Netarkivet – mere påkrævet end nogensinde

Tale ved 10-års-jubilæet for det officielle danske webarkiv, Netarkivet.dk, afholdt ved RESAW-konferencen 'Web Archives as scholarly Sources: Issues, Practices and Perspectives', Aarhus Universitet 8.6.2015.

Konferencen var arrangeret af RESAW, Aarhus Universitet, Statsbiblioteket, Det Kongelige Bibliotek, l'Institut des sciences de la communication du CNRS, Université de Lille 3, Institute of Historical Research (University of London), University of Amsterdam, British Library og Leibniz Universität Hannover/ALEXANDRIA.

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

Det er en stor fornøjelse at være her for at fejre 10-års-jubilæet for det nationale danske netarkiv, Netarkivet.dk, som officielt åbnede i 2005. Altså, det er det *officielle* jubilæum. Det Kongelige Bibliotek og Statsbiblioteket begyndte arkivering af web-materiale tilbage i 1997.

Det var imidlertid kun statisk materiale, og uden nogen form for forpligtelser til at levere materialet til arkivet og uden en strategi for indsamling af disse materialer. Hovedfokus var på elektroniske kopier af trykte materialer. Hverken nettet selv, eller interaktive og hypertekstsammensætninger, blev betragtet som relevante.

Jeg er nødt til at indrømme, at jeg som forsker dengang endnu ikke var klar over strategiens utilstrækkelighed, men i slutningen af 1999 fik jeg en brat opvågnen, da en af mine studerende – som skrev på sit speciale om Røde Kors' 'hjemmeside' – opdagede, at sitet var blevet udskiftet til et nyt i løbet af natten til den dag, hun skulle aflevere sit speciale.

Jeg husker hende komme ind på mit kontor for at fortælle om katastrofen. Hendes dokumentation var fuldstændig forsvundet bortset fra nogle få skærmprint og udskrifter. Den dag blev det meget tydeligt for mig, at web-arkivering handler om umiddelbar arkivering, mens ting foregår. 'On the fly'.

Dette er den reelle nye og fundamentale betingelse for vores håndtering af web-materiale og digitale materialer generelt, og det medfører en lang række konsekvenser for forskere og for samfundet som sådan. Det er fundamentalt, fordi det betyder, at så længe samfundet er afhængigt af digitale materialer, så må disse bevares, hvis historien overhovedet skal skrives. De bevarer ikke sig selv, og de forbliver ikke nødvendigvis stabile. Man ved aldrig om dagens materialer bliver slettet eller afgørende ændret eller pludseligt omredigeret.

Faktisk kunne man sige, at vi burde ændre vores fundamentale forståelse af computeren. Det er en maskine, vi bruger til regelbaseret datamanipulation, ja, men det er kun muligt, fordi det grundlæggende ikke er en regelbaseret maskine, men en valg-maskine (forresten døbt sådan af Alan Turing i hans skelsættende artikel fra 1936 om beregnelige tal).

Det er en maskine baseret på tilfældig adgang og hypertekstuel sammenkædning af hvilket som helst stykke information, man måtte ønske at forbinde – eller frakoble. Computeren beregner ikke; den gennemfører søgning og mønstergenkendelse, selv når den beregner. Det kan man i alt fald argumentere for – men jeg vil gemme denne argumentation til en anden lejlighed.

Altså, Røde Kors' site forsvandt tidligt i 2000, og et par uger senere – mens jeg talte med min, på det tidspunkt, nye kollega, Niels Brügger, – fandt jeg ud af, at han var vild med arkivering og planlagde at arkivere web-materialet med relation til verdensmesterskabet i fodbold, som skulle finde sted i sommeren 2000. Og endnu værre, han overvejede også at arkivere hele web'en.

På det tidspunkt var vi ved at opdage, at The Internet Archive i USA faktisk allerede var etableret, men også at det langt fra var tilstrækkeligt, når det gjaldt danske web-materialer.

I de følgende uger blev vi enige om, at der var behov for at etablere et nationalt web-arkiv, og – må jeg indrømme – vi troede også fuldt og fast, at opgaven ville kræve en ny slags institution. Vi gjorde det til et af målene for vores nært forestående etablering af Center for Internetforskning, og vi benyttede åbningen af centret i efteråret 2000 til at komme med en offentlig udtalelse om behovet for at arkivere internettet.

Vi var overbeviste om, at bibliotekerne sad fast i trykkulturen, ligesom de gamle medier gjorde. Jeg må vist undskylde. Jeg er virkelig ked af, men også glad for, at vi tog fejl, når det gælder de nationale biblioteker.

Vores offentlige udtalelse blev blandt andet citeret i *Jyllands-Posten*, og kun få dage senere blev vi inviteret til møde med Statsbiblioteket og Det Kongelige Bibliotek. De viste sig at være 'med på beatet' og har været fremragende samarbejdspartnere lige siden.

Som et første resultat blev der etableret et pilotprojekt, finansieret af DEFF – bibliotekerne leverede ekspertise i arkivering og vi påtog os opgaven at opstille kriterier for relevans set fra en forskers synsvinkel.

Resultatet var blandt andet en ny arkivstrategi, som kombinerede regelmæssige øjebliksbilleder fra det danske domæne med selektiv høst af hyppigt opdaterede sites (inkl. nyhedssites, typiske websites og særligt originale websites), suppleret med høst af særlige begivenheder, hvor materialer forventes at dukke op på nye sites og uventede sites.

Pilotprojektet blev efterfulgt af en international konference, hvor Statsbibliotekets daværende udviklingschef, Birte Christensen-Dalsgaard, og jeg blev udnævnt som medlemmer af en regeringskomité, som skulle komme med forslag til bevarelse af elektronisk kulturarv.

Idéerne, som blev udviklet under pilotprojektet, blev senere grundlaget for arkivstrategien for det nu 10 år gamle Netarkivet.dk – og pilotprojektet førte derfor direkte ind i den fremtidige lovgivning, der inkluderede etableringen af dette arkiv. Selvom samlingen langt fra var komplet, var det, så vidt jeg ved, den mest omfattende strategi, der var kendt på det tidspunkt. Bibliotekerne blev indflydelsesrige organer inden for det internationale web-arkiv-fællesskab, som det også viser sig f.eks. ved netop denne konference.

Vores samarbejde fortsatte. Vi deltog i Netarkivet.dk's redaktionsgruppe og

vi indgik i forskellige udviklings- og forskningsprojekter i de nationale danske forskningsinfrastruktur-projekter *LARM* og *DigHumlab*. I 2011 etablerede vi *NetLab* som en forskningsenhed direkte forbundet til Netarkivet.dk. I 2012 påbegyndte vi etableringen af RESAW – en europæisk forskningsinfrastruktur til studie af arkiverede web-materialer – som foreslår et "integrationsinitiativ" inden for rammerne af EU's *Horizon 2020*. Forslaget blev vurderet og anbefalet som et af emnerne med det højeste potentiale og med fortrin ved fremtidige Horizon 2020 forskningsinfrastrukturtiltag.

Arkivet udvikler sig selv, og samlingerne udvides: Nu rummer Netarkivet.dk mere end 500 terabytes. Fantastisk arbejde udføres for at udvide adgangen og søgemulighederne, og for at forbedre dokumentationen mv. Senest er der således åbnet for fritekstsøgning.

Alligevel er resultatet kun en fragmenteret samling af rester af, hvad der engang var på internettet. Vi er langt fra at have en komplet samling, endsige blot et repræsentativt udsnit. Det skyldes ikke, at bibliotekerne ikke er i stand til at udføre arbejdet. Det skyldes, at webmaterialer hører til blandt de mest komplekse digitale materialer, vi har, da de kan indeholde en række dynamiske karakteristika som scripts, hypertekst og interaktivitet. Disse arkiver vil aldrig blive komplette.

Så lad os vende blikket fremad og spørge, hvad vi skal gøre? Og spørge, om den eksisterende strategi er dækkende?

Lad mig begynde med at sige, at internettets samfundsmæssige rolle er langt mere udtalt i dag end dengang. I det 21. århundrede vil webmateriale ofte være den primære eller sågar den eneste kilde, da samfundet i dag flytter sig selv ind på web'en.

Begrundelserne for arkivering af webmaterialer er således stærkere end nogensinde, og – må jeg tilføje nu 10 år senere – det er begrundelserne for at åbne Netarkivet for offentligheden også.

Vi bliver imidlertid også nødt til at revurdere, om vores strategier er dækkende, når vi tager det faktum i betragtning, at digitalisering i dag er trængt igennem til alle dele af samfundet; fra aflæsning af det ydre rum til scanninger af vores kroppes indre og alt ind imellem, biosfæren, klimaet og kultur og samfundet inklusive.

Vi diskuterer Big Data. Nyere estimater antager, at produktionen af digitale materialer fra de seneste fem år mere end svarer til den totale mængde af information, der hidtil er produceret i den menneskelige historie. Den eksponentielle vækst i vidensproduktion dateret tilbage fra det 18. århundrede og registreret siden midten af det 20. århundrede har nået nye højder, nu hvor vi er gået ind i det 21. århundrede. Væksten i mængden af producerede digitale materialer er overvældende; men endnu vigtigere: det er ikke blot 'mere af det samme'.

Nutidens data er mere heterogene, og vi har set en mangfoldighed af nye formater til organisering af viden og nye genrer, som spænder over mobilkommunikation, sociale medier til fremkomsten af hvad vi kan betegne som *flerkilde-vidensystemer.* Nogle af dem baseret på globalt spredte kilder – nogle af dem leverer det samlede materiale i et synkroniseret format, som f.eks. nyhedsopsamlingsservices, Google-søgninger eller specialiserede services til det finansielle marked. Mange af disse systemer leveres som kommercielle

online-løsninger (Google tilbyder en vifte af sådanne) andre er etableret som ny uafhængig forskningsinfrastruktur og dataregistre på nationalt eller internationalt niveau, andre som digitaliserede enheder inden for eksisterende arkiver, biblioteker og museer, og endnu andre fremkommer inden for traditionelle medier (i Damnark f.eks. Infomedia og DR's arkiver) og der er bare en voksende række af kommercielle forskningstidsskrifts-arkiver.

Lad os indrømme, at det måske ikke er alle disse data, som er værd at bevare, men lad os også tilføje, at mange af dem er, men at ingen – ingen – i dag har et klart overblik over disse kilder eller den ringeste anelse om, hvorvidt vi faktisk bevarer de mest værdifulde data.

Faktisk har vi ikke en klar forståelse for variationen af formater, og de måder de kan blive analyseret eller visualiseret på, eller på den såkaldte "messiness" i disse datasæt.

Nu må jeg komme til min jubilæumsgave. Det er en idé gemt i et spørgsmål. I får det gratis. Spørgsmålet er nu 10 år senere, hvorvidt det er tiden for Netarkivet og dets forældre at bede om en mere gennemgribende afbildning af hele det digitale univers og stille det fundamentale spørgsmål: Hvilke dele af vores digitale arv kunne overlades til private services med risiko for, at de forsvinder eller kun er bevaret som en del af en bestemt forretningsmodel, og hvilke dele kunne gøres til generel public service, der bevares til kommende generationer?

Jeg indrømmer, det er en meget enkel idé, kun et spørgsmål. Men hvorfor ikke prøve at besvare det?

Det er min gave. Ikke fordi I ikke allerede har gjort et fantastisk stykke arbejde, tværtimod. Jeg stiller det præcis til jer, fordi I har gjort det og derfor også kvalificerer jer selv til at tage næste skridt.

Tillykke – til Netarkivet og forældrene, Universitetet og Statsbiblioteket og Det Kongelige Bibliotek. I har gjort et fantastisk arbejde og åbnet sporet til fremtidige studier af fortiden og nutiden. Og vi, som forskere, må vedstå at vi er blevet forelskede i vores digitale børn. Vi behøver dem. Og i fremtiden kan vi ikke leve uden.

Tak.

Talen er oversat fra engelsk.