, at kortformer her vinder frem, at ordforråd måske forandrer sig hurtigere, og at den sproglige del af kommunikationen er mere integreret med *emoticons*, *likes*, *links* og *scripts* og andre visuelle supplementer. Om det gør forskel, at noget er produceret som trykt tekst eller er født digtaliseret tekst, som er indlejret i tekstbehandlingsprogram og www, ved vi ikke.

Men vi ved to ting om Google Translate: 1) Der kommer ofte nogle stærkt misvisende oversættelser som resultat

– mange sprogforskere rejser rundt med sådanne eksempler – og 2) vi har ofte stor nytte af dette redskab, selv om det er upræcist og mangelfuldt.

Der er imidlertid en tredje dimension, der også må tages i betragtning. For Google Translate udvikles også løbende ved hjælp af frivillige og ufrivillige bidrag fra *crowdsourcing*. Hvordan dette materiale indarbejdes, ved jeg ikke, jeg kender heller ikke dets omfang, men det peger på et metodologisk potentiale, der ligger og venter på at blive prøvet af humanistiske forskere. Hvorvidt Google Translate vil blive væsentligt forbedret med crowdsourcing over tid må vise sig. Hvis befolkningsunderlaget er stort nok, kan det sikkert få betydning (jfr. forskellen på Wikipedias nationale varianter), og det vil i alle tilfælde kunne bidrage til den løbende ajourføring i forhold til sprogbrugs udvikling.

Er Google Translate bygget på sprogforskning? Selv mener Google ja (Alon Halevy et al. (2009)). Man bygger på 'memorering' af rigtig mange fraser frem for generelle regler. Erfaringen er, at simple modeller med mange data er bedre end komplekse modeller med færre data. Data hentes især via webkilder (men som nævnt også EU-korpora og andre tekstkorpora, der foreligger i flere sprog). Som begrundelse for denne fremgangsmåde, henvises der på den ene side til en erfaring, der siger, at et materiale på måske en million eksempler oftest er nok, selv om antallet af sætninger er ubestemmeligt stort, mens man på den anden side antager, at sprogets kompleksitet er for stor til regelbaseret analyse. Nye ord dannes hele tiden, og ordenes betydning varierer. Ingen regel uden undtagelse – heller ikke i grammatik og syntaks. Man skal således heller ikke, siger Harvey et al., smide de sære undtagelser væk: Det, der er sært i den individuelle sprogbrug, er ofte hyppigt i det store korpus.

Det er sværere at formalisere almensproget end det formelle sprog. Almensproget er ganske enkelt mere komplekst. Mens det endnu ikke er lykkedes at indbygge sprogets regler i tekstbehandlingen, leveres de finitte regneregler uden videre i færdigpakkede regneark.

Personligt tror jeg, man må give Googleforskerne ret i, at sproget i sidste ende er uregerligt, at regelbestemte oversættelsesprogrammer a la Noam Chomskys transformationsgrammatik aldrig vil komme til at fungere. Det hænger i mine øjne sammen med, at ethvert medie, inklusive tale- og skriftsprogene, først og fremmest skal formidle et indhold, der er distinktivt og netop ikke forudsigeligt, for hvorfor skulle vi så høre efter? Hvis sproget var helt regelbestemt, ville alle sproglige udtryk være regelbestemte fra begyndelsen til enden, og vi ville ikke kunne bruge sproget til at meddele noget, der ikke er bestemt af systemet.

Der er således en slags skjult indre og nødvendig sammenhæng mellem støj, undtagelsen fra regler og muligheden for meningsfuld artikulation. I den forstand har Googleforskerne fat i en lang ende. Jeg tror også, de har ret i, at det store sprogmateriale giver mulighed for at bringe oversættelser af mere sjældent forekommende ytringer. Det er det mønster, vi i forvejen kender som "the power law of distribution" – eller som "the long tail effekt": Den spredte brug af mange forskellige ord er ofte tilsammen mere omfattende end den fælles brug, vi gør af de samme

få ord. De sjældne ord er også ofte mere informationsrige end de almindelige. Det er denne lange hale, der systematisk tabes ved brugen af samples.

Brugen af "messy" data i store mængder ('alle de data, der kan fås') kan altså i nogle tilfælde måske fange mere differentierede, spredte fordelingsmønstre end samples eller cases. Den har imidlertid også sine begrænsninger, for Google Translate vil aldrig kunne fortælle os, hvorfor den oversætter, som den gør. Vi ved således allerede fra 1980'ernes eksperimenter med neurale netværk, at systemer, der trænes med eksempler, somme tider kan håndtere et materiale selv, men uden at vi selv kan lære af det. Vi får med andre ord ikke nogen viden om sproget eller nogen mulighed for validering af oversættelsen med Google Translate. Den må vi selv tjekke. Heller ikke Google får en sådan viden; de får bare en statistik, der måske forbedres, så længe den bliver fodret med nyt materiale indtil et punkt, hvor det måske går ned af bakke igen, som det muligvis er tilfældet for *Google Flu Predictor* efter otte års succes (Lazer et al. 2014).

Webmaterialer

Mit tredje eksempel i dag, webmaterialer, er et fuldblods født digitalt materiale.

Webmaterialer studeres i dag ud fra mange vinkler, men det bør også blive et primært kildemateriale for historikere, lingvister, litterater, kunsthistorikere, etnografer, kulturforskere og medieforskere af enhver art, samt for enhver, der vil vide noget om i går, i dag og i morgen.

Derudover er www imidlertid også allerede et uomgængeligt forskningsredskab, der vil få voksende betydning i takt med, at forskningen bliver mere netværksbaseret og international. Her er materialerne og metoderne tæt sammenflettet: Det er to aspekter af det samme indeholdt i hinanden. Det skyldes i sidste ende, at webmaterialer, ligesom alle digitale materialer, udelukkende og altid må tilgås via en eller anden søgemetode. Vi kan ikke se dem: De skal opsøges fra serveren eller harddisken og fremvises på en eller anden måde. Allerede derfor er webmaterialer og andre digitale materialer forskellige fra trykte materialer og skrevne materialer. Men de har også en række andre, specifikke træk. Lad os starte med forskellen på digitale kopier af trykte tekster og født digtalt materiale, der også i høj grad er tekster.

Både digitale kopier og digitalt fødte materialer er normalt indlejret i hypertekstuelle, interaktive og multimodale kontekster, og de kan håndteres med *scripts*, *tools* osv. Digitalt fødte materialer kan imidlertid også indeholde disse features som del af materialet, da de indgår i digitale mediers grammatiske repertoire. Også programmer og scripts kan være del af datamaterialet. Med internettet og de netværksforbundne digitale medier bliver digitale materialer yderligere trinløst variable i henseende til lokal-global, privat-offentlig, hvem-med-hvem (Brügger og Finnemann 2013).

Det distinktive kriterium er, at hypertekstuelle, interaktive features og indholdet af scripts kan være del af den digitale original. Disse "grammatiske features" danner grundlag for en lang række genremæssige nydannelser, der fører til en stadigt mere heterogen mængde af forskellige digitale materialer og genrer. Det gælder både i forhold til:

- Typer af forfatterskab
- Filformater
- Genrer
- Redaktionelle standarder
- Viral spredning
- Indlejring og *remix*
- *Expressions* 'given off'/*tracking data*

Endelig kan digitale medier være følsomme over for og respondere på individuelle meddelelser. Som konsekvens heraf ser vi også nye blandingsformer, hvor redaktionelle, censurerende instanser realiseres i realtid, f.eks. i form af Facebooks filtrering af indlæg i borgernes nyhedsstrøm. For forfatterskabers vedkommende kan der (jfr. Finnemann 2014a) anføres en lang række nye eller mere fuldt udfoldede former som:

- Personlige profiler, der over tid kan aggregeres som en form for selv-biograferende logning;
- *Statement*-kultur i nærsynkron skrift via chatfora, debatfora, Facebook- og Twitter-tråde m.v.;
- Uafsluttede tekster (f.eks. bøger, der fortsætter på blogs);
- Kollektive forfatterskaber (Wikipedia og andre);
- Computergenerede tekster (der er personligt genererede, men trækker materiale fra mange forskellige ressourcer på anfordring);
- Viral remix, indlejring af andres meddelelser i nye meddelelser;
- Kommercielle 'medforfattere' og redaktører (Facebook redigerer rækkefølger og skubber til folk baseret på aktivitetsovervågning);
- Personalisering af services;
- Efterladte spor, der indsamles og aggregeres af andre via *tracking software*, bl.a. til profilering.

Det er ikke hensigten at gå nærmere ind på disse former her, men blot at beskrive konturer af det heterogene materiale, der hober sig op omkring os. Vi kan imidlertid roligt konkludere, at webmaterialer er 'vilde' og ikke-parametriske, på samme måde som almensproget, sammenholdt med det formelle sprog, er 'vildt' og ikke-parametrisk. Webmaterialer hører til blandt de mest komplekse sæt af datamaterialer, vi har, men de forsvinder hurtigt og skal arkiveres i realtid 'as is'. Imidlertid er der også mange andre typer af digitale materialer. Der er grund til også at nævne det dybe web (der enten er afspærret med password eller sikkerhedsmure), som omfatter både forsknings-, virksomheds- og myndighedsressourcer.

Digitale materialer deler sig efter proveniens og mediebestemte egenskaber

Jeg har prøvet at vise, at datamaterialer på den ene side altid er præget af deres proveniens og formålet med deres produktion samt på den anden side af de digitale mediers særlige egenskaber. Jeg har samtidig argumenteret for, at der i forståelsen af datamaterialer også altid er en faglig dimension, der udspringer af det sted, det hjørne af naturen, hvori digitaliseringen tager sit afsæt og i det det hjørne, hvor den fører hen.

Jeg har endvidere søgt at pege på de digitale materialers betydning som kildemateriale og antydet det store ubenyttede metodiske repertoire og skitseret nogle kritiske indfaldsvinkler. En del af denne argumentation retter sig mod vores generelle forståelse af digitale medier og deres betydning, og en del retter sig specielt mod humaniora, men

jeg ser også et stærkt voksende behov for at udvikle en vifte af mere specialiserede og forbundne fagkompetencer, der skal til for at tage vare på en kvalificeret, professionel videns- og kulturformidling baseret på de hastigt voksende mængder af heterogene digitale materialer, der tegner til at blive det dominerende kulturgrundlag i det 21. århundrede. Vi skal måske endda gå et skridt videre og ikke bare flytte centrum fra bogen til netforbundne PC'er, tablets og køleskabe, men tage afsæt i den mobile søgemaskine, der i dag er vores mest intime ejendel.

Men der er også grund til at understrege, at der er god plads til kritiske forskere, der kan reflektere over og analysere tvetydighederne og bidrage til at uddybe analysen af de drivkræfter, der gør sig gældende. Herunder også de helt åbenlyse og store spørgsmål om overvågning, sikkerhed, persondatabeskyttelse, myndighedsmisbrug, og privat monopolisering af vidensressourcer og forskningsetik. På den anden side skal vi måske passe på, at vi ikke melder os ind i bekymringsindustrien med alt for mange meninger, der er dannet, før vi har haft fingrene nede i de digitale materialer og bevæget os helt ind i det digitale felt.

Data strømmer over alle grænser – også dem mellem fakulteterne

Digitale materialer produceres af mennesker, der bruger digitale medier, der også er menneskeligt frembragte kulturprodukter. De udgør som sådan i al deres forskellighed et genuint humanvidenskabeligt forskningsfelt. Her finder vi bare ikke ret mange humanister. Man må vel mene, at digitale medier befinder sig hinsides det alleryderste Thule, udenfor civilisationen, helt ovre i naturvidenskabens rige, mineralriget måske? Hvordan kan det gå til?

Et bud kunne være, at humanvidenskaben er blevet fanget i sine rødder i den filosofiske idealisme, hvor det fysisk-materielle, og derfor også vores medier og teknologier, befinder sig uden for den ideale verdens synsfelt. Vi kender faghierarki-lagkagen: I bunden fysikken, så biologien, så psykologien og så sprog, litteratur, historie, samfund og kultur. Verden som en lagkage, hvor man i fred og ro kan beskæftige sig med sit eget lag. I det 20. århundrede flyttede man ganske vist *res cogitans*, den tænkende substans, ind i *res extensa*, i tidens og rummets udstrakte verden, men man opretholdt ideen om separate lag. En af pionererne i computerforskningen, John von Neumann, talte f.eks. om psykofysisk parallelisme: Det psykiske og fysiske var to adskilte lag uden kontinuerlig interaktion. Det gjorde han, skønt han var en af pionererne i udviklingen af de digitale medier, hvor alle symboler flyder rundt i en redigerbar, elektromagnetisk form.

Det er rigtigt, at humaniora bliver mobbet af andre, men spørgsmålet er, om vi ikke også har muret os selv inde i vores eget domæne og gjort os relativt uvedkommende for de andre i tillid til det 20. århundredes dannelsesbegreber, der var forankret i det 20. århundredes medier?

Hvorom alting er: De datastrømme, der i dag rækker fra det ydre kosmos til kroppens indre, er kulturelle strømme, og de strømmer i dag med en sådan kraft, at det 20. århundredes lagkagemodel bryder sammen. Vi er på vej fra den psykofysiske lagkagemodel til hvad vi kan kalde psykofysisk interaktionisme.

Jeg støtter mig her også til en argumentation, jeg låner af filosoffen Hans Fink (2006), der plæderer for, at vi forlader det lille naturbegreb (eller rettere en hel række natur-som-afgrænsning-forestillinger), der sætter kulturen uden for og i modsætning til naturen, til fordel for det store *Naturbegreb*, der indbefatter det hele – til og med kulturen inklusive både Hitler og Stalin, Cykelslangen i Københavns Havn og Beethovens musik.

Det lyder nemt, men det er en ganske besværlig proces, fordi dikotomien natur-kultur er så dybt indlejret i vores tankemønstre. Ved at holde sig til natur-kultur-distinktionen, har humaniora forbudt sig selv at forholde sig til vores biologiske og fysiske eksistensvilkår og fortrængt medierne ud i periferien. Fakultetsgrænsen er blevet en intellektuel spærrebom. Lad os følge datastrømmene, hvor de er. Dér er nemlig også kulturen og det mere menneskelige.

Mens forskningen i stigende grad får sin magt og legitimitet for det, den påstår at vide, er forskningen selv i sidste ende funderet i forholdet til det, den ikke ved. De helt store udfordringer ligger her måske ikke så meget i de enkelte fagvidenskaber, som i det forhold, at de alle handler om et og samme univers. Et skridt kunne være at indrette teorierne derefter. Man kan måske også formode, at jo flere data vi får, jo mere går det i den retning.

Teorien. En fusion af informations- og mediebegreberne

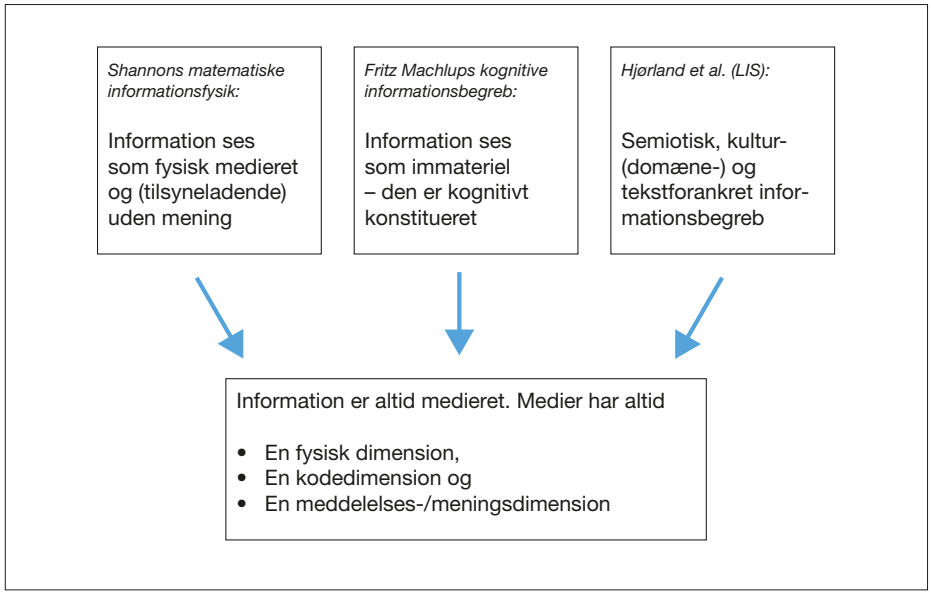
Min teoretiske tilgang til denne indledende analyse af digitale materialer tager afsæt i et forsøg på at fusionere vore medie- og informationsbegreber.

Der er efterhånden en solid tradition for at betragte informationsbegrebet som kontekstuelt betinget eller forankret. Men information er også altid fysisk medieret på en mediespecifik måde. Mediet udgør en særskilt del af konteksten, dels fordi medier rækker ud over tidens og stedets enhed (den situerede kontekst), dels fordi de enkelte medier har forskellige egenskaber, der har konsekvenser for, hvad der kan udtrykkes, hvordan det kan udtrykkes, og hvor det kan udtrykkes. Også når det gælder viden.

Man kan ikke sige, at informations- og mediebegreberne har rendt hinanden på dørene. Men lad os prøve at sætte dem sammen (se modellen på næste side):

Fysisk udtryk: Medieringen mellem det fysiske udtryk og det meningsmæssige indhold. Alle meddelelser er manifesteret i fysisk form. Der er altid og uden undtagelse en mediering mellem materien og bevidstheden. Fra Claude Shannon hentes begreberne om fysisk støj, støj i koden og indholdsmæssige støj samt begrebet om redundans som meningsstabiliserende feature.

Kodens sociale bånd. Medieringen mellem dem, der kommunikerer: mennesker indbyrdes. Mellem bevidstheder. Den fysiske form kan ikke selv garantere betydningen. Al betydningsdannelse forudsætter et fælles kodesystem for betydningsgenkendelse. Kognitiv individualisme går ikke. Fra Fritz Machlup (kognitive paradigmer) hentes den mentale kodning, men tilføjet en social dimension. Tanken kan måske være individuel, men sproget = mediet og koden er social.



Meddelelsens indhold/mening. Medieringen mellem mennesker og vores ideer/forestillinger om verden. Det er meddelelsens indhold, der giver mediet sin mening, men mediet præger samtidig de mulige udtryksformer for mening. Fra Hjørland et al. (semiotik) henter jeg, at information har en kulturel dimension, med den tilføjelse, at den kulturelle dimension også indbefatter mediet, eller den samlede mediematrice, hvori meddelelsen er virksom. Den kulturelt specifikke dimension kan undertiden aflæses gennem mediet, men oftere gennem genren, og også ofte gennem den enkelte meddelelses indhold.

Digitale materialer er altid forbundet med støj på alle tre niveauer. Der er fysisk støj, der er kodestøj, og der er støj på informationsplanet. Ved denne kobling kan vi knytte de mediebestemte og medie-specifikke egenskaber – der undertiden overlapper – direkte til de digitale materialer. Vi kan identificere støjkilderne mere præcist. Vi kan især gruppere digitale materialer på en meningsfuld måde relativt til mulige metoder og søgeredskaber.

Drivkræfter

Det er ikke IT, der driver udviklingen. Hvis det var tilfældet, ville hele verden allerede være pakket ned i en enkelt database, der blev håndteret af et generelt logisk programmeringssprog.

Det er samfundets behov, der driver IT-udviklingen, der derfor tegner sig som en historie om stadigt flere forskellige former for datamaterialer. Der er stadigt flere forskellige genrer, der artikuleres på stadigt flere forskellige medieplatforme og skræddersyede *devices*, der håndteres

ved hjælp af stadigt flere forskellige typer søge-, analyse- og fremvisningssoftware.

Digitaliseringens historie er et resultat af, at stadig flere forskellige aktører udtrykker stadig flere forskellige behov i stadigt flere genrer på stadigt flere forskellige digitale medieplatforme.

Der er tre grundlæggende kilder:

Verdens forskelligheder: IT udvikles på stadigt flere områder fra kosmos til kroppens indre og alt der imellem – herunder kultur og samfund.

Aktørernes forskelligheder: IT udvikles af stadigt flere forskellige faglige ekspertiser og en stadigt voksende kreds af civile.

De menneskelige behovs forskellighed: Fra den eksponentielt voksende produktion af viden til den moderne, veluddannede middelklasses umættelige behov for underholdning – eller er det i sidste ende ...:

Referencer

Aiden, E. & Michel, J.B. 2013. *Uncharted. Big Data as a Lens on Human Culture*. New York: Riverhead Books.

Boyd D. & Crawford, K. 2012. Critical Questions for Big Data. *Information, Communication & Society.* 15:5, 662-679, DOI <dx.doi.0rg/10.1080/1369118x.2012.678878>.

Brügger, N., & Finnemann, N.O. 2013. The Web and Digital Humanities: Theoretical and Methodological Concerns. *Journal of Broadcasting and Electronic Media*, *57*(1), 66-80. 10.1080/08838151.2012.761699

Burrows, R. & Savage, M. 2014. After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology. Commentary. *Big Data and Society*, 2014 1: <bds.sagepub.com/content/1/1/2053951714540280> DOI: 10.1177/2053951714540280.

Busa, R. et al. 2003. *Index Thomisticus* <www.corpusthomisticum.org>. Tilgået 20.11 2014.

Cox, M. and Ellsworth, D. 1997. Application-controlled demand paging for out-of-core visualization. In *Proceedings of the 8th conference on Visualization '97* (VIS '97), Roni Yagel and Hans Hagen (Eds.). IEEE Computer Society Press, Los Alamitos, CA, USA, 235ff.

Einstein, A. 1921. Geometry and experience. Tale i det Preussiske akademi 27.1 1921. Citeret efter Engelsk oversættelse p. 9: <todayinsci.com/E/Einstein_Albert/EinsteinAlbert-MathematicsAndReality.htm>.

Fink, H. 2006. Three Sorts of Naturalism. *European Journal of Philosophy* 14:2 ISSN 0966-8373 pp. 202–221

Finnemann, N.O. 2014. Digitization: New trajectories of mediatization? In K. Lundby (Ed.), *Mediatization of Communication.* (pp. 297-321). Berlin: Mouton de Gruyter. (Handbooks of Communication Science; No. 1, Vol. 21).

Finnemann, N.O. 2014a. Digital Humanities and networked digital media. *Mediekultur* vol. 30, No. 57 – Special Issue on Digital Humanities.

Gitelman, L. (ed.). 2013. *Raw Data is an Oxymoron*. Cambridge Mass: MIT press.

Google Books & Library. <www.google.com/googlebooks/library>. Tilgået 20.11 2014.

Google Flu Trends. <www.google.org/flutrends>. Tilgået 20.11 2014.

Google Translate. <translate.google.com>. Tilgået 20.11 2014.

Google Trends. <www.google.com/trends/?hl=en-GB> eller <www.google.dk/trends/explore>. Tilgået 20.11 2014

Halevy, A. Nordvig, P. Pereira, F. 2009. *The unreasonable Effectiveness of Data.* IEEE Computer Socierty. 2009: 1541-1672/09.

Hjørland, B. manuscript, u.å. *Theoretical development of information science: A Brief history.* Copenhagen: IVA.

Lazer, D. Kennedy, R. King, G. & Vespiagnani A. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science vol. 343*. Policy Forum. 14.3 2014.

Levi, Amalia S. 2013. *Humanities 'Big Data'. Myths, challenges, and lessons.* IEEE International Conference on Big Data. <bighumanities.files.wordpress.com/2013/09/2_5_levi_paper.pdf>. Tilgået 20.11 2014.

Machlup, F. & Mansfield, U. (Eds.). 1983. *The Study of Information*. New York: John Wiley & Sons.

Mayer-Schönberger, V. & Cukier, K. 2013. *Big Data. A revolution That Will Transform how We Live, Work and Think*. London: John Murray.

Normandeau, K. 2013. *Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity.* <insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity>. Tilgået 20.11 2014

OED. *Oxford English Dictionary* <www.oed.com/view/Entry/18833#eid301162178>. Tilgået 20.11 2014.

Press, G. 2014. 12 Big Data Definitions: What's Yours? *Forbes* 9 March 2014. <www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours>. Tilgået 20.11 2014.

Shannon, C. & Weawer, W.1949. *The Mathematical Theory of Information.* Urbana: University of Illinois Press.

Turing, A. 1936. On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, Series 2, 42 (1936-7), pp 230–265. Errata appeared in Series 2, 43 (1937), pp 544–546. Online version f.eks.: <www.dna.caltech.edu/courses/cs129/caltech_restricted/Turing_1936_IBID.pdf>. Tilgået 20.11 2014.

Wikipedia 2014. *Big Data*. <en.wikipedia.org/wiki/Big data>. Tilgået 20.11 2014.