DIGITAL HISTORIE OG ARKIVERET WEB SOM HISTORISK KILDE

NIELS BRÜGGER

DET DIGITALE OG HISTORIESKRIVNINGEN

Mængden af digitalt lagrede data er vokset eksplosivt inden for de seneste 10 år: i 2000 var 75 % af al verdens data lagret i ikke-digital form, i 2007 var dette tal 7%, mens det i 2012 var faldet til 2%. I samme periode vokser mængden af født digitalt materiale, eksempelvis behandler Google 24 petabytes om dagen, der uploades 10 millioner fotos til Facebook i timen og en times video til YouTube i sekundet.[1]

Det voksende digitale kildegrundlag ændrer betingelserne for historieforskning, og fremtidens historikere vil med stor sikkerhed komme til at bevæge sig rundt i et kildemæssigt landskab, der i stigende grad er digitalt, og i mange tilfælde kun digitalt. Imidlertid må den enorme mængde af digitalt materiale ikke overskygge, at blot fordi noget er digitalt, så er det ikke nødvendigvis digitalt på samme måde. Der er meget stor forskel på en digitaliseret avis, et Facebook feed, og en arkiveret webside, idet de hver især er digitale på forskellig vis, og de vil derfor skulle behandles forskelligt som kildemateriale.

Blandt de digitale medier har internettet i form af World Wide Web, eller blot 'web' siden slutningen af 1990erne været en stadig mere uomgængelig del af vore samfunds kommunikative infrastruktur, og det i en sådan grad at studier af emner inden for politik, kultur og samfund i bredeste forstand sandsynligvis på et eller andet tidspunkt vil involvere web som kilde.[2] Det vil eksempelvis være vanskeligt at skrive den ekstreme højrefløjs eller venstrefløjs historie omkring årtusindskiftet uden at inddrage disse grupperingers aktivitet på web. Der er derfor god grund til som historiker at beskæftige sig med de metodiske implikationer ved at bruge arkiveret web som kilde, herunder at udvikle en digitalt funderet kildekritik.

Med afsæt i et fokus på vores samtids digitale kildemateriales særlige væremåde og kvalitet og ikke kun på dets mængde vil denne artikel argumentere for en mere differentieret tilgang til det digitale materiale. Først præsenteres en

1 Mayer-Schönberger and Cukier: *Big Data*, 8-10.

2 I denne artikel skelnes der mellem internettet, som et netværk af computernetværk, og 'web' brugt som synonym for 'World Wide Web', der er en bestemt type software, som fungerer 'oven på' internettet, og som er baseret på særlige protokoller og programtyper, primært http, html og URL. Internet og web er altså ikke det samme, men de er forbundne.

overordnet typologisering af det digitale materiale, dernæst zoomes ind på én bestemt type materiale, nemlig arkiveret web. Dette gøres ved at introducere til webarkiveringsprocessen samt dens indvirkning på den type dokument, der bliver resultatet af processen med henblik på at diskutere, hvilke metodemæssige konsekvenser det får for den forskningsmæssige brug af denne kildetype, herunder overvejelser over behovet for en kildekritik, der tager afsæt i denne type digitalt materiales særlige karakteristika. Dernæst introduceres til samlinger af arkiveret web, som kan være relevante for danske historikere, henholdsvis det nationale danske webarkiv Netarkivet, Rigsarkivet samt det amerikanske Internet Archive. Og endelig afsluttes der med overvejelser over, hvordan denne artikels perspektiv på det digitale kan informere diskussioner om historieforskning og Big Data, samt hvilke nye udfordringer de nye mobile medier rejser.

DIGITALISERET, FØDT DIGITALT OG GENFØDT DIGITALT MATERIALE

Når man arbejder med digitalt materiale, så er det som nævnt vigtigt at holde sig for øje, at alt digitalt materiale ikke nødvendigvis er ens, bare fordi det er digitalt.[3] Digitale materialer vil typisk være digitale på forskellig måde, baseret på forskelle der angår både soft- og hardware samt den brugs- og/eller bevaringsmæssige praksis, materialet er indfældet i. Til at betegne hver digital materialetypes specifikke måde at kombinere soft- og hardware samt praksis på, og dermed at være digital på, kan man bruge ordet 'digitalitet'. I forlængelse heraf kan der skelnes mellem tre overordnede typer af digitalt materiale med hver sin digitalitet: Digitaliseret, født digitalt og genfødt digitalt materiale. Forskellen mellem de tre hovedtypers digitalitet ligger ved, hvordan materialet er blevet digitalt.

Det digitaliserede materiale er materiale, der tidligere har eksisteret i en ikke-digital form som for eksempel håndskrevne eller trykte medietyper (håndskrevne dokumenter, aviser, bøger, magasiner o.lign.), eller elektroniske medier såsom radio- og tv. Ved denne type materiale vil der i mange tilfælde stadig eksistere en original at gå tilbage til, enten i form af dokumenter, trykte medier eller optagne bånd med det udsendte materiale.

I modsætning hertil så findes det født digitale materiale kun i digital form, hvorfor der ikke er nogen original at gå tilbage til ud over det født digitale materiale selv. Født digitalt materiale kan fx være indhold på CD-ROM, DVD, eller i en hvilken som helst form for computernetværk, fx et Facebook feed, en YouTube video eller et tweet.

Og endelig så er det genfødt digitale materiale født digitalt materiale, som er blevet indsamlet og bevaret, og som er blevet ændret i denne proces, så det ikke længere er identisk med det født digitale materiale, der lå til grund for indsam-

3 Pointerne i dette afsnit refererer Brügger: 'Digital humanities'.

ling og bevaring; her kan der eksempelvis være tale om gemte computerspil eller arkiveret web.

Inden for hver af disse tre hovedformer gør der sig ligeledes forskelle gældende, for eksempel er der stor forskel på en digitaliseret avissamling og en digitaliseret radio- eller tv-samling, ligesom for eksempel internet, web og mobile medier er meget forskellige, til trods for de alle er født digitale; og selv inden for hver af disse er forskellene markante, eksempelvis er Facebook, Twitter og et almindeligt websted forskellige, til trods for de alle er på web.

Distinktionen mellem tre hovedtyper af digitalt materiale er væsentlig, fordi den kan lede opmærksomheden hen på, at de enkelte materialetypers forskellige digitalitet har indflydelse på, hvordan der kan interageres med dem, både i forbindelse med indsamling, bevaring og tilgængeliggørelse og i den forskningsproces, hvor materialet bruges. Der er for eksempel stor forskel på at digitalisere en avissamling, på at indsamle Facebook-opdateringer og på at arkivere websider, og disse forskelle er indlejret i materialet på en sådan måde, at de også i vid udstrækning opstiller muligheder og grænser for den efterfølgende forskningsmæssige brug heraf, herunder de digitalt understøttede analysemetoder, der kan bringes i anvendelse. Vi kan med andre ord ikke generelt regne med uproblematisk at kunne bruge samme søge-, annoterings- og analysesoftware på hver af de tre ovennævnte typer digitalt materiale, selvom genbrug i nogle tilfælde vil være mulig.

Blandt de genfødt digitale materialer er arkiveret web den undertype, der dels har haft den største samfundsmæssige udbredelse, dels den der findes de største samlinger af. Der er derfor god grund til, at arkiveret web fremover vil blive en væsentlig kilde for historikere, og derfor ligeledes god grund til at se nærmere på, hvad der kendetegner arkiveret web som kilde, og hvilke konsekvenser det arkiverede webs digitalitet måtte have for den forskningsmæssige brug af det. For at få det arkiverede webs særlige kendetegn til at fremstå tydeligt, bliver der i det følgende løbende sammenlignet med en samling digitaliseret materiale, nærmere bestemt digitaliserede aviser.[4]

WEBARKIVERING

Det er relevant kort at opholde sig ved, hvorfor man i det hele taget gemmer web, det findes jo derude, online. Grundlæggende gemmer man web, fordi det i sin online form er en meget flygtig kilde, hvilket tydeligt illustreres ved, at den gennemsnitlige levetid for en webside anslås at være mellem to og fire måneder, samt

4 Dette fokus på web som kilde i sin egen ret er forskelligt fra den tilgang til web, der eksempelvis kendes inden for traditionen 'digital history' (Cohen and Rosenzweig: *Digital History*), hvor web primært ses som formidlings- eller kommunikationsplatform (jf. Brügger: 'When the Present', 104-106).

at ud af 10 millioner websider arkiveret i 2001 var over 90 % væk i 2013.[5] Denne flygtighed udgør en væsentlig forskel til den digitaliserede samling, hvor der som nævnt ovenfor eksisterer en original. Denne original kan også være truet af at forsvinde, fx nedbrydes surt papir produceret i perioden ca. 1800-1985,[6] men i modsætning til online web er der typisk tale om en længere 'forsvindingsperiode', ligesom man ved præcist, hvad der forsvinder, hvilket samlet set betyder, at arkiveringsinstitutionen har lang tid til at forberede digitaliseringen, og i den tid, dette forberedes, ændrer originalen sig kun langsomt.

Arkivering af web

Før der gås i detaljer med det arkiverede webs karakteristika, er det relevant kort at forklare, hvordan webarkiveringsprocessen konkret foregår. Indsamling af det online web kan ske på flere måder, men den mest udbredte blandt store webarkiver som fx det danske Netarkivet (jf. nedenfor) er den såkaldte web crawling.[7] Web Crawling foregår ved, at arkiveringssoftwaren fyldes med en række webadresser som fx dr.dk, jp.dk, politiken.dk osv., hvorefter den kontakter de pågældende webservere og henter det materiale, der ligger på dem. Som det vil fremgå af det følgende, så betyder en række forhold ved denne proces, at det, der arkiveres, ikke nødvendigvis er en 1:1 kopi af det online web, men derimod en transformering heraf. Derfor kan arkiveret web betragtes som genfødt digitalt materiale.

Er et webarkiv et arkiv?

Når man taler om webarkivering og webarkiver, så er det desuden væsentligt at holde sig en vis terminologisk uklarhed for øje. Normalt vil kulturarvsinstitutioner, der indsamler og gemmer dokumenter i bredeste forstand, være delt i på den ene side arkiverne, der indsamler materiale, der ikke har været offentligt tilgængeligt, og på den anden side bibliotekerne, der indsamler dokumenter, der har været tilgængelige for offentligheden (denne skelnen opretholdes naturligvis ikke altid lige strengt). Langt de fleste webarkiver i verden indsamler det offentligt tilgængelige web, hvorfor betegnelsen 'webarkiv' grundlæggende er misvisende, men ikke desto mindre den, der bruges. Og på samme måde er verbet for selve indsamlingshandlingen – webarkivering – den term, der bruges i dag og har været brugt siden midten af 1990erne. For at undgå misforståelser er det væsentligt at minde om denne terminologiske uklarhed, ikke mindst i forhold til forskere, der er vant til at arbejde i arkiver (i klassisk forstand). Webarkiver er i langt de fleste tilfælde samlinger af fortidens offentligt tilgængelige web (enkelte institutioner arkiverer det ikke-offentlige web, se nedenfor om Rigsarkivet). Den følgen-

5 Brügger, 'Web History', 318, samt Agata, Miyata, Ishita, Ikeuchi and Ueda: 'Life span'.

6 Bibliotek og Medier: *Bevaring*.

7 For en oversigt over andre arkiveringsformer, se Brügger: 'Web Archiving', 27-29.

de karakteristik af arkiveret web gælder dog generelt for alle typer samlinger af bevaret web.

WEBARKIVERINGSPROCESSEN OG DET ARKIVEREDE WEB

Som antydet ovenfor, så er indsamling og arkivering af online web ikke en neutral proces. Derimod så omskabes det, der arkiveres, på et utal af måder, både i selve arkiveringsprocessen og i den senere proces med at fremfinde og vise det arkiverede i arkivet.

En digitaliseret samling er som nævnt baseret på en eksisterende original, som man i en vis udstrækning vil kunne gå tilbage til, mens et webarkiv bygger på en flygtig original, et 'moving target', der med en vis sandsynlighed ændrer sig uophørligt, og tendentielt forsvinder hurtigt.[8] I modsætning til den digitaliserede samling, så skal web arkiveres her og nu, man kan ikke regne med at kunne vente med indsamlingen.

Hvad og hvordan

Enhver samling bygger på til- og fravalg. Noget tages med, mens andet ikke gør, og det gælder naturligvis også for eksempel en digitaliseret avissamling. Arkiveringsinstitutionen skal vælge, *hvad* der skal digitaliseres, fx hvilke aviser der skal overføres til digital form.

I et webarkiv skal der ligeledes vælges, hvad der skal indgå i samlingen, men derudover skal der også tages stilling til, *hvordan* der skal arkiveres. Ved digitalisering tages der naturligvis også stilling til, hvordan digitaliseringen skal gennemføres, og der er visse tekniske valg, der skal tages, fx med hensyn til genkendelse af layout på siderne, OCR-genkendelse o.lign. Men omend disse valg har stor betydning for det efterfølgende resultat, så er de få og gennemskuelige, og alle andre arkiveringsinstitutioner, der gør det samme, vil få samme resultat. Til sammenligning er rækken af valg i webarkiveringsprocessen langt mere komplekse og gennemgribende, eksempelvis skal der vælges, om bestemte filtyper skal medtages eller ej fra de webservere, arkiveringssoftwaren besøger, ligesom der skal tages stilling til, om arkiveringssoftwaren må hente materiale fra andre webservere, hvor mange niveauer under forsiden, den skal gå ned på et website, og hvilken maksimumsgrænse der er for mængden af arkiveret materiale, for blot at nævne nogle af de utallige valg, der skal træffes. Disse mange valg betyder, at det er meget sandsynligt, at det samme websted arkiveret på samme tid af to webarkiver, vil se forskelligt ud.

8 Dette afsnit refererer og udbygger dele af Brügger: 'Digital humanities'.

Manglende gennemsigtighed

Processen med at skabe en digitaliseret samling er generelt gennemsigtig, og i de tilfælde, hvor den ikke er, vil den manglende gennemsigtighed være systematisk. For eksempel ved man ved digitalisering af aviser præcis hvilke aviser, der digitaliseres, og hvilke indstillinger der er brugt, og hvis der er en fejl et sted, så vil den typisk være systematisk og forekomme alle steder.

Anderledes med arkiveret web. Webarkiveringsprocessen er af mange grunde betydeligt mindre gennemsigtig end digitaliseringsprocessen. For det første vil der løbende kunne opstå tekniske problemer, som ikke umiddelbart er forklarlige, og derfor heller ikke systematiske, og som skyldes forhold ved selve det webmateriale, man ønsker at arkivere, og som man ikke selv er herre over på samme måde som ved digitaliseringsprocessen; eksempelvis kan bestemte typer software på det websted, der skal arkiveres, få arkiveringen til at gå i stå. De såkaldte 'crawler traps' kan have samme virkning, som det for eksempel ses på websider, hvor arkiveringssoftwaren forsøger at hente en kalender, der som regel er uendelig. Arkiveringssoftwaren vil derfor blot gå i stå, når den når sin maksimumsgrænse for mængden af arkiveret materiale (problemet med 'crawler traps' kan løses, men her blot nævnt som eksempel). For det andet gælder det for det online web, at det ikke blot er flygtigt, det er også dynamisk, dvs. det kan (potentielt) ændre sig hele tiden, også under selve arkiveringsprocessen. Som hvis der pludselig under digitaliseringen af en avis blev skudt en ny side ind, eller en overskrift på side 2 pludselig blev ændret, mens man var ved at digitalisere side 5. Denne 'opdateringens dynamik' betyder, at man aldrig kan vide med sikkerhed, om det, der arkiveres, ændrer sig, mens arkiveringen foregår, ligesom man i bekræftende fald heller ikke kan vide, præcist hvor det ændrer sig og hvornår.

Version uden original

Som nævnt så vil den originale webside eller det websted, der oprindelig var online under arkiveringen, typisk være enten forsvundet eller ændret efter (eller under) arkiveringsprocessen. Det betyder, at det materiale, der findes i et webarkiv, ikke er kopier, som i en digitaliseret samling, men derimod en samling versioner uden tilgængelig original. Derudover er det i langt de fleste tilfælde ikke muligt blandt versionerne at identificere en mest 'korrekt' version, der ville kunne fungere som 'original'. Der er kun versioner, og ud fra disse må man forsøge at sandsynliggøre, hvordan originalen kan have set ud på det online web.

For lidt og for meget

I en digitaliseret samling er der typisk kun én udgave af hver enhed, det være sig ét eksemplar af et håndskrevet dokument eller en avis, en radioudsendelse, osv. Og går man ind i det enkelte eksemplar, er der – naturligvis – kun én udgave af hver avisside i dagens avis. På dette punkt forholder det sig også anderledes i et webarkiv.

På den ene side er der typisk 'for lidt', nemlig for så vidt alt det, der var på det online web, ikke nødvendigvis er kommet med i arkivet. Noget kan være valgt fra som en konsekvens af den valgte arkiveringsstrategi (jf. nedenfor om det danske Netarkivet), mens andet af både forklarlige og uforklarlige arkiverings- eller servertekniske grunde ikke er kommet med, f. eks kan der mangle billeder, streamet video, indhold i diskussionsfora mm. Som hvis nogle eksemplarer i en digitaliseret avissamling indeholdt alle sider, andre kun forsiden, nogle havde kun overskrifter, andre manglede alle billeder, o. lign. Et webarkiv er altså ofte en samling fuld af huller, men i modsætning til andre typer samlinger, hvor mangler naturligt også forekommer, enten som en konsekvens af bevidste og systematiske fravalg, eller fordi materialet bare ikke har været tilgængeligt, så er manglerne i et webarkiv i langt overvejende grad usystematiske, tilfældige og uforklarlige.

Men på den anden side så er der i mange tilfælde også for meget i et webarkiv, nemlig for så vidt én webside på et websted eller ét foto på en webside kan være gemt flere gange dagligt, mens andre websider eller fotos kun er gemt én gang inden for samme tidsrum, eller måske slet ikke. I disse tilfælde vil der altså være for mange udgaver af 'det samme', uden at man helt med sikkerhed kan afgøre, i hvilken udstrækning det faktisk er det samme. Som hvis nogle eksemplarer i en digitaliseret avissamling forekom som 10 delvist forskellige udgaver fra samme dag, eller som hvis side 3 i en avis skulle hentes fra dagen før og side 5 fra dagen efter.

Et samling arkiveret web vil altså i mange tilfælde være en rodet samling med både for lidt og for meget på samme tid, og begge dele vanskeliggør en eventuel rekonstruktion af, hvordan et websted så ud på et givet tidspunkt i fortiden, da det var online, fordi det skal sammenstykkes af stumper fra delvist overlappende, men ikke identiske versioner. Et websted vil derfor ofte kun kunne rekonstrueres inden for et tidsrum og ikke på et tidspunkt.

Hyperlink og inkonsistens

Digitaliserede samlinger kan i visse tilfælde indeholde hyperlinks, der gør det muligt at springe fra fx én avisside til en anden. Det er imidlertid en tilføjelse, som arkiveringsinstitutionen kan vælge at lægge oven på den digitaliserede samling for at lette brugerens bevægelser rundt i samlingen. Og uanset om en samling digitaliserede aviser er beriget med hyperlinks eller ej, så er den i tid- og rummæssig henseende konsistent, det vil sige eksemplarerne kommer i kronologisk rækkefølge, og der er lige mange af hver.

I et webarkiv er hyperlinket derimod en integreret og uadskillelig del af web – hvis hyperlinket fjernes fra web, holder det grundlæggende op med at være web. Men i kombination med alle de ovennævnte forhold, så betyder tilstedeværelsen af hyperlinket, at en samling arkiveret web fremtræder som både tids- og rummæssigt inkonsistent.

Tidsmæssigt er webarkivet inkonsistent, fordi det, hyperlinket linker til, ikke nødvendigvis er arkiveret på samme tidspunkt som den side, der linkes fra, ofte

er der tale om en tidsforskel på flere måneder. Det ville svare til, at en avisartikel henviste til en senere opfølgende artikel dagen efter, men den artikel, der var blevet gemt, var i stedet fra to måneder senere. Dette forhold har stor betydning, hvis man vil lave analyser baseret på hyperlinks, for eksempel hyperlinkanalyse af linknetværket mellem en række websteder, idet hele linkstrukturen vil være tidsmæssigt inkonsistent.

Rummæssigt er webarkivet inkonsistent for så vidt ikke alle websteder er arkiveret lige dybt under forsiden. I nogen tilfælde er kun forsiden gemt, andre har alle sider to niveauer herunder, og atter andre mange flere niveauer. Som hvis den digitaliserede avissamling bestod af forsider i nogle tilfælde, 3 sider i andre og hele eksemplarer i atter andre. På denne måde er webarkivet rummæssig inkonsistent.

Web som kildekode og fragment

En digitaliseret samling vil på det mere tekniske niveau typisk bestå af en samling enkeltfiler, hvor én fil oftest vil svare til ét eksemplar, fx svarer én fil til ét dokument, én udgave af dagens avis eller til én radio- eller tv-udsendelse. Det betyder for eksempel, at man kun vanskeligt kan udskille skrevet tekst fra billeder, og hvis man fra arkivinstitutionens side vælger at gøre det, så er det noget, der er foregået i digitaliseringsprocessen, idet det af indlysende årsager ikke er en integreret del af det, der digitaliseres: papiravisen adskiller ikke i materiel forstand skrevet tekst fra billeder, idet de begge består af tryksværte på papir.

Anderledes i webarkivet, hvor det, der gemmes, ikke er selve den synlige webside, sådan som vi ser den i vores webbrowser som for eksempel Internet Explorer eller Firefox, men derimod sidens kildekode. Det, der sker, når vi skriver en webadresse ind i webbrowserens adressefelt – fx jp.dk – er, at den webserver, vi hermed kontakter, sender den sides kildekode, som vi beder om i form af html-kode – in casu forsiden på jp.dk – og denne kildekode oversætter vores webbrowser med det samme til den webside, vi ser, hvor tekst, fotos mm. er hentet og placeret på bestemte steder. Som webbruger ser vi med andre ord aldrig kildekoden i sig selv, men kun den oversatte form, der præsenteres for os i webbrowseren som en læsbar webside. Men det, der arkiveres i et webarkiv, er kildekoden samt de webobjekter, den måtte pege på, fx at et billede, et stykke grafik, et Facebook feed eller en overskrift skal hentes. Indholdet i et webarkiv er udelukkende html-sidernes kildekode samt stumper og stykker, som html-koden bygger siden op af, kort sagt: en samling fragmenter. Så i modsætning til den digitaliserede samling, hvor én avis svarer til én fil (groft sagt), og hvor opsplitningen i mindre fragmenter sjældent forekommer, og hvis den gør, så foretages det kontrolleret og systematisk, så svarer én webside eksempelvis til et patchwork af mange små fragmenter, der allerede i udgangspunktet er splittet op i et utal af filtyper og indholdsformer, og som desuden i mange tilfælde hentes forskellige steder fra, for eksempel på andre webservere.

WEBARKIVET OG VISNINGEN AF DET ARKIVEREDE WEB

Alle ovennævnte forhold har stor betydning for, hvordan webarkivet fremtræder for den forsker, der sidenhen skal bruge det, samt for de kildekritiske overvejelser historikeren i den forbindelse må gøre sig. Men der gør sig også særlige forhold gældende i selve visningen af det arkiverede web, efter at det er kommet ind i arkivet.

Mange mulige tilgange til det arkiverede materiale

Som nævnt ovenfor så består et element i en digitaliseret samling typisk af én fil, fx dagens avis, og denne fil kan man søge i, hvis der er lavet OCR-genkendelse, mens billedsøgning sjældent er muligt, fordi det er vanskeligt at lave billedgenkendelse på en indscannet side. Den efterfølgende forskertilgang til digitaliseret materiale begrænser sig derfor oftest til søgning og annotering samt analyser af skreven tekst. Den digitaliserede tekst er med andre ord relativt stabil, og af samme årsag er der et relativt begrænset register af mulige tilgange til den.

Et webarkiv består grundlæggende af en meget stor spand med milliarder af digitale fragmenter, der enten er eller hører til de enkelte html-sider. Det betyder, at der er et utal af mulige forskertilgange til webarkivets samling, alt efter hvilke fragmenter, forskeren ønsker at fokusere på, og hvordan vedkommende ønsker at få dem kombineret og præsenteret. Et par eksempler. Man kan vælge at få en websides fragmenter præsenteret, så det ligner det, man ville have set på det online web, hvis man havde kigget på det via sin webbrowser i fortiden; dette er den gængse præsentationsmåde i webarkiver, der bruger den såkaldte Open Wayback præsentationsmåde, såsom det danske webarkiv Netarkivet eller det amerikansk baserede Internet Archive. Men man kunne også have interesse i udelukkende at analysere hyperlinknetværk, dvs. hvem der linker til hvem, og man vil i så tilfælde bare skulle have samtlige hyperlinks på de websider, man ønsker at studere og ikke andet. Eller man kunne ønske at lave billedanalyser af samtlige billeder på en samling websteder, og man vil udelukkende skulle have fat i de relevante billeder. Så i modsætning til den digitaliserede samling så er webarkivets materialer allerede i udgangspunktet fragmenterede og 'beriget' med metadata fx i form af html-kode, filtyper mv., hvilket gør det muligt at finde og samle fragmenterne efter forskerens behov. Arkiveret web er med andre ord både mindre stabilt som kilde, fordi én fil på ingen måde svarer til ét eksemplar af for eksempel en webside, og mere fleksibelt, fordi det er født fragmenteret, og dets fragmenter kan trækkes ud og kombineres på utallige måder.

Tidslighed på selve den arkiverede webside

Ovenfor blev den indbyggede tidslige inkonsistens, der er knyttet til hyperlinks, berørt. Der er som nævnt tale om en inkonsistens, der hidrører fra selve webarkiveringens tidsmæssige udtrækning. Men i forbindelse med det ovennævnte visningssoftware Open Wayback gør der sig også en anden form for tidsmæssig

inkonsistens gældende, nemlig den inkonsistens der knytter sig til selve sammenstykningen af en websides mange små elementer, når de i webarkivet samles som én webside. Som nævnt så sammenstykkes en webside på det online web ved, at en html-kode fortæller webbrowseren, hvilke stumper, der skal hentes, hvorfra, og hvor de skal placeres, og noget tilsvarende foregår i visningssoftwaren Open Wayback, bortset fra at der hentes stumper fra webarkivet og ikke fra det online web. Men i modsætning til det online web, hvor der på et givet tidspunkt kun er én udgave af hvert fragment, fx et foto, så kan der i webarkivet være flere – eller der kan mangle et fra det pågældende tidspunkt. Hvis det sidste er tilfældet, så vil Open Wayback softwaren automatisk hente det element, der ligger nærmest i tid, hvilket kan være alt lige fra en time tidligere til flere måneder senere. Det betyder, at den webside, man som forsker kigger på i webarkivet, og som i udgangspunktet stammer fra den dato, hvor sidens html-kode blev arkiveret, ikke nødvendigvis *in toto* er fra samme dato; det er nemlig muligt, og endog sandsynligt, at websiden henter nogle af dens elementer fra tidligere eller senere tidspunkter, hvis ikke de af en eller anden grund blev arkiveret samtidig med selve websiden (hvilket de ofte ikke gør). Med andre ord så gør der sig en tidslighed gældende på den webside, der for en umiddelbar betragtning ser 'samtidig' ud, og det er vel at mærke ikke noget, der gøres opmærksom på.

WEBARKIVERINGENS FORSKNINGSMÆSSIGE KONSEKVENSER

Man kan ikke fortænke selv den mest webinteresserede historiker i at miste lysten til at forske i arkiveret web med alle de udfordringer, som materialet i sig selv lægger i vejen. Ikke desto mindre vil det i de kommende år være stadig vanskeligere at undgå web som historisk kildemateriale, hvis man arbejder med historieskrivning for perioden fra slutningen af 1990erne og frem. Der vil derfor være behov for generelt at udvikle en historiefaglig 'digital literacy', der kan hjælpe historikeren med at få det fornødne materialekendskab til det digitale for dermed at kunne reformulere den traditionelle kildekritik i et digitalt landskab. I forhold til arkiveret web vil et godt udgangspunkt være at have viden om, hvad webarkiveringsprocessen gør ved det, der ender med at komme i webarkivet, og hvad det kan have af implikationer for den senere forskningsmæssige brug af materialet. En sådan viden er en nødvendig forudsætning for at opnå den fornødne metodebevidsthed i brugen af arkiveret web som historisk kilde.

Digital proveniens

Et væsentligt element i kildekritikken er proveniens. Behovet for at fastslå, hvorfra en kilde kommer, er ikke blevet mindre med det digitale, derimod er det blevet både mere påtrængende og på nogle punkter også lidt nemmere at håndtere.

For at komme dette lidt nærmere må man kigge på, hvad man kunne kalde den digitale teksts dobbelthed, som antydet ovenfor i forbindelse med en websides kildekode. Et digitalt dokument vil altid i tekstmæssig henseende være dobbelt

(tekst forstået i bred forstand, ikke kun som skrift), idet det på den ene side består af den digitale tekst og på den anden side består af den for os læsbare tekst. Den digitaliserede avis består af en tekst, der i sidste ende er skrevet med det digitale alfabets to bogstaver 1 og 0, der omsættes til henholdsvis sorte og hvide prikker på en skærm, hvilket danner linjer og former, der for os udgør skrift og billeder.[9] Og på samme måde består den arkiverede webside af en række digitale tekstlag, hvor html-koden er det øverste niveau, dvs. den digitale tekst, der generer den skrift og de billeder, vi ser på skærmen.

I en ikke-digital verden er det at fastslå proveniens i væsentlig grad knyttet til et dokuments synlige tekst, selvom både fundomstændigheder og dokumentets materielle underlag, fx pergament eller papir, kan spille en rolle og være med til at datere en kilde. Men i en verden af digitale kilder vil det være relevant at udstrække undersøgelsen af proveniens til også at gælde de lag af digital tekst, der muliggør den synlige tekst. Så mens det digitale på mange måder kan gøre omgangen med digitale dokumenter mere kompliceret, som det for eksempel ses ved webarkivet, så kan det samtidig være med til at åbne nye kildekritiske veje, fordi født og genfødt digitale dokumenter ofte kommer med informationer om deres egen herkomst – i tekstlig form – som vi ikke kender fra ikke-digitale kildetyper. Med andre ord har digitale dokumenter ofte indlejret deres proveniens i sig, det være sig websiders kildekode, der kan indeholde oplysninger om, hvilke elementer siden består af, hvornår den er lavet, og af hvem, eller det kan være digitale fotos, der indeholder en guldgrube af oplysninger om, hvor fotoet er taget, hvilket apparat eller hvilken mobiltelefon, der er brugt osv.

Set på den baggrund, så er der behov for en udvidet kildekritik i form af, hvad man kunne kalde et digitalt materialekendskab, hvor fokus er på, hvad der kendetegner de digitale tekstlag inden for de tre hovedtyper af digitalt materiale samt deres underformer. Historikere skal med andre ord lære at læse de i ordets egentlige forstand digitale tekster.

Mod en webfilologi

I forbindelse med arkiveret web så er første trin i historikerens metodebevidsthed at forlige sig med, at det webmateriale, der findes i webarkivet, ikke nødvendigvis svarer til det, der var online i fortiden, samt at der meget vel kan findes delvist overlappende versioner af det samme.

Med henblik på at etablere et dokuments proveniens eller genese har sammenligninger mellem varianter længe været en del af filologiens værktøjskasse, og på nogle måder er det samme fremgangsmåde, der kan bringes i anvendelse i arbejdet med arkiveret web. Ligesom den klassiske filologi sammenligner varianter af håndskrifter med henblik på at etablere en tekst eller måske ligefrem iden-

9 Om den digitale tekst, se Finnemann: 'Modernity'.

tificere en 'original', så kan en webfilologi sammenligne versioner med henblik på at sandsynliggøre, hvordan et givent websted har set ud i fortiden på det online web. Der er imidlertid også forskelle. For det første at ingen af de arkiverede webversioner kan siges at udgøre en original for en af de andre, og for det andet at mens filologien typisk sammenligner versioner bagud i tid – den ene variant kom før den anden variant – så sammenligner webfilologien oftest i samtidighed, det vil sige versioner der alle er indsamlet inden for et relativt begrænset tidsrum.[10]

I modsætning til den klassiske filologi så vil det arkiverede webs digitale tekst gøre det muligt for webfilologen at få udviklet en ny type værktøjskasse. Det kan eksempelvis være værktøjer, der kan hjælpe med sammenligning af to eller flere websider, eller redskaber som på en let tilgængelig måde kan vise, hvornår et givent element på en webside blev arkiveret, hvilket ville være en stor hjælp i forbindelse med den ovennævnte tidslighed, der er indlejret i en vist webside. Og endelig værktøjer der gør det muligt at præsentere de mange oplysninger, der automatisk genereres under en webarkivering som for eksempel, hvornår arkiveringen begyndte og sluttede, om der forekom fejl, hvorfor materiale ikke kom med, o.lign.

ARKIVERET WEB I DANMARK – OG USA

Har man som historiker valgt at trodse udfordringerne og sat sig for at bruge arkiveret web som historisk kildemateriale, så er det næste spørgsmål, hvor man finder det arkiverede web.

Som historiker i Danmark, der ønsker at studere arkiveret web, vil især tre samlinger være relevante, nemlig det nationale danske webarkiv Netarkivet, Rigsarkivets samling af webmateriale fra myndigheders og virksomheders intranet, samt det amerikanske Internet Archive.[11]

Netarkivet

Fra midten af 2005 er den offentlige del af det danske web blevet arkiveret i det nationale danske webarkiv Netarkivet (netarkivet.dk). Netarkivet blev grundlagt af Statsbiblioteket og Det Kongelige Bibliotek i fællesskab (fra 2017 sammenlagt som Det Kgl. Bibliotek), og det er baseret på Pligtafleveringsloven, ifølge hvilken Netarkivet skal indsamle og gemme dansk materiale i computernetværk. Dansk materiale afgrænses på to måder, dels gemmes alt materiale, der befinder sig på det danske internetdomæne .dk, dels alt materiale på andre internetdomæner som fx .com, .nu osv., der er rettet mod en dansk offentlighed, eller som ejes af danske borgere. Dette sidste materiale kaldes 'Danica', og i modsætning til ma-

10 I Brügger: 'Web Archiving' er formuleret et bud på generelle metoder og regler, der kan være vejledende for en webfilologi.

11 For oversigter over webarkiver internationalt, se Truman: *WebArchiving*, IIPC Member Archives, List of Web archiving initiatives.

terialet på .dk, så opspores det manuelt/semi-automatisk.[12] Som nævnt arkiverer Netarkivet alt offentliggjort materiale, hvilket betyder, at man også kan indsamle materiale, der ligger på passwordbeskyttede områder, hvis enhver ville kunne få et password til materialet (enten ved blot at anmode om det, eller ved at købe det). Det betyder, at fx betalingsbelagt materiale på en netavis skal arkiveres af Netarkivet, mens en virksomheds intranet eller en personlig Facebook-side ikke skal.

Af både tekniske og ressourcemæssige grunde er det ikke muligt at arkivere hele det danske web uophørligt, så derfor benytter Netarkivet tre strategier med henblik på at få så stor dækning som muligt. Den første strategi kaldes tværsnitsarkivering, og her indsamles hele det danske internetdomæne .dk (i 2015 ca. 1,2 million domænenavne, ca. 30TB) samt Danica-materialet, hvilket gøres fire gange årligt. En sådan tværsnitsarkivering tager mellem to og fire måneder at gennemføre, hvorfor den er mindre anvendelig som strategi i forhold til websteder, der ændrer sig ofte. Den anden strategi kaldes selektiv arkivering, og den skal sikre, at de websteder, der opdateres hyppigt, bliver arkiveret meget ofte; her er udvalgt omtrent 100 websteder, typisk nyhedswebsteder, der indsamles dagligt, nogle af deres sider flere gange dagligt. Og endelig kaldes den tredje strategi begivenhedsarkivering, dvs. indsamling af webmateriale i forbindelse med nationale begivenheder, såsom politiske valg, katastrofer, sportsbegivenheder, o.lign.; her arkiveres færre websteder end ved tværsnitshøstningen, men der arkiveres oftere.

Ifølge Pligtafleveringsloven er der kun adgang til Netarkivet for forskere, og den adgang, der gives til materialet, er primært den tidligere omtalte Open Wayback software, der præsenterer indholdet, så det ligner websiden, som den så ud i fortiden.[13]

Netarkivet er blandt de bedste webarkiver i verden, og materialet i arkivet er af meget høj kvalitet. Men det betyder ikke, at alt dansk materiale kan findes i arkivet. Som nævnt startede Netarkivet først i 2005, hvilket betyder, at de første 10 års danske web uigenkaldeligt et væk. Derudover vil eksempelvis en lokalhistoriker have vanskeligt ved at finde en provinsbys lokalaviser dagligt, eller webaktiviteten i forbindelse med en lokal begivenhed. Det skyldes, at de tre ovennævnte strategier er nationalt funderet, hvorfor lokalt materiale falder igennem strategierne, hvis ikke det bliver opfanget af én af de fire årlige tværsnitsarkiveringer.[14]

Rigsarkivet

Rigsarkivet indsamler og arkiverer materiale fra myndigheders, institutioners og virksomheders intranet, det vil sige materiale, der ikke har været offentligt

12 For en kort indføring i Netarkivets indsamlingsstrategier, se Schostag and Fønss-Jørgensen: 'Webarchiving'.

13 Retningslinjerne for forskeradgang kan ses på http://netarkivet.dk/adgang (20-03-2017).

14 Disse problemstillinger er uddybet nærmere i Brügger: 'Web som'.

tilgængeligt.[15] Rigsarkivets webarkivering ligger i forlængelse af arkivets øvrige virksomhed, hvorfor samlingen består af det interne webmateriale, som vurderes at udgøre historisk værdifuldt materiale, og som Rigsarkivet fastsætter skal afleveres, baseret på en drøftelse med den pågældende myndighed.

Materialet afleveres i de formater, som er fastsat i Bekendtgørelse nr. 1007 om arkiveringsversioner, og det foregår ved, at myndighederne konverterer data og dokumenter til Rigsarkivets bevaringsformat. Formålet med konverteringen er, at Rigsarkivet får data og dokumenter uafhængigt af det system, de er skabt i, hvilket skal sikre et ensartet format, som kan langtidsbevares. I de tilfælde, hvor der er tale om materiale, som allerede findes i andre formater end et webintranet, så kan myndighederne lægge bevaringsværdige dokumenter i deres ESDH-system (Elektroniske Sags- og Dokumenthåndteringssystemer), hvorfra det afleveres til Rigsarkivet. Hvis Rigsarkivet fastsætter en sådan løsning, vil informationsindholdet være bevaret i disse dokumenter, men selve intranetsiden, hvor indholdet også har været på, vil ikke.

Webintranet gemmes således ikke i et webformat som fx html, men derimod som kopier af enkeltsider, eksempelvis i pdf-format. Lyd og video kan dog gemmes, men som selvstændige filer og ikke som det oprindeligt var indlejret på en webside.

Ovennævnte forhold betyder, at de indholdstyper på et webintranet, som udelukkende findes på et intranet, sandsynligvis ikke vil være bevaret; det kan være mere uformel, efemerisk og interaktionsbaseret kommunikation i diskussions- og chatfora, i tests og afstemninger, på opslagstavler, lyd og video i streamet form, o.lign.

Da et webintranet ikke er gemt i sit oprindelige html-format, så vil forskeren ikke kunne navigere rundt i det ved hjælp af hyperlinks, som var det online, interaktive elementer vil ikke virke, og eventuelle lyd- og video vil skulle findes som enkeltfiler og kobles med de websider, de måtte høre hjemme på. Derudover er det ikke muligt at foretage en målrettet og struktureret søgning på intranet som specifik materialetype og således finde en myndigheds eller virksomheds intranet i Rigsarkivets arkivdatabase, Daisy. Men som nævnt ovenfor, så kan informationsindholdet på et intranets websider være blevet lagret i myndighedernes ESDH-systemer, og det kan derfor søges ved at fremsøge disse. Af disse grunde er det vanskeligt at vurdere, hvor mange intranet der er bevaret i Rigsarkivet samt fra hvilke myndigheder. Da virksomheder ikke er omfattet af samme afleveringspligt som offentlige myndigheder, så er der endnu ikke afleveret intranetsider fra private virksomheder.

15 Rigsarkivet har været behjælpelig med oplysninger om dets bevaring af webintranet.

The Internet Archive

Hvis ikke det materiale, man leder efter, findes i Netarkivet, så kan der være god grund til at søge efter det i verdens største webarkiv, det amerikanske Internet Archive (archive.org).[16] The Internet Archive er en non-profit institution, der siden 1996 har indsamlet web på verdensplan. I modsætning til Netarkivet, så baserer Internet Archive ikke sit virke på en national lovgivning, og det dækker derfor heller ikke webaktiviteten inden for et geografisk afgrænset område som en nation. Derimod baseres deres arkiveringsstrategi overvejende på at forfølge hyperlinks kumulativt, ud fra det allerede arkiverede, hvorfor arkivet principielt arkiverer alt, hvortil der linkes (og dermed ikke det, der ikke linkes til). Derfor har Internet Archive også en del dansk materiale, og for så vidt angår perioden før 2005, hvor Netarkivet åbnes, så er Internet Archive uden sammenligning den største samling dansk web.

Imidlertid er kvaliteten af det arkiverede meget svingende, for eksempel gemmes ofte kun forsiden og måske to niveauer herunder, og især på de tidlige udgaver mangler der ofte billeder og grafik. Derudover så betyder den manglende nationalt orienterede indsamlingsstrategi, at det danske område er dækket meget usystematisk, både hvad angår hvilke websteder der er arkiveret, og hvor ofte de arkiveres. Og endelig så skal Internet Archive omgås med en vis portion stringens, idet arkivets indhold vises gennem den tidligere omtalte Open Wayback software, der præsenterer indholdet, så det ligner websiden, som den så ud i fortiden, men dels kan elementerne på siden være hentet både tidligere og senere end den pågældende webside, og man kan således efter et par klik befinde sig flere måneder væk fra, hvor man startede, dels kan det også være hentet fra det nu online web, idet Internet Archive tilgås frit online. Det betyder, at hvis det, der linkes til, ikke findes i arkivet, så sendes brugeren hen til det pågældende websted, som det ser ud i dag. Det finder man som oftest ud af, men mekanismen kan også foregå mere skjult, hvilket er tilfældet, hvis et givent element på den arkiverede webside kalder en webserver for at få information herfra – for eksempel dagens vejrudsigt – og denne webserver stadig er online og har indhold, så vises dette indhold på den arkiverede webside, og man ser således en strålende sommervejrudsigt på en side, der er arkiveret i december. I Internet Archive kan den arkiverede webside altså ikke blot være stykket sammen af elementer fra forskellige tidspunkter i webarkivet, men også af elementer fra dagens online web.

NYE MULIGHEDER, NYE UDFORDRINGER

Arkiveret web er blot én kildetype i en større digital kildeøkologi, der indbefatter både digitaliserede og født digitale kildeformer samt utallige andre former for genfødte digitale kilder. Med afsæt i denne artikels fokus på de forskellige digitale

16 Internet Archive er frit tilgængeligt online. Om Internet Archive, se Kimpton and Ubois: 'Year-by-Year'.

kilders digitalitet skal der afslutningsvis reflekteres dels over forholdet mellem historieskrivning og et af samtidens store modeord, Big Data, dels over de nye udfordringer, der anes i horisonten med fremkomsten af nye typer digitale kilder, som supplerer web.

Historieskrivning og Big Data

I 2014 udgav de to historikere Jo Guldi og David Armitage bogen *The History Manifesto*, der hurtigt skulle vise sig at blive meget omdiskuteret.[17] Med afsæt i Fernand Braudels overvejelser over *la longue durée* er det bogens hovedargument, at historieforskningen skal genintroducere historiske analyser af lange tidsstræk, i modsætning til det spøgelse, der efter forfatternes opfattelse "is haunting our time: the spectre of the short term."[18] Som et led i dette forehavende forsøger bogen også at tage livtag med et af de væsentligste nye kendetegn ved det 'new *longue durée*', der argumenteres for, nemlig "the abounding sources of big data available in our time – data ecological, governmental, economic, and cultural in nature, much of it newly available to the lens of digital analysis".[19] Denne samtidskarakteristik uddybes i bogens afsnit om 'Big questions, big data', hvor det bliver tydeligt, at det væsentligste for Guldi og Armitage er, at der i dag er meget data til rådighed, og at big data netop muliggør *longue durée* analyser, hvilket ikke er så overraskende, deres overordnede argument taget i betragtning.

Men måske er det væsentligste ikke (kun) mængden af data i dag, men derimod den form, data foreligger i, nemlig at de er digitale. Guldi og Armitages hovedargument overskygger således tendentielt, at det væsentlige ikke er, at vi har store mængder data til rådighed, men derimod at vi har store mængder *digital* data til rådighed. Og dem har vi – som nævnt i denne artikels første linjer – ganske rigtigt rigtig mange af. Men at mængden af ikke-digitale kilder stagnerer samtidig med at mængden af digitale kilder vokser eksplosivt, er ikke ensbetydende med, at de store mængder digitale data nødvendigvis skal bruges som 'big data', de vil lige så vel kunne bruges som 'small data', for det afgørende er, at de er digitale. Det overdrevne fokus på 'big data' kommer således nemt til at overskygge, at skiftet ikke kun er en bevægelse fra 'small data' til 'big data', men derimod at det mere grundlæggende er et skifte fra ikke-digital til digital.

I en ikke-digital verden vil historikerens valg af kilder typisk være begrænset af en kombination af, hvad der var tilgængeligt, og hvad det var praktisk muligt at overkomme at læse. Og af den sidste grund har tilgangen til kilderne typisk været en vis grad af nærlæsning af hvert enkelt dokument. Følger man den italienske litteraturforsker Franco Morettis distinktion mellem 'close reading' og 'distant rea-

17 For overblik over de diskussioner bogen afstedkom, se http://historymanifesto.cambridge.org/media. (20-03-2017).

18 Guldi and Armitage: *History*, 1.

19 Guldi and Armitage: *History*, 9.

ding', så har tilgangen til kilderne for det meste været 'close reading', dvs. nærlæsning af et relativt begrænset antal dokumenter.[20] Som enhver anden kildetype så kan webarkiver bruges til at understøtte et utal af historieforskningsprojekter, der minder om dem, der blev lavet uden digitale kilder, ud fra en 'close reading' tilgang. Men som enhver anden digital samling så åbner webarkiver også for nye typer forskningsprojekter, der ligger tættere på 'distant reading'. Med andre ord tillader en digital kildesamling som et webarkiv dels at stille (og besvare) forskningsspørgsmål, som minder om dem, vi tidligere har stillet, dels at formulere forskningsspørgsmål, som man ikke kunne forestille sig tidligere, men som materialet nu gør mulige.

At fokusere på et begrænset antal websteder og tilgå dem analytisk på traditionel nærlæsningsvis vil stadig være en oplagt mulighed for en forsker, der vil bruge arkiveret web.[21] Men med webarkivernes store digitale samlinger åbnes også nye muligheder i stil med dem, Guldi og Armitage fremhæver, hvor tekstmasser, der hidtil havde været uoverkommelige at 'læse', nu kan analyseres, baseret på delvist automatiserede analysemetoder. Dette skridt fra 'close' til 'distant reading' er taget i flere historiske projekter, der alle bygger på arkiveret web som kilde, eksempelvis i analyser af hele nationale webdomæner,[22] af udviklingen af abortdebatten i Australien fra 2005 til 2015,[23] eller af the Church of England's seneste udvikling i Storbritannien.[24]

Digitale kilder post-web – nye udfordringer

Web har i dag eksisteret i 25 år, hvilket i internetsammenhæng gør det til et modent, om ikke ligefrem gammelt medie. Og nye medietyper har allerede i et årti suppleret web, frem for alt mobile medieplatforme såsom smartphones og tavlecomputere, der i vid udstrækning bygger på såkaldte apps.

Ikke overraskende kommer disse nye typer digitale kilder med en digitalitet, der er væsensforskellig fra webs digitalitet, eksempelvis er de små, mobile, og i modsætning til web så henter en app løbende information fra databaser. Men samtidig har de overlappende digitalitet med web, det gælder på selve platformens niveau, hvor smartphones og tavlecomputere har både en webbrowser og apps, og det gælder på selve indholdets niveau, hvor det samme indhold på eksempelvis Facebook og lignende både kan ses på web og via en app. Der er her tale om en generel tendens ved digitale kilder, hvor dele af en medietypes digitalitet integreres i et andet medies digitalitet, det ses fx i mødet mellem digital radio eller tv, som udsendes som henholdsvis radio eller tv via broadcast eller kabler, men

20 Moretti: 'Conjectures'.
21 Eksempler herpå er Brügger: 'The Idea', samt Nanni: 'Reconstructing'.
22 Brügger: 'Probing'.
23 Ackland and Evans: 'Using'.
24 Webster: 'Religious'.

som samtidig udsendes på web som del af den pågældende stations website. På denne måde indlejres digital radio og tvs digitalitet i webs digitalitet.

Denne høje grad af kompleks integration udfordrer fremtidens historikeres brug af digitalt kildemateriale, ligesom den stiller store udfordringer til de institutioner, der skal indsamle og gemme disse kilder. Skal eksempelvis digital radio og tv på web gemmes som digital radio/tv eller som web, og hvordan får man det til at 'spille sammen', hvis det gemmes i to adskilte arkiver? Og skal eksempelvis en offentlig Facebookside gemmes som websiden, som den så ud, eller skal man gemme det indhold, der løbende blev vist i en app, uden at kunne gemme selve appen?

Men overlappene til trods så er der allerede i næsten et årti blevet produceret indhold, som kun er tilgængeligt via en app, det være sig interaktive 'bøger', diverse self-tracking apps, der er knyttet til mobilmediernes geolokalisering, eller interaktive kort, der samler alt fra fotos til anmeldelser om en given lokalitet. Disse mobile født-digitale medietyper vil givetvis i mange tilfælde være en uvurderlig kilde for fremtidens historikere. Men hvordan disse medier skal arkiveres – og dermed hvordan de efterfølgende kan bruges som historiske kilder – er stadig uafklaret, til trods for at de kan siges at være omfattet af den gældende Pligtafleveringslovs formulering om at indsamle og gemme dansk materiale i computernetværk. Sikkert er det dog, at vi allerede i dag har mistet det indhold, der er skabt i løbet af de første 10 år af mobilmediernes levetid, fordi det ikke er indsamlet og bevaret noget sted. Desværre ser webarkiveringens historie ud til at gentage sig.

LITTERATUR

Ackland, Robert and Ann Evans: 'Using the Web to Examine the Evolution of the Abortion Debate in Australia 2005-2015'. I N. Brügger and R. Schroeder (red.): *The Web as History: Using Web Archives to Understand the Past and the Present*, London: UCL Press, 2017, 159-189.

Agata, T., Y. Miyata, E. Ishita, A. Ikeuchi, and S. Ueda: 'Life span of web pages: A survey of 10 million pages collected in 2001', *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2014, 463-464.

Bibliotek og Medier: *Bevaring af surt papir i de statslige samlinger: Rapport fra Arbejdsgruppen vedrørende masseafsyring af papir*, Rapporter fra Bibliotek og Medier, nr. 6, København 2008, http://www.bs.dk/publikationer/rapporter/6/pdf/Bevaring_af_surt_papir_i_de_statslige_sa.pdf (20.03.2017).

Brügger, Niels: 'Digital humanities in the 21st century: Digital material as a driving force', *Digital Humanities Quarterly* 10 (3), 2016, http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html (20.03.2017).

Brügger, Niels: 'The Idea of Public Service in the Early History of DR Online'. I M. Burns and N. Brügger (red.): *Histories of Public Service Broadcasters on the Web*, New York: Peter Lang, 2012, 91-104.

Brügger, Niels: 'Probing a nation's web domain: A new approach to web history and a new kind of historical source'. I G. Goggin and M. McLelland (red.): *The Routledge Companion to Global Internet Histories*, New York: Routledge, 2017, in press.

Brügger, Niels: 'Web Archiving – between Past, Present, and Future'. I M. Consalvo and C. Ess (red.): *The Handbook of Internet Studies*, Oxford: Wiley-Blackwell, 2011, 24-42.

Brügger, Niels: 'Web History and the Web as a Historical Source', *Zeithistorische Forschungen* 9 (2), 2012, 316-325.

Brügger, Niels: 'Web som lokalhistorisk kilde – hvad er udfordringerne?'. K.H. Andersen and C.R. Jansen (red.): *Lokalhistorie: Fortid, nutid og fremtid*, Højbjerg: Forlaget Skippershoved, 2014, 279-295.

Brügger, Niels: 'When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies', *Historical Social Research* 37(4), 2012, 102-117.

Cohen, Daniel J. and Roy Rosenzweig: *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*, Philadelphia: University of Pennsylvania Press, 2006.

Finnemann, Niels Ole: 'Modernity Modernised: The Cultural Impact of Computerisation'. I P.A. Mayer (red.), *Computer, Media and Communication*, Oxford: Oxford University Press, 1999, 141-159.

Guldi, Jo and David Armitage: *The History Manifesto*, Cambridge: Cambridge University Press, 2014.

IIPC Member Archives (n.d.), http://netpreserve.org/resources/member-archives (20-03-2017).

Kimpton, Michele and Jeff Ubois: 'Year-by-Year: From an Archive of the Internet to an Archive on the Internet'. I J. Masanes, *Web Archiving*, Berlin: Springer, 2006, 201-212.

List of Web archiving initiatives (n.d.), https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives (20.03.2017).

Mayer-Schönberger, Viktor and Kenneth Cukier: *Big Data: A revolution that will transform how we live, work, and think*, New York: Houghton Mifflin Harcourt Publishing Company, 2013.

Moretti, Franco: 'Conjectures on world literature', *New left review*, *1*, Jan.-Feb., 2000, 56-58.

Nanni, Federico: 'Reconstructing a website's lost past – Methodological issues concerning the history of www.unibo.it', *Digital Humanities Quarterly*, 2017, in press.

Schostag, Sabine and Eva Fønss-Jørgensen: 'Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective', *MDR*, 41, 2012, 110-120.

Truman, Gail: *WebArchiving Environmental Scan*, Harvard Library Report, 2016, http://nrs.harvard.edu/urn-3:HUL.InstRepos:2565831 (20.03.2017).

Webster, Peter: **'**Religious discourse in the archived web: Rowan Williams, archbishop of Canterbury, and the sharia law controversy of 2008'. I N. Brügger and R. Schroeder (red.): *The Web as History: Using Web Archives to Understand the Past and the Present*, London: UCL Press, 2017, 190-203.

NIELS BRÜGGER
PH.D, PROFESSOR (MSO)
INSTITUT FOR KOMMUNIKATION OG KULTUR
AARHUS UNIVERSITET
EMAIL: NB@CC.AU.DK

ABSTRACT

Niels Brügger: Digital history and the archived web as historical source

Within the last decade the amount of digitally stored data has exploded, and in the same period of time the amount of born-digital material is growing, such as content on social media and the web. Future historians must find their way through a source landscape in which more and more sources become digital, and in many cases digital only. This article argues that all digital sources are not alike just because they are digital which leads to a distinction between digitised, born-digital and reborn-digital sources. Then follows an introduction to one specific type of reborn-digital material, the archived web, that is compared to digitalised

newspaper archives. Finally, it is discussed what the consequences are of the specific characteristics of the archived web for its use as a historical source.