Webarkivering

Af Birgit Nordsmark Henriksen

> The Universe, as has been observed before, is an unsettlingly big place, a fact which for the sake of a quiet life most people tend to ignore.
>
> Douglas Adams, *The Restaurant at the End of the Universe*

Udskiftes 'universet' med 'webben' i dette citat, kunne sætningen stadig være sand. Webben, som vi kender den i dag, er så stor, at vi i det daglige ignorerer det. Det er først, når vi forsøger at gøre noget, der omslutter den, som f.eks. at arkivere den, at vi må se i øjnene, at den er enorm, og at vi har sat os selv en voksen opgave for.

Internettet giver mulighed for en lang række services som f.eks. e-mail, instant messaging, fildeling, chat og nyhedsgrupper, og webben er således blot en af internettets mange services. Arkivering af internettet omfatter derfor langt mere end arkivering af webben. Denne artikel omhandler imidlertid kun arkivering af webben og vil forsøge at give et overblik over, hvad det er for en opgave, man står overfor, hvis man vælger at arkivere dele af den. Artiklen vil derudover give et vue over webarkiveringens udvikling i Danmark og det øvrige Norden, og hvordan denne udvikling nu orienterer sig i international retning.

I begyndelsen af 1990'erne var internettet og webben stadig et svært tilgængeligt forskningsredskab, og det var først ved fremkomsten af de grafiske browsere (Mosaic fra NCSA i 1993 hhv. Netscape i slutningen af 1994), hvor det blev muligt at surfe og se billeder, at internettet ændrede karakter og blev det alment tilgængelige og vidt udbredte medie, vi kender i dag.

Webben indeholder den absolut største kendte samling af informationer, (u)organiseret i en stor blanding af informationsdatabaser, publikationer, kom-

munikation, personlige tilkendegivelser og det, vi traditionelt kalder 'småtryk' som f.eks. varekataloger, annoncer og medlemsblade. Den udgør dermed et væsentligt element af vores nutidige kulturarv, som vi ønsker at dokumentere og bevare for vores eftertid.

Webben er ikke så ligetil at arkivere af flere årsager. Jeg vil i det følgende blot fremhæve de mest fremtrædende:

Først og fremmest viser alle undersøgelser, at indholdet ændrer sig med stor hast:[1]

- halvdelen af nettets websider er mindre end 100 dage gamle
- kun 20 % af de sider, der er tilgængelige i dag, er det også om et år
- 40 % af websiderne i .com-domænet ændrer sig dagligt
- der dannes 8 % nye sider om ugen, og de største ændringer sker gennem tilvækst af nye sider og ikke gennem ændringer i eksisterende
- den samlede volumen er meget stor. Google, der indekserer større dele af nettet, indekserede således 1 mia. websider i 2000, 3 mia. i 2002 og 8 mia. i 2004.

Dertil kommer, at webbens organisering ændrer sig endnu hurtigere, idet linkstrukturerne er væsentlige mere dynamiske end dokumenterne. Efter et år er 80 % af alle links på webben erstattet med nye. Dette betyder f.eks., at håndtering af dubletter er en stor udfordring.[2]

Store dele af webbens materiale findes på det, der kaldes den dybe web: Materiale, som befinder sig i søgbare databaser, som kun kan nås gennem direkte søgeforespørgsler, og hvor resultatet leveres som dynamiske websider, genereret *on-the-fly* på unikke, men ikke-blivende webadresser. Andre problemer i forbindelse med den dybe web kan være passwordbeskyttelser, personaliseringer eller konstruktioner som f.eks. kalendere, der kan trække indsamlingsværktøjerne ind i uendelige løkker.[3]

Endelig udvikler webben sig teknisk hele tiden mht. til såvel dataformater som anvendelses- og interaktionsformer, hvorfor indsamlingsmetoder og -værktøjer løbende skal justeres for at tilpasse sig disse ændringer. Således dukkede f.eks. streaming af lyd og video op i hhv. 1995 og 1997, og samtidig blev det muligt at afvikle multimedieprogrammer med animationer og vektorbaseret interaktion bl.a. ved hjælp af Macromedia's Shockwave i 1995 og Flash i 1997. Onlinehandel og netauktioner slog for alvor igennem i 1998, og i maj 2004 gennemførtes den største onlineafstemning nogensinde, da 370 mio. af Indiens 675 mio. vælgere stemte elektronisk til delstats- og nationalvalg.

Hvad skal der til for at skabe den historiske web?

Der er mindst tre indsatsområder, der kommer i spil, når man beslutter sig for at arkivere webben: Først skal det materiale, man vil arkivere, lokaliseres og indsamles. Dernæst skal det bevares, og endelig skal det gøres tilgængeligt.

I webarkiveringssammenhænge praktiseres i dag to radikalt forskellige politikker for indsamling: Enten begrænser man det, man indsamler, ud fra et sæt af udvælgelseskriterier, eller også forsøger man at indsamle så meget som muligt i et forsøg på at undgå at begrænse eftertidens muligheder med materialet mere end højst nødvendigt. Uanset hvilken politik man vælger, vil man dog ikke kunne få det hele med. Webbens konstante foranderlighed samt de tekniske begrænsninger i indsamlingsværktøjerne gør, at man altid kun vil arkivere øjebliksbilleder. Den indsamlingsstrategi, man vælger, skal derfor tage højde for, at man opnår den repræsentativitet og den volumen, der er behov for i forhold til arkivets formål.

Da ophavsretsloven betyder, at man skal indgå aftaler med rettighedshaverne, hvis man arkiverer kopier af digitalt materiale i institutionelle rammer som f.eks. et nationalbibliotek, vil man oftest finde en meget selektiv indsamlingspolitik dér, hvor denne ikke kan sikres af en pligtafleveringslov, ganske enkelt fordi alt andet bliver praktisk uoverkommeligt. For hvert website eller webside, der indsamles, skal rettighederne afklares og beskrives, hvorfor der ofte følger omfattende registrerings- og katalogiseringsaktiviteter med. I de lande, hvor indsamlingen kan sikres af en pligtafleveringslov, ser man oftest indsamlingspolitikken suppleret med den anden tilgangsvinkel, motiveret af et ønske om at være så dækkende som mulig. Danmark og New Zealand er således lande, som via pligtafleveringslove forsøger at sikre tilstrækkelig bred indsamlingspolitik gennem anvendelse af flere parallelle indsamlingsstrategier.[4]

Tre indsamlingsinitiativer startede i 1996 i Sverige, Australien og USA (det private initiativ, Internet Archive (IA) se herom nedenfor). I de øvrige nordiske lande påbegyndtes aktiviteter i 4-års perioden frem til årtusindskiftet, og i de seneste fire år har vi set en lang række nye initiativer overalt i verden: USA (Library of Congress), Storbritannien, Frankrig, Slovenien, Tjekkiet, Østrig, Japan, Litauen, New Zealand og senest Kina og Grækenland. Indtil dato er det kun IA, som forsøger at indsamle den globale web – de øvrige initiativer fokuserer på områder af webben, der har særlig national interesse, skønt alle har vanskeligheder ved at afgrænse det nationale fra helheden. Der findes endnu ikke et overblik over hvilke dele af webben, der faktisk arkiveres hvor og af hvem.

To forskellige indsamlingsmetoder praktiseres i dag: Man indsamler i bredden, for at have så repræsentativt et billede af nettets indhold som muligt, eller man indsamler i dybden, for at have så præcist et billede som muligt af, hvad der var på et givet site på et bestemt tidspunkt. De to indsamlingsmetoder indgår på forskellig vis i de tre mest udbredte indsamlingsstrategier:

1. Tværsnitshøstning, som går ud på at give et øjebliksbillede af hele webben, og som anvender breddehøstning
2. Den selektive indsamling, som går ud på at have meget høj dækning på nogle udvalgte områder eller websites, hvorfor man anvender dybdehøstning eller afleveringer
3. Den begivenheds- eller tematisk orienterede høstning, hvor man forsøger at dække så meget som muligt om en bestemt begivenhed eller et bestemt emne under anvendelse af en kombination af bredde- og dybdehøstning.

Alle tre høstningsstrategier har været anvendt og vil blive omtalt senere i artiklen.

Ud over indsamling er viden om benyttelsen af materialet også en vigtig information at gemme for eftertiden. For trykt materiales vedkommende kan oplagstal i nogen grad afspejle udbredelsen. Når informationer er publiceret på webben, har i princippet alle adgang til det, og brugen bliver derfor mere relevant end udbredelsen. Megen af den information, der findes på webben i dag, er måske aldrig set af andre end den, der publicerede det. Således har Microsoft et kæmpe website med dokumentation af al deres software, samtidig med at kun en meget lille del heraf benyttes af brugerne. Benyttelsesmønstre i en eller anden form vil muligvis skulle indsamles eller etableres sammen med webmaterialet, hvis forskere i fremtiden skal vide noget om materialets brug.

Bevaringsopgaven består af to grundlæggende opgaver, der begge skal udføres for, at den samlede opgave er løst: Først skal man have en sikker opbevaring af de bitstrømme, man indsamler, og dernæst skal man sikre, at man nu og i tiden fremover kan fortolke og tilgængeliggøre de opbevarede data i takt med, at dataformater og tilgængeliggørelsesprogrammer ændrer sig. Det første kalder man bitbevaring – det sidste logisk bevaring. Stort set alle webarkiver udfører i dag udelukkende bitbevaring, men fokus nu og de kommende år bliver i fællesskab at finde metoder og løsninger til at gennemføre den logiske bevaring, herunder at afklare rækkevidden af opgaven: Bevarer vi kun de nøgne informationer, eller fokuserer vi også på at bevare deres organisering, grafiske layout, funktionalitet(er) eller deres „look & feel".[5]

I et webarkiv er der behov for et værktøj til at gennemføre den manuelle inspektion i arkivet, der skal sikre og kontrollere, at kvaliteten af det indsamlede materiale er på plads. Ret hurtigt derefter opstår et behov for at kunne lade brugere få adgang til den historiske web på en måde, så man får en oplevelse, der så godt som muligt ligner den, som brugere fik på indsamlingstidspunktet. Hertil kræves et program, som både giver arkivmaterialet mulighed for at simulere den indsamlede del af webben, og som giver brugeren mulighed for at lokalisere materiale gennem en struktureret adgang. Udformningen af dette adgangsprogram afhænger bl.a. af det ambitionsniveau, man lægger i omfanget af adgang: Skal man f.eks. blot slå en URL op og kunne se de informationer, der har været formidlet på denne web-adresse over tid? Skal man kunne navigere i det historiske webarkiv i tid og rum med fuldt bevaret funktionalitet ved hjælp af værktøjer svarende til dem, vi kender fra det aktive net? Eller skal det være muligt at lave store datamining-operationer på materialet til forsknings- og statistikformål? Udformningen af adgangsprogrammet vil også være påvirket af, hvorledes man har valgt at gennemføre sin logiske bevaring – om man har satset på at flytte de arkiverede bitstrømme til dataformater, der fortsat understøttes af de gængse tilgængeliggørelsesprogrammer (migrering), eller om man har satset på at etablere en fremtidssikret understøttelse af de oprindelige dataformater (emulering).

Den disciplin, som kommer tættest på webarkivering, er uden tvivl søgemaskinernes indeksering af det aktive net. Siden lanceringen af AltaVista i 1995 har der været en heftig kamp mellem forskellige initiativer på dette område, oftest opstået i universitetsmiljøerne for derefter hurtigt at flytte over i det kommercielle område. Det mest succesfulde til dato er uden tvivl Google, der via sine særlige ranking mekanismer siden 1998 har gjort det praktisk muligt for brugere at finde relevant materiale blandt de mere end 8 mia. websider, som de i slutningen af 2004 indekserer. Webindeksering og indsamling af webmateriale har flere fælles problemstillinger. Behovet for et effektivt værktøj til at kunne 'høste' webben ved at følge de enkelte dokumenters links samt behovet for gode algoritmer til at kunne lokalisere det materiale, der mest sandsynligt har ændret sig, siden sidst man høstede, er blot nogle af dem. Men hvor søgemaskinernes formål er at lokalisere materialet for at kunne vise hen til det, når efterspørgslen opstår, er arkivernes behov helt anderledes. Her ønsker man i så høj grad som muligt at kunne give brugeren af arkivet et autentisk billede af, hvad der var på et website på et bestemt tidspunkt, og man tilstræber derfor, at der f.eks. er konsistens mellem de arkiverede versioner af et websites enkelte sider.

Webarkivering i Danmark

Bevaring af kulturarven er i fokus her ved årtusindskiftet – ikke bare internationalt, men også i Danmark. De seneste danske eksempler herpå er Kulturministeriets oprettelse af Kulturarvsstyrelsen i 2003 samt Kulturministeriets kulturarvsudredning for hhv. den fysiske og den digitale kulturarv samme år.[6]

Den danske pligtafleveringslov havde oprindelig til formål at sikre gratis kopier til udvekslingsformål af alt trykt materiale til rigets monark, men først 300 år senere i forbindelse med lovrevisionen fra 1997 blev det slået fast (i bemærkningerne), at loven har til hensigt at bevare den del af den danske kulturarv, som udgøres af publicerede værker. Både pligtafleveringsloven og dens formål har været revideret flere gange siden 1697. Nogle af de senere lovrevisioner har været tæt forbundet med ændringer i anvendte publiceringsteknikker. Lovændringen i 1902 fandt sted med den industrielle revolutions indtog i trykkeindustrien. I 1994 påbegyndtes arbejdet med at udforme en ny pligtafleveringslov, der havde til formål at indarbejde de ændringer, som karakteriserede informationssamfundet, herunder elektronisk publicering via databaser og internettet. Den nye lov i 1997 blev, bl.a. efter lobbyvirksomhed fra it-branchen, begrænset til kun at omfatte statiske netpublikationer (afsluttede og uafhængige dokumenter, hvor mindst et eksemplar har været sat til salg eller har været distribueret til offentligheden – uanset medie) og ikke internettets eller webbens indhold som helhed.[7]

I 1994 slog nettet igennem herhjemme. 10 år senere oplyser Danmarks Statistik:[8]

- At 98 % af danske virksomheder har internetadgang, 81 % er på nettet med en hjemmeside, 27 % modtager ordre fra deres kunder via nettet, og 19 % anvender det til e-læring.
- At de statslige myndigheder i 64 % af tilfældene vurderer, at de i høj eller nogen grad har digitaliseret de blanketter, der er rettet mod borgere og virksomheder og gjort dem tilgængelige via nettet.
- At antallet af brugere med adgang til nettet steg hurtigt efter 1995, hvor en række nystartede, små internetudbydere begyndte at sælge grafisk internetadgang til private, således at 2/3 af alle danske familier nu har adgang til nettet hjemmefra:

Årstal	**1996**	**1997**	**1998**	**2000**	**2001**	**2004**
% familier med adgang til nettet	5	10	22	45	60	66

- At hele 96 % af alle unge mellem 16 og 19 år, ca. 90 % af de 20-59 årige, men kun 54 % af dem, der er ældre, har adgang til nettet på en eller anden måde.
- At 65 % af den danske befolkning bruger nettet til kommunikation, 70 % til informationssøgning, 55 % til køb/salg og bankforretninger, 43 % til kontakt med offentlige myndigheder, 17 % til kurser og undervisning, 16 % til jobsøgning.

Som omtalt tidligere viser undersøgelser, at websider ikke har så lang en levetid på nettet. Hvis nogen på et senere tidspunkt vil gå tilbage og følge de ovennævnte udsagn fra Danmarks Statistik for at se, hvad det var for en ny kultur, disse mange danskere var en del af i perioden 1994-2004, vil man derfor være afhængig af at have adgang til indsamlet dansk materiale. Dette gælder, hvis man f.eks. vil undersøge e-handelens etablering og udbredelse eller den indflydelse, jobportalernes indmarch i 1996 fik på processerne omkring jobsøgning og rekruttering.

Indsamling og arkivering af netmateriale finder p.t. sted i mange sammenhænge, bl.a. som en integreret del af forskningsresearch eller forvaltningspraksis i den offentlige sektor. I perioden 1994-2004 er der imidlertid kun foretaget systematisk indsamling af dansk materiale i to sammenhænge, hvor der gives offentlig adgang til det indsamlede.

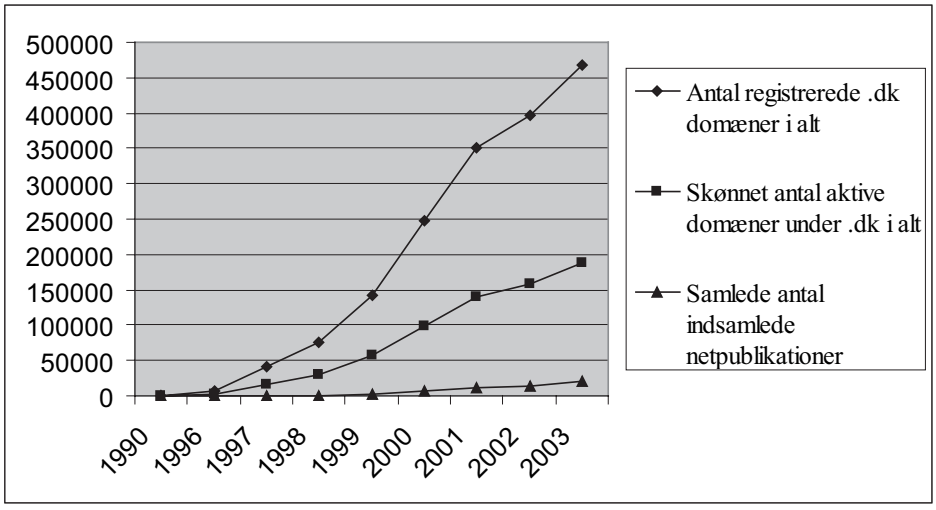
IA har fra 1996, som en del af deres arkivering af den globale web, indsamlet spredte dele af danske websites. Det skønnes, at IA i 2004 akkumuleret har ca. 0,5 Tbyte data fra .dk-domænet liggende i deres arkiv. Da vi i Danmark skønner, at en enkelt tværsnitsarkivering af det danske .dk domæne i 2004 fylder ca. 2 Tbyte, er der således tale om en mindre, om end historisk meget vigtig, samling. IA giver online adgang til det indsamlede danske materiale på linie med alt andet materiale med 6 måneders forsinkelse via deres website web. archive.org.

Webarkivering har været praktiseret i Danmark siden 1997 som følge af den reviderede danske pligtafleveringslov fra 1997. Allerede fra 1. januar 1998 skulle danskere, der publicerede et afsluttet værk på nettet, anmelde dette via et nyoprettet website: www.pligtaflevering.dk, hvorefter netpublikationen blev indsamlet og arkiveret to steder: I et værkarkiv på Det Kongelige Bibliotek og tilsvarende på Statsbiblioteket i Århus. Det Kongelige Bibliotek og Statsbiblioteket har givet adgang til de pligtafleverede netpublikationer via dedikerede maskiner på deres læsesale.

Af de godt 21.000 værker, der fra 1998 til 2003 er anmeldt og som konsekvens heraf indsamlet, udgør ca. 1/3 monografier og ca. 2/3 numre fra ca.

750 periodicatitler. Blandt de anmeldte værker findes kun meget få fra .org- og .com-domænerne. Erfaringer fra Sverige viser, at 25-40% af svenske websider ligger uden for.se-domænet, og der er grund til at formode, at dette også gælder for det danske materiale, herunder statiske netpublikationer, jf. tabellen med de svenske tal senere i artiklen.

Udviklingen i .dk-domæner 1990-2004 og indsamlede netpublikationer 1998-2003[9]



Som det fremgår af figuren, blev lovrevisionen i 1997 gennemført på et tidspunkt, hvor den danske del af webben udviklede sig meget hurtigt. Analyser gennemført af Danmarks BiblioteksCenter (DBC) og Det Kongelige Bibliotek/ Statsbiblioteket har vist, at kun ca. 42 % af de registrerede danske domæner er aktive, hvilket betyder, at der i juli 2004 var skønnet ca. 213.000 aktive domæner mod ca. 17.000 i 1997, hvor den nye pligtafleveringslov blev vedtaget.[10] Det Kongelige Biblioteks og Statsbibliotekets egne stikprøver i 2004 viser, at langt den overvejende del af de aktive domænenavne tilhører private virksomheder – en sektor, hvis netpublikationer kun i ringe grad er repræsenteret i de eksisterende værkarkiver.

Den kraftige vækst skete imidlertid ikke kun volumenmæssigt, men også indholdsmæssigt – et forhold der hurtigt blev klart for både de institutioner, der stod bag indsamlingen, og for nogle af de forskere, der har en interesse i, at indsamlingen finder sted. Dette førte til, at man i 2001 igangsatte en række initiativer med det formål, at få revideret pligtafleveringsloven endnu engang.

I juni 2001 afholdtes den første internationale webarkiveringskonference i Danmark, finansieret af Danmarks Elektroniske Forskningsbibliotek (DEF): *Preserving the Present for the Future – Strategies for the Internet*. Konferencens formål var at bringe udenlandske forskere og biblioteksfolk med erfaring inden for webarkivering sammen med danske forskere, biblioteksfolk og teknikere for at skabe en debat om forskellige indsamlingsstrategier. Det viste sig overraskende under konferencen, at der var enighed om, at der var behov for afvikling af flere parallelle indsamlingsstrategier for at opnå en dækkende indsamling og sikre vigtigt kildemateriale til fremtidens forskere; de, der foretog selektiv indsamling, efterspurgte tværsnitshøstning og omvendt. Det eneste, der afholdt de enkelte aktører fra at udvide aktiviteterne, var økonomien. Flere af de danske indlæg tog fat på det forhold, at nettet teknisk og indholdsmæssigt havde udviklet sig meget kraftigt i perioden fra tidspunktet for loven i 1997 og frem, hvilket betød, at store dele af denne danske kulturarv slet ikke var blevet indsamlet og dokumenteret: områder som e-handel, onlinemedier og udviklingen af webben til et interaktivt kommunikationsmedie var blot eksempler herpå. De ovenfor refererede tal fra Danmarks Statistik dokumenterer de karakteristika ved den almindelige borgers brug af nettet, som flere af konferencens indlæg fremhævede, var fuldstændigt fraværende i den eksisterende webarkivering.[11]

Konferencen blev fulgt op af et etårigt pilotprojekt: *Netarkivet* i 2001-2002, ligeledes finansieret af DEF. Projektet skulle gennem konkrete indsamlingsforsøg i forbindelse med kommunalvalget november 2001 undersøge, hvordan en kommende og mere dækkende indsamlingsstrategi i Danmark kunne udformes og realiseres. Projektpartnerne var Center for Internetforskning, Aarhus Universitet, Det Kongelige Bibliotek og Statsbiblioteket. Projektet var et af de første webarkiveringsprojekter, der inddrog forskere, i dette tilfælde medieforskere, i det indledende indsamlingspolitiske arbejde. Resultatet blev et forslag om en hybrid strategi bestående af 4-årlige tværsnitshøstninger, intensiv indsamling af ca. 80 særligt udvalgte sites og 1-2 årlige begivenhedsorienterede indsamlinger. Navnlig erfaringerne fra den sidste type indsamling (kommunalvalget) betød, at forskerne lagde vægt på, at indsamlingen burde udvides til at omfatte andre dele af internettet end webben: Shoutboxes, chat, quickpools og online-spil viste sig at indgå i politikernes forsøg på at fange de unges interesse.[12]

Projektets arbejde indgik efterfølgende i Kulturministeriets arbejde i forbindelse med *Udredning om bevaring af kulturarven*, april 2003. Her foreslås en ændring af pligtafleveringsloven for materiale i elektroniske kommunikationsnet, da kun en del af de afleveringspligtige materialer er blevet afleveret, dels

fordi det i loven af 1997 påhviler fremstilleren selv at anmelde udgivelsen af et afleveringspligtigt værk, dels fordi lovens sondring mellem statisk og dynamisk internetmateriale ikke længere er meningsfuld. Lovens udgivelseskriterium for afleveringspligt kombineret med lovens definition af værksbegrebet har bevirket, at kun afsluttede, statiske værker, der udgør en stadig mindre del af nettet, er omfattet af afleveringspligten. Udredningen anbefaler derfor, at alt materiale offentliggjort i elektroniske net for fremtiden bør være afleveringspligtigt. Offentliggørelse forstås her på samme måde som i ophavsretsloven, hvor et værk anses for offentliggjort, når det lovligt er gjort tilgængeligt for almenheden med eller uden modydelser som f.eks. betaling eller oplysning af identitet.

Det var udredningens vurdering, at den foreslåede indsamlingsstrategi vil sikre en volumen og en repræsentativitet, der er forsvarlig ud fra et forskningsog bevaringssynspunkt. Udredningen anbefaler, at indsamlingen afgrænses ved hjælp af Danica-princippet, dvs. 1) at materiale der offentliggøres fra .dk domæner, 2) er udarbejdet af danskere på dansk, på andet end dansk eller som fremføres af danske kunstnere fra andre domæner end danske domæner eller endelig 3) materiale om Danmark fra andre domæner end danske domæner, indsamles.

Udredningen blev fulgt op af endnu et projekt i 2003-2004, finansieret af Kulturministeriet med det formål, at Det Kongelige Bibliotek og Statsbiblioteket kunne levere den nødvendige viden, herunder også økonomi, til ministeriet med henblik på en revision af pligtafleveringsloven. For både indsamling, bevaring og tilgængeliggørelse af det danske internet lykkedes det at finde brugbare og realistiske løsninger på tekniske problemer og organisatoriske spørgsmål i forbindelse med forskellige indsamlingsproblematikker. Ny høstersoftware, der er i stand til at indsamle internetmateriale i arkivkvalitet, blev udviklet i internationalt samarbejde. På bevaringssiden blev der udarbejdet et teknisk arkivformat, der efterfølgende blev koordineret med International Internet Preservation Consortium (IIPC). I forbindelse med færdiggørelsen af Statsbibliotekets og Det Kongelige Biblioteks fælles bryllupsgave til kronprinseparret – en samling af relevante websider indsamlet i ugerne omkring brylluppet 14. maj 2004 – udviklede de to biblioteker en enkelt løsning for tilgængeliggørelse, der efterfølgende har vakt stor interesse hos en række andre nationale internetarkiver.

I november 2004 fremlagde Kulturministeriet en revision af pligtafleveringsloven og ophavsretsloven i Folketinget. De reviderede love giver de to biblioteker mulighed for at indsamle alt materiale fra den offentlige del internettet. Kun dokumenter fra områder, der betragtes som ikke-offentlige, dvs. områder,

hvortil man enten kun har tjenstlig adgang som f.eks. intranet, eller hvortil man skal inviteres, vil ikke være omfattet af pligtaflevering af netmateriale. Som noget nyt bliver oplysninger om registrerede danske domæner afleveringspligtige med henblik på, at man kan lokalisere alle sites, der er omfattet af lovgivningen.

Anderledes restriktivt ser det ud mht. adgang til det indsamlede. Da de nye indsamlingsmetoder ikke kan forhindre, at der indsamles særligt personfølsomme data, kan der ikke gives almen adgang til det indsamlede. Derfor vil kun forskere med godkendte forskningsprojekter kunne forvente at få adgang til det indsamlede materiale i den nærmeste fremtid. Pligtafleveret materiale sidestilles hermed for første gang med arkivmateriale og underlægges samme restriktive adgang.

I Danmark har man gennem pligtafleveringen af netpublikationer siden 1999 forsøgt at understøtte udbredelsen af Dublin Core Metadata i statiske dokumenter på internettet ved at lette anmeldelsen af dokumenter, der indeholdt disse data. Disse metadata er senere blevet genanvendt både ved strukturerede søgefaciliteter i arkivet samt som supplement til den registrering af netpublikationerne, der finder sted hos DBC til nationalbibliografien. Det er endnu ikke besluttet, hvorledes en sådan registrering skal videreføres ved omlægning af arkiveringen.

Webarkivering i Norden

I 1993-1994 gennemførte Lunds Universitetsbibliotek og Danmarks Tekniske Bibliotek et Nordinfo-sponsoreret projekt, hvor man udviklede en ny service til søgning i webbaserede informationsressourcer, der på det tidspunkt kun kunne lokaliseres via browsing. Projektet blev fulgt op af endnu et nordisk projekt i 1996-1997, Nordic Web Index, der resulterede i en eksperimentel søgetjeneste fra 1997 til 1999, hvor alt nordisk webmateriale blev høstet af den egenudviklede høster, 'Combine', indekseret og gjort tilgængeligt fra fem nationale tjenester på nettet. Projekterne lagde grunden til nordisk forskningsbiblioteksdeltagelse i en lang række EU-projekter, men først og fremmest blev de katalysator for udviklingen af webarkivering i nordisk regi.

I oktober 1996 igangsatte Kungliga Biblioteket i Stockholm (KB.se) et projekt, som havde til formål at teste metoder til indsamling og bevaring af publicerede svenske elektroniske dokumenter. Biblioteket havde konstateret, at mens man havde indsamlet stort set alle trykte svenske aviser siden 1643, var man ikke i besiddelse af de første svenske webbaserede aviser fra 1993-1995.

Kulturarw[3], som projektet hed, var det første projekt i nationalbiblioteksregi, hvor man tværsnitshøstede netmateriale.

Som det fremgår af tabellen nedenfor, vokser også det svenske net hurtigt, og mængden af materiale, der indsamles, vokser betragteligt. Det samme gør antallet af forskellige dataformater, man skal håndtere. Her otte år efter projektets start tværsnitshøster KB.se stadig to gange årligt den svenske del af nettet og har oparbejdet det til dato største webarkiv i nationalbiblioteksregi.

Oversigt over volumen i webarkiveringen i Sverige[13]

	1997 (forår)	1998 (forår)	1999 (forår)	2000 (sommer)	2001 (sommer)	2003 (efterår)
#arkiverede websites i.se-domænet	15.675	26.342	32.325	48.528	56.598	76.948
#arkiverede websites uden for.se		7.254	20.189	39.072	68.997	58.760
Data-mængde pr. tværsnits-høstning i GByte	161	196	258	773	1.132	2.088
# forskellige filtyper	295	323	351	438	424	456

I forbindelse med et NORDUnet møde i Island i sommeren '97 og en tilstødende Nordic Cooperative Indexing Workshop opstod der enighed om at nedsætte et uformelt forum for teknisk samarbejde og erfaringsudveksling om bevaring af elektroniske publikationer. Dette uformelle forum navngav man Nordic Web Archive (NWA), og man mødtes regelmæssigt i de følgende år. Man drøftede de aktuelle indsamlingsprojekter i Sverige (Kulturarw[3], 1996ff), Finland (Eva, 1997ff) og Danmark (pligtaflevering af netpublikationer, 1997ff), og man påbegyndte arbejdet med at specificere nyt indsamlingssoftware, der i modsætning til f.eks. Combine, som anvendtes i Sverige, skulle målrettes til arkiveringsopgaverne.

NEDLIB-høsteren, som blev udviklet på Helsinki University Library og CSC i Finland i forbindelse med EU-projektet NEDLIB, så dagens lys i 2000 og var den første høster, der blev udviklet til arkiveringsformål.[14] Denne og senere releases har været anvendt på eksperimentel basis af en lang række nationalbiblioteker i årene 2001-2003.

I oktober 1998 formaliseredes NWA i de nordiske nationalbibliotekarers regi. Man traf beslutning om at påbegynde et fælles projekt og valgte at fokusere på udvikling af værktøjer, så man kunne søge og navigere i det indsamlede materiale. Strategien var at eksportere indholdet fra de forskellige webarkiver til et fælles format, hvorefter alle disse dokumenter kunne indekseres og gøres søgbare til fritekst. Ved at anvende samme format og samme indekseringssoftware fra Fast, der muliggjorde distribuerede indekser, var det intentionen, at det skulle være muligt at navigere på tværs af alle de nordiske landes nationale webarkiver.

Udviklingsopgaven kom til at strække sig over to projektperioder i årene 2000-2002 og 2003-2004 og blev støttet af Nordinfo og Nordinfo's projekt, 'NORDUnet2'. Resultatet blev 'NWA Toolset' – et sæt af softwareværktøjer til søgning og navigering i tid og rum blandt arkiverede webdokumenter. Værktøjet blev frigivet i oktober 2004 som Open Source.[15]

Billedet er et eksempel på fremvisning af en webside ved hjælp af NWA Toolset, hvor der er mulighed for at følge et dokuments udvikling over tid.

Samtlige nordiske lande har således siden 1996 afviklet forsøg med henblik på at bringe sig teknisk i stand til at indsamle relevante dele af internettet ved hjælp af automatiserede høstninger som en del af de respektive landes pligtafleveringsaktiviteter. Lovgivningen er ikke alle steder fuldt på plads endnu. Norge (1989), Island (2003) og meget snart også Danmark (2004) har fået en lovgivning på plads, som sikrer en indsamling. KB.se har indsamlet svensk internet materiale via høstning siden 1996-1997 og fik lovliggjort sine indsamlinger i 2002. Den svenske pligtafleveringslov er dog endnu ikke på plads mht. det digitale internetmateriale. Finland forventes først i 2005 at have en revision af pligtafleveringsloven klar, der omfatter høstning af det finske internet.

Mht. tilgængeliggørelse er der stadig et stykke vej til, at der kan gives fri adgang via de udviklede værktøjer. Således har kun Island p.t. mulighed for at give fri online adgang til det indsamlede materiale. I Sverige kan alle få adgang til alt det indsamlede materiale ved at møde op på KB.se's læsesal, mens man p.t. ikke kan få adgang til det indsamlede i Norge. Også i Danmark synes adgangen at blive begrænset til forsknings- og statistikformål. I såvel Norge som Danmark begrundes den restriktive eller manglende adgang med hensynet til persondataloven.

Webarkivering uden for Norden

Verdens største webarkiv med webmateriale fra hele verden findes hos det amerikanske, privatejede non-profit Internet Archive (IA) i San Francisco, som blev grundlagt i 1996 af Brewster Kahle. Arkivet er bygget på dataleverancer fra først og fremmest Alexa Internet, et site til webnavigering (katalogisering af webressourcer), der ligeledes blev etableret i april 1996 af Brewster Kahle og Bruce Gilliat. Fra september 2002 findes en kopi af arkivet på Bibliotheca Alexandrina i Egypten, og fra maj 2004 endnu en kopi i Amsterdam. IA havde efter salget af Alexa Internet til Amazon.com i 1999, og dette firmas partnerskab med Google i 2002, et behov for på sigt at kunne gøre arkivets opbygning uafhængig af dataleverancer fra Alexa. Derfor startede man selv i foråret 2003 udviklingen af en ny høster, *Heritrix*.

IAs webarkiveringsaktiviteter 1997-2004

Tidspunkt	Foråret 1997	Marts 2000	Marts 2001	2002	2004
Størrelse		1 mia. sider, 13.8 Tbyte	4 mia. sider, 40 Tbyte	120 Tbyte	500 Tbyte
Vækstrate	Ca. 1 Tbyte / Måned	2 Tbyte / måned		12 Tbyte / måned	30 Tbyte / måned

Nordic Web Archives krav til en ny høster blev nu bragt ind i debatten med IA og en række andre nationalbiblioteker og resulterede i, at de nordiske lande i efteråret 2003 valgte i fællesskab at sende to af bibliotekernes softwareudviklere et halvt år til IA for at fremme udviklingen af *Heritrix*.[16] Udviklingen går støt fremad, og *Heritrix* anses nu for at være den høster, som alle nationalbiblioteker vil basere deres indsamling på. Her i Danmark har vi også forsøgsvis anvendt den i forbindelse med den indsamling af websider, som Statsbiblioteket og Det Kongelige Bibliotek foretog i ugerne omkring kronprinsebrylluppet 14. maj 2004, og som senere blev overrakt kronprinseparret i gave.

Mange andre, først og fremmest nationalbiblioteker, har afviklet webarkiveringsprojekter i de senere år, og jeg skal her blot nævne nogle få af dem, vi her i Danmark har kontakt til.[17]

Tidligst ude var National Library of Australia (NLA), der siden 1996 har foretaget en selektiv indsamling af mere end 7000 titler/websites.

I Storbritannien indførtes muligheden for pligtaflevering af digitalt materiale ved en ny lov af 31. oktober 2003, og i juni 2004 etableredes et nationalt Web Archiving Consortium, (UKWAC) bestående af seks institutioner, der forsøgsvis skal forestå selektiv indsamling af forskellige typer af websites, foreløbig 6000, regelmæssigt over de næste to år, baseret på harvesting efter aftaler med siteejerne.[18]

I Frankrig har man også afviklet mindre arkiveringsforsøg siden 2001. 12. november 2003 fik man en ny pligtafleveringslov, hvor Bibliothèque nationale de France (BnF) får rettigheder til at indsamle nettet. BnF er netop i efteråret 2004 gået sammen med British Library (BL) om en videreudvikling af *Heritrix*-høsteren, så den kan anvendes til at lokalisere og indsamle dele af de nationale webber vha. særlige prioriteringsmekanismer.[19] Derudover har BnF i 2004 indgået en 2-årig aftale med IA, hvor IA gennem forsøg med forskellige indsamlingsstrategier skal finde den bedste måde at lokalisere webmateriale med relation til Frankrig.

Library of Congress (LC) har siden 2000 samarbejdet med først og fremmest

IA, men også med forskergruppen Archivist.org og computerfirmaet Compaq om opbygning af begivenhedsorienterede websamlinger ud fra en filosofi om, at netop disse dele på webben var i særlig fare for at forsvinde hurtigt. Det første projekt var præsidentvalget i 2000, senere fulgte en samling om 11. september-tragedien, hvor man arkiverede fra 30.000 websites i perioden september til december 2001 for derefter at katalogisere 2300 af disse websites.[20] Senest er LC begyndt at undersøge mulighederne for at opbygge tematiske samlinger om f.eks. terrorisme og sundhedsvæsen.

Gennem IA's samarbejdet med LC, havde IA erkendt, at de havde et interessefællesskab med de nationalbiblioteker, som var begyndt at interessere sig for indsamling fra nettet: IA havde den tekniske knowhow, og nationalbibliotekerne havde stor ekspertise i samlingsopbygning. Dette førte i 2002 til, at IA rettede henvendelse til flere nationalbiblioteker, heriblandt BnF og Det Kongelige Bibliotek, for at sondere mulighederne for et tættere samarbejde omkring arkivering af webben.

Initiativet førte året efter til dannelsen af International Internet Preservation Consortium (IIPC). Konsortiet ledes af det franske nationalbibliotek, Bibliothèque nationale de France, og dets øvrige medlemmer er nationalbibliotekerne i hele Norden, Australien, Canada, Italien, Storbritannien og USA samt IA.[21] Det er konsortiets målsætning at identificere, udvikle og fremme implementeringen af løsninger til at etablere webarkiver samt at skabe en international lobby for initiativer, som fremmer dette. Konsortiet har nedsat seks arbejdsgrupper, som blandt andet skal udarbejde politikker, standarder og værktøjer. Arbejdet har i skrivende stund været i gang et år, og man begynder at nærme sig konsensus om opgavens indhold. Danmark har i det første år været mest aktiv i to af arbejdsgrupperne: framework, hvor vi bidrager til udarbejdelsen af et teknisk arkivformat og access, som Det Kongelige Bibliotek har haft ansvaret for, og hvor fokus primært har været på udarbejdelse af brugsscenarier og afklaring af, hvilke opgaver fremtidige adgangsværktøjer skal kunne løse.

Webarkivering – en grænsedisciplin?

Arkivering af webben og andre dele af internettet ligger biblioteksfagligt i grænseområdet mellem traditionel samlingsopbygning og it. Der er tale om et samlingsobjekt, der er skabt gennem udbredelsen af et elektronisk publiceringsmedie, men som også kræver væsentlige it-mæssige tiltag, hvis man vil kunne indsamle, lagre, bevare og give adgang.

Placeringen mellem det biblioteksfaglige og det rent tekniske ses bl.a. i den måde, hvorpå webarkivering som disciplin er blevet indordnet i faglige konferencer gennem de seneste 5-6 år. Mens problemstillingerne ved webarkivering endnu ikke har været et særskilt tema på World Wide Web Consortium's (W3C) store årlige internationale konferencer, ser det lidt anderledes ud på to andre årlige konferencer, IFLA og ECDL.

Johan Mannerheim fra Kungliga Biblioteket har været en drivende kraft i forbindelse med introduktionen af webarkivering i internationale biblioteks-sammenhænge. Ved IFLA-konferencen i 1997 lykkedes det ham at få indarbejdet webarkivering i 4-årsplanen for sektionen for Bevaring & Konservering. Fire år senere tog andre af IFLA's sektioner temaet op, *og indsamlingsstrategier, de lovgivningsmæssige aspekter og muligheder for at etablere adgang blev præsenteret.* Brewster Kahles engagerede indlæg om webarkiveringens pionerarbejde og hans efterfølgende tilstedeværelse ved et tilknyttet møde for en række af verdens største nationalbiblioteker blev det endelige skub til dannelsen af IIPC.

Siden 1997 har der årligt været afholdt en europæisk konference med fokus på forskellige aspekter af digitale biblioteker, ECDL. Det er først og fremmest universiteter med støtte fra nationalbiblioteker, der står bag konferencen, og emnerne er langt overvejende af datalogisk og informationsteknologisk art. Fra 2001 har en af konferencens fast tilknyttede heldagsworkshops været helliget webarkivering: IWAW, International Web Archiving Workshops med Julien Masanès, Bibliothèque nationale de France, som afgørende drivkraft.[22]

Webarkivering har i de seneste år også været genstand for selvstændige internationale konferencer i nationalbiblioteksregi, senest i Australien 2004 med et meget bredt program fra det samlingsorienterede over teknisk organisering til forretnings- og samarbejdsmodeller samt forskning.[23]

Forskellen i tilgang fra de to forskellige fagområder ses måske mest tydeligt i diskussionen om registrering af det arkiverede materiale. I den ene ende af spektret finder vi den traditionelle biblioteksmæssige tilgang med udvælgelse af 'kvalitetsressourcer', som beskrives af bibliotekarer, og som stilles til rådighed gennem strukturerede søgninger, således som det f.eks. kommer til udtryk i det australske Pandora-projekt: Her indsamles kun fra et site, hvis man kan træffe aftaler med siteejerne om, at materialet kan gøres tilgængeligt fra bibliotekets arkiv nu og for eftertiden, og der lægges tilsvarende en del bibliotekarressourcer i registreringsarbejdet.

I den anden ende af spektret findes IA's tværsnitshøstning af webmateriale fra hele verden, indsamlet og arkiveret uden hensyntagen til nationale

lovgivninger, og hvor alt det indsamlede udelukkende stilles til rådighed via maskingenererede indekser af forskellig art.

At tværsnithøstning ikke i sig selv udelukker overvejelser om muligheder for etablering af mere traditionelt strukturerede adgangsveje, er det svenske SVESØK-projekt et eksempel på. Da projektet startede i 1998, anvendte man det materiale, der blev indsamlet via Kulturarw[3]-projektet til at generere to indekser: et maskingenereret fritekstindeks til alle svenske websider på det aktive net og en linksamling med et udvalg af 'kvalitetsressourcer' på det svenske net, organiseret, katalogiseret og kvalitetsvurderet ved hjælp af 15 bibliotekarer fra hele Sverige. I 2000 ophørte man imidlertid med den manuelle katalogisering af økonomiske årsager. Fra 2004 fjernedes både det manuelt baserede indeks og mulighederne for at indrapportere Dublin Core Metadata, således at kun det maskingenererede indeks nu stilles til rådighed.

At den viden og knowhow, der findes på it-området, er afgørende for fremdriften inden for dette felt, ses også klart af de alliancer, jeg har nævnt ovenfor: LC har siden 2000 anvendt IA som teknisk samarbejdsparter, og BnF har i 2004 indgået en 2-årig samarbejdsaftale med IA om gennemførelse af forsøg på indsamlingsområdet. IIPC kan ligeledes ses som et forum, hvor disse to forskellige tilgange til opgaven mødes, synspunkter brydes, og man forsøger at anvende det bedste fra hver af de to meget forskellige tilgange i forbindelse med løsning af den fælles opgave.

Webarkivering er en ny disciplin først og fremmest i vores nationalbiblioteker – en disciplin, der ofte fremhæves for sine mange udfordringer. Den gode nyhed er, at der efterhånden er mange, der samarbejder om at løse opgaven. Danmark er i dag en aktiv del af denne udvikling, og så længe vi vedbliver med at være dette, har vi også mulighed for at sikre, at de løsninger, vi selv gennemfører, kvalitetsmæssigt og funktionelt er på højde med internationale standarder, når det gælder indsamling, bevaring og adgang og ikke mindst, at vi også kan sætte vores fingeraftryk på disse standarder.

Noter

1 B.E. Brewington og G. Gybenko: „How dynamic is the Web?" I: *[Proceedings of the] Ninth International World Wide Web Conference*. Amsterdam May 15-19, 2000. <www9.org/w9cdrom/264/264.html> (set oktober 2004); Alexandros Ntoulas, Junghoo Cho and Christopher Olston: *What's New on the Web? The Evolution of the Web from a Search Engine Perspective*. 2004, <www.2004.org/proceedings/docs/1p1.pdf> (set oktober 2004); J. Cho og H. Garcia-Molina: *The evolution of the Web and implications*

for an incremental crawler. 2000, <rose.cs.ucla.edu/~cho/papers/cho-evol.pdf> (set oktober 2004).

2 Alexandros Ntoulas, Junghoo Cho and Christopher Olston: *What's New on the Web? The Evolution of the Web from a Search Engine Perspective*. 2004, <www.2004.org/proceedings/docs/1p1.pdf> (oktober 2004).

3 Birte Christensen-Dalsgaard: „Web Archive Activities in Denmark", *RLG DigiNews* June, 15, 2004 <www.rlg.org/en/page.php?Page_ID=17661#article0>. I samme artikel gives en grafisk fremstilling af, hvorledes ændringshastighederne på nettet og variansen i interaktivitet påvirker valget af indsamlingsmetode.

4 I Danmark er der netop fremsat lovforslag om ny pligtafleveringslov, der skal sikre mulighed for gennemførelse af tre forskellige indsamlingsstrategier parallelt, <www.kum.dk/graphics/kum/billeder/Pressemeddelser/Pligtaflevering_041104/L_77-forslag_til_lov_om_pligtaflevering.pdf>, og det samme ses i New Zealand: <www.nla.gov.au/padi/topics/92.html#NZ>.

5 Birte Christensen-Dalsgaard: „Web Archive Activities in Denmark", *RLG DigiNews* June, 15, 2004, <www.rlg.org/en/page.php?Page_ID=17661#article0> (set oktober 2004).

6 *Udredning om bevaring af Kulturarven*. Kulturministeriet 2003. <kum.inforce.dk/graphics/kum/downloads/publikationer/bevaring_af_kulturarven.pdf>. (set oktober 2004).

7 Grethe Jacobsen*:* „Pligtaflevering 1850-1997". I: *Den trykte Kulturarv. Pligtaflevering gennem 300 år*. Redigeret af Henrik Horstbøll og John T. Lauridsen. 1998, s. 103-140. Lovtekst af 10. juni 1997 og betænkning på <www.pligtaflevering.dk>.

8 *Statistisk Årbog/Danmarks Statistik*, 2004, s. 336-338. *NDS*, nr. 81 25/2 2004, *Danmarks Statistik, Tabeller til Informationssamfundets Danmark, Itstatus* 2004, <www.dst.dk/Statistik/ags/IT/Informationssamfundet/Tabeller2004.aspx?>.

9 <www.dk-hostmaster.dk/dkhostcms/bs?pageid=101&action=cmsview&language=da>, <ftp.ripe.net/ripe/hostcount/History/RIPE-Hostcount.90-Oct-10>, <www.dkuug.dk/content/view/12/29/>

10 Dansk Bibliotekscenter: *Domæne.dk og nationalbibliografi? Registrering af netpublikationer, estimering af mængde,* <www.dbc.dk/nationalbibliografi/domaenedk.htm>, 2002 og rapporter fra Det Kongelige Bibliotek og Statsbibliotekets internetbevaringsprojekt.2004, <www.netarkivet.dk> (set oktober 2004).

11 <www.deflink.dk/upload/doc_filer/doc_alle/846_Trykt%20proceeding.pdf> (set oktober 2004).

12 Pilotprojektets slutrapport: <www.netarkivet.dk/rap/index-da.htm> (set oktober 2004).

13 Kulturarw^3 projektet offentliggør sin statistik på dets website: <www.kb.se/kw3/Statistik.htm> (set november 2004).

14 NEDLIB projekt: <www.kb.nl/coop/nedlib/> (set oktober 2004).

15 Projektbeskrivelse og det udviklede software kan findes via projekt websitet: <nwa.nb.no>.

16 <crawler.archive.org/team-list.html> (set november 2004).

17 Et godt overblik over projekter fås på NLA's website: <www.nla.gov.au/padi/topics/92.html>.

18 UK Government Web Archive: <www.nationalarchives.gov.uk/preservation/webarchive/> og UK Web Archiving Consortium <www.webarchive.org.uk/> (set oktober 2004).
19 BnF og BL's beskrivelse og EU-udbud af deres fælles Smart Archiving Crawler project <www.bl.uk/procurement/eucontracts/pqqweb11.doc> (set november 2004).
20 Gå ind til alle samlingerne via MINERVA: Mapping the Internet Electronic Resources Virtual Archive: <www.loc.gov/minerva/> (set november 2004).
21 International Internet Preservation Consortium (IIPC): <www.netpreserve.org> (set november 2004).
22 Samtlige bidrag fra IWAW workshops kan findes via <bibnum.bnf.fr/ecdl/index.html> (set november 2004).
23 National Library of Australia's konference november 2004: <www.nla.gov.au/webarchiving/> (set oktober 2004).

Til verifikation af en lang række forhold, hvor der henvises til et årstal og f.eks. et produkt eller et projekt, er næsten overalt anvendt IA's webarkiv: <web.archive.org>.