Erfaringer med høstning af det danske net 2005 - 2008

Af Grethe Jacobsen

I foråret 2006[1] kunne DF Revys abonnenter læse om de første erfaringer med nethøstning. Der er nu gået 3 år siden den lov, der gjorde nethøstningen lovlig, trådte i kraft og Netarkivet.dk synes, det kunne være passende med en opdatering.

Grethe Jacobsen
Førstebibliotekar
Det Kongelige Bibliotek
gja@kb.dk

Loven

Vi arbejder under Lov nr. 1439 af 22. december 2004 om pligtaflevering af offentliggjort materiale, der trådte i kraft 1. juli 2005. Loven omfatter også værker i fysisk form, radio og TV samt film, og for disse materialer er adgangen reguleret af ophavsretsloven. Vi har oprettet et særligt websted om pligtaflevering: www.pligtaflevering.dk med lovteksterne og andre relevante oplysninger, især rettet mod de borgere, der har en forpligtelse i henhold til loven.

Loven skulle revideres i folketingssamlingen 2007/08, specielt med henblik på at ændre de restriktive forhold vedrørende adgang til netarkivet, som er begrundet i, at arkivets materiale kan indeholde personfølsomme data. Da vi imidlertid endnu ikke har nogle tekniske løsninger, der kunne identificere sådanne data, er revisionen af loven udsat til folketingssamlingen 2010/11 (Lov nr. 89 af 20/02/2008 om ændring af lov om pligtaflevering af offentliggjort materiale).

Organisation

Netarkivet.dk, der tager sig af nethøstning, er en virtuel institution dannet af Det Kongelige Bibliotek og Statsbiblioteket og med en daglig leder, der hidtil har været ansat på halv tid, men fra 1. september arbejder 30 timer om ugen. Lederen refererer til styregruppen, der består af tre afdelingsledere fra Statsbiblioteket og tre fra Det Kongelige Bibliotek med ekspertise i samlingsopbygning, it-udvikling og digital bevaring. Styregruppen mødes 3-4 gange årligt og kommunikerer desuden via en intern wiki.

Den daglige leder har indtil 1. juli 2008 været Bjarne Andersen, der med sikker hånd har styret Netarkivet.dk gennem de første tre år. Bjarne fortsætter sin tætte forbindelse til Netarkivet.dk, idet han bliver medlem af netarkivets styregruppe fra 1. oktober 2008. Ny daglig leder er Claus Lomborg, der er ansat ved Statsbiblioteket. Foruden Claus beskæftiger Netarkivet.dk i alt 20 medarbejdere (bibliotekarer, biblioteksassistenter, en medieforsker, IT-ingeniører og dataloger) fordelt på 5 ÅV, der regelmæssigt mødes enten personligt eller via videomøder.

Som hjælp til indsamlingen har Kulturministeriet udnævnt en redaktionsgruppe, bestående af fem repræsentanter fra IT-branchen, medie- og forlagsverden. Medlemmerne er beskikket frem til juni 2009, hvorefter Kulturministeriet vil udpege en ny redaktion.

Teknik

Netarkivet råder over 9 servere med Linux som system og 55 pc'er med Windows XP. Programmerne, der anvendes, er dels høstningsprogrammet Heritrix, udviklet af et konsortium bestående af Internet Archive og de nordiske nationalbiblioteker, dels webarkiveringsprogrammet NetArchiveSuite, udviklet af Det Kongelige Bibliotek og Statsbiblioteket, og som nu også anvendes af det skotske og det østrigske nationalbibliotek. Softwaren er frigivet under Open Source, hvilket betyder, at det er tilladt alle at bruge det – og bidrage til videreudviklingen.

Hvad har vi så samlet ind og hvordan?

I juli 2008 var der data på i alt 71 Terabytes (71.000 Gigabytes) i arkivet. De er indsamlet ved hjælp af de tre strategier for indsamling, der blev lagt som resultat af de pilotprojekter, de to biblioteker afviklede forud for loven af 2004. De tre strategier er:

- Tværsnitshøstning (der fylder i alt 56 Tb)
- Selektiv høstning (9 Tb)
- Begivenhedshøstning (6 Tb)

Rent praktisk er det organiseret sådan, at Det Kongelige Bibliotek har ansvaret for tværsnitshøstningen, Statsbiblioteket for den selektive høstning og de to biblioteker er fælles om begivenhedshøstningen.

Tværsnitshøstning af domænet <.dk>

Formålet med denne strategi er at indsamle et billede af det danske internet, både de sider, der findes på domænet <.dk> og de sider, der findes på andre domæner (<.com>, <.org>, <.nu> etc.), men som er rettet mod et dansk publikum.

Websider, der findes på <.dk> høstes efter en liste over danske domæner, som vi har fået fra Hostmaster DK, der bestyrer DK-domænet. Da vi begyndte nethøstningen i juli 2005 var der 607.000 domæne navne, hvoraf de 479.000 var aktive. I juli 2008 var over 900.000 domænenavne, hvoraf de 660.000 var aktive. Nogle domænenavne er aliasser, dvs. henvisninger til et websted under et andet navn, fx www.detkongeligebibliotek.dk, der er et alias for www.kb.dk. Andre er navne på websteder, der er beregnet for en lukket kreds (fx Netarkivets egen wiki, der kun er for medarbejdere og ledere knyttet til Netarkivet.dk) og som derfor ikke er omfattet af pligtafleveringsloven. Endelig har en del websider kun adgang via log-in. De er ikke blevet høstet endnu under tværsnitshøstningen. En stikprøveundersøgelse udført i juli 2008 har vist, at ca. 16 % af disse sider er omfattet af pligtafleveringsloven. Netarkivet skal nu undersøge nærmere, om disse sider kan identificeres af høsteprogrammet, så de kan indsamles, uden at vi får sider, vi ikke skal have.

[1] Andersen, B. "DK-domænet i ord og tal". I: *DF Revy*, nr. 1, 2006, s. 4-7

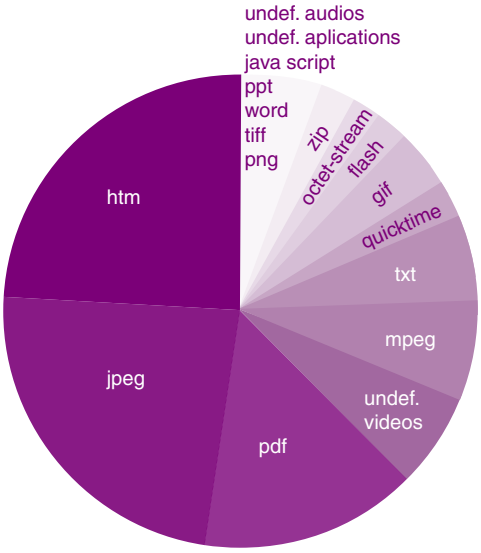
Webstederne bliver i første runde høstet op til 10 Mb, hvorved over 90 % af alle danske websteder bliver indsamlet. Dem, der er større end 10 Mb bliver så høstet op til 1 Gb. Vi har erfaret, at det går hurtigere og med færre problemer at høste i to omgange end at tage alt op til 1 Gb i første omgang. De sidste 2.400 websteder, der er større end 1 Gb, bliver manuelt gennemgået af medarbejdere i Pligtafleveringsafdelingen på Det Kongelige Bibliotek, der kontrollerer, om webstedet har indhold, der fylder mere end 1 Gb eller, om der er tale om en såkaldt "crawlertrap", dvs. en side med lænker, der sender maskinen videre til lignende sider uden indhold i en uendelighed (fx en kalender). Tilsvarende er det heller ikke heldigt, når vores høsteprogram putter tusindvis af varer i indkøbskurven i en webshop. Når vi opdager crawlertraps, sørger vi for, at de fremover bliver blokeret, så disse sider ikke indsamles ved næste tværsnitshøstning.

Det var planen at lave fire tværsnitshøstninger om året. Det har i praksis vist sig umuligt, så vi har kun nået seks tværsnitshøstninger i alt, hvilket er halvdelen af, hvad vi havde planlagt. Dette skyldes dels mangel på ressourcer (ÅV og maskiner), dels problemer med programmel. Ved at sætte flere maskiner ind, håber vi på at kunne nå tre høstninger i år og være oppe på fire næste år.

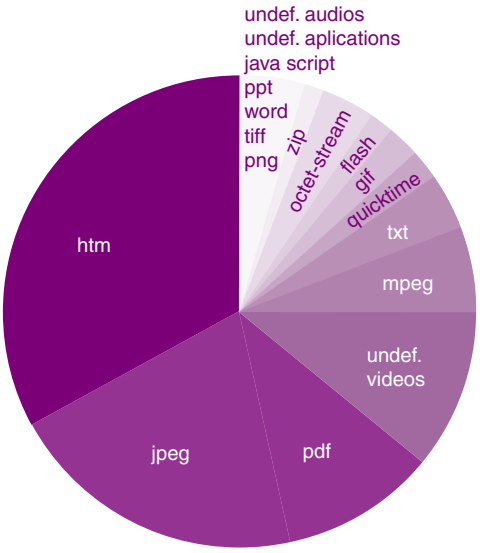
Tværsnitshøstning – sider uden for domæne <.dk>

Loven dækker også sider uden for <.dk> og dem må vi selv finde frem. Det kan vi gøre automatisk ved at undersøge alle de sider på andre domæner, hvortil der er et link fra en <.dk> side ved hjælp af et geo-ip program. Programmet (en IP-lokalisator) kan med rimelig sikkerhed afgøre, i hvilket land et givent navn hører til. Hvis serveren fysisk befinder sig i Danmark, formoder vi, at indholdet vil være beregnet for det danske marked og indlemmer disse url'er i vores høster. I 2006 lavede vi desuden en manuel søgning i Google på danske stednavne og fagtermer, der fandtes uden for <.dk> domænet og fandt mange sider, der ligeledes blev indlemmet. Endelig har vi på vores websted, www.netarkivet.dk, en side, hvor man kan angive websteder uden for <.dk> domænet, som vi så kontrollerer og hvis indholdet er beregnet for det danske marked, bliver også disse url'er indføjet i høsteren.

Filtyper indsamlet efterår 2005[2]



Filtyper indsamlet forår 2008



[2] Statistikken dækker høstede filer og diagrammet er behæftet med visse usikkerheder

Disse tre metoder har givet mere end 42.000 websteder på andre domæner, som Netarkivet for første gang høstede i foråret 2007. Disse nye domæner er større end DK-domænerne og fylder ca. 46 Mb i gennemsnit (mod ca. 16 Mb i gennemsnit for et domæne på <.dk>). Det havde vi forventet, da mange af disse websteder tilhører firmaer, der har valgt at lægge deres websted på <.com>.

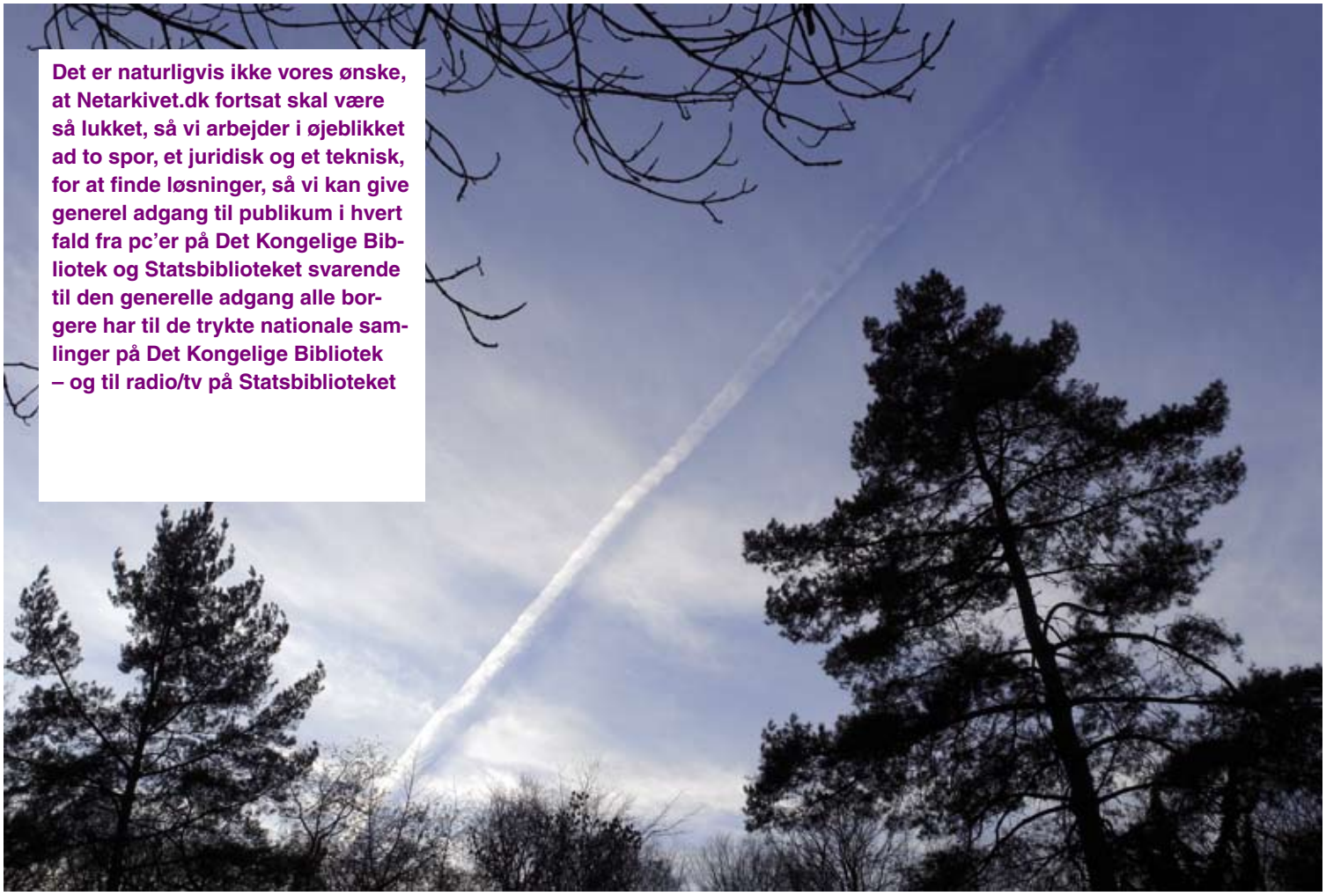
Indholdet af det materiale, vi har indsamlet, er selvfølgelig meget varieret, men vi kan konstatere, at de tendenser, vi så ved den første indsamling, har holdt: de fleste websteder er stadig små (90 % er under 10 Mb), mens de store er blevet større. Internettet er blevet opbevaringssted for familien Danmarks fotoalbum, inklusive videoer. Vi kan også se en stigende tendens til, at flere og flere netsteder indlejrer materiale fra andre netsteder (fx videomateriale fra youtube.com). Det er et tydeligt resultat af Web 2.0 tendensen, der slår igennem. Hvad angår filtyper, så er det stadig htm og jpeg filer, der dominerer.

Selektiv høstning

Formålet med den selektive høstning er at indsamle sider, der ændrer sig hurtigere end tværsnitshøstningen kan nå at fange (selv når vi kommer op på fire høstninger om året). Vi skal ifølge bemærkningerne til loven indsamle:

1. Netsteder, der fungerer som nyhedsmedie for den danske offentlighed. Hertil hører især de landsdækkende avisers og tv-stationers netsteder, de større regionale avisers netsteder, Folketingets netsted samt et mindre udvalg af nyhedsmedier
2. Netsteder, der dækker et repræsentativt udsnit af typiske netsteder, dvs. meget besøgte netsteder (fx portaler af typen netdoktor), og/eller netsteder, der repræsenterer typiske anvendelser (fx virksomheders netsteder, byportaler og kommunale portaler, netsteder, der anvendes som lokalt medie)
3. Netsteder, der er særligt karakteristiske, eks perimentelle eller unikke (fx åbne debatmiljøer, netfællesskaber, personlige sider, netkunststeder) i forhold til indhold og udtryksformer.

Da vi begyndte nethøstningen i juli 2005 var der 607.000 domæne navne, hvoraf de 479.000 var aktive. I juli 2008 var over 900.000 domænenavne, hvoraf de 660.000 var aktive.



Det er naturligvis ikke vores ønske, at Netarkivet.dk fortsat skal være så lukket, så vi arbejder i øjeblikket ad to spor, et juridisk og et teknisk, for at finde løsninger, så vi kan give generel adgang til publikum i hvert fald fra pc'er på Det Kongelige Bibliotek og Statsbiblioteket svarende til den generelle adgang alle borgere har til de trykte nationale samlinger på Det Kongelige Bibliotek – og til radio/tv på Statsbiblioteket

I bemærkningerne er også angivet, at der skal høstes ca. 80 websteder med denne strategi, og Statsbiblioteket gik straks i gang med at finde websteder, der faldt inden for disse kategorier. Kandidater til kategori 1 (nyhedsmedier) er oplagte, og danske avisers websteder blev da også hurtigt indlagt i systemet, efter vi begyndte nethøstningen. I øjeblikket indsamler Statsbiblioteket mellem 30 og 35 nyhedswebsteder. Kandidater for kategori 2 og 3 krævede flere overvejelser og især undersøgelser. Statsbibliotekets medarbejdere har vurderet flere hundrede websteder, førend de fandt frem til de i alt 82 websteder (inklusive nyhedswebsteder), der indsamles i dag. Hvert websted skal løbende vurderes, både om det fortsat skal indsamles selektivt, hvor ofte det skal indsamles og hvor mange sider af websteder, der skal indsamles selektivt, så det er en strategi, der kræver en del manuelt arbejde. Udvælgelsen af de 82 websteder er løbende blevet drøftet i møderne med Redaktionsgruppen, der også har bidraget med forslag. Derudover er webstederne fundet via aviserne og fra lister over "mest besøgte steder" samt omtale i internetforskningen. Det skal bemærkes, at blandt de mange websteder, der har været på tale, er der adskillige, som opfylder kriterierne ovenfor, men som også har en arkivfunktion knyttet til webstedet, hvorfor informationerne fanges af tværsnitshøstningen.

På Netarkivets websted http://netarchive.dk/selektiveHostninger.html kan man se en liste over de websteder, der pt. høstes hyppigt (nogle op til flere gange dagligt).

Begivenhed:

Den tredje strategi er begivenhedshøstning. Hvad er så en begivenhed i Netarkivets øjne? Det er i Redaktionskomiteen blevet defineret som noget, der:

- skaber debat i befolkningen og menes at have en betydning for Danmarks historie eller udvikling
- udløser nye websteder
- også behandles i stort omfang på eksisterende websteder

En "begivenhedshøstning" kan være en kortvarig ekstra indsamling fra et enkelt eller nogle få domæner, eller det kan være en længerevarende indsamling, hvor høstningsfrekvensen nedtrappes, efterhånden som aktiviteten på nye websteder aftager.

Nogle begivenheder, fx et valg, er forudsigelige, men de fleste begivenheder vil ikke kunne forudsiges. Det er først, når en sådan begivenhed er i gang, dvs. når der er opstået nye websteder, når begivenheden har skabt debat, og når den behandles på eksisterende websteder, at den bliver en begivenhed i Netarkivets øjne. For at få fanget en sådan begivenhed, så tidligt som muligt, kan det være nødvendigt at sætte høstninger i gang af en eller flere websteder, der ser ud til at blive en begivenhed. Denne høstning bliver så stoppet, hvis der alligevel ikke sker noget, der ikke bliver indsamlet af de to andre høstningsstrategier.

Den første begivenhed, vi høstede, var en af de planlagte, nemlig lokalvalgene i november 2005, som jo var af speciel betydning, idet der var valg til pladser i bestyrelsen af de nye kommuner og regioner, der opstod som følge af strukturreformen. I modsætning til valget i 2001, der blev indsamlet som et pilotprojekt, var der denne gang sket en voldsom vækst i antallet af websteder for de enkelte kandidater. Mere end 1.000 kandidater havde deres eget websted, og mange af disse sider forsvandt igen efter valget.

Den næste begivenhed var uforudset, nemlig krisen i forbindelse med Muhammed-tegningerne i 2006, hvor vi høstede danske og internationale websteder fra februar til juni 2006. Den tredje høstning, i december 2006, var afledt af strukturreformens ikrafttræden i januar 2007, idet vi høstede dels de gamle kommuners og amters websteder, dels websteder med oplysninger om overgang til ny kommune, da vi – korrekt viste det sig – forudså, at de hurtigt ville forsvinde.

Derudover har vi små begivenheder, der løbende er indsamlet, fx rydningen af ungdoms-

Den første begivenhed, vi høstede, var en af de planlagte, nemlig lokalvalgene i november 2005, som jo var af speciel betydning, idet der var valg til pladser i bestyrelsen af de nye kommuner og regioner, der opstod som følge af strukturreformen

huset i København i marts 2007 og dannelse af partiet Ny Alliance i maj 2007. En stor begivenhed i dette år var selvfølgelig valget i november 2007, hvor høstmaskiner var i gang tre timer efter, at valget blev udskrevet. Da valgrygterne begyndte tidligere på året, begyndte Det Kongelige Bibliotek og Statsbiblioteket at indsamle og opdatere listen fra forrige valg over websteder, der skulle indsamles, såsom partiernes hjemmesider og politiske debatfora. Derefter fik medarbejderne travlt med at finde frem til alle kandidaterne og undersøge, om de havde deres egne websteder, hvad de fleste havde. Nyt i valgkampen var brug af videoklip på webstederne. Her fik vi hjælp af Internet Archive til at indsamle disse klip. Også blogspots blev indsamlet, og det lykkedes os desuden at fange lidt af valgaktiviteterne på Facebook, inden de lukkede for adgangen. Det var tydeligt, at meget af valgkampen nu havde skiftet til nettet, hvor den foregik i multimedieform og med interaktion med vælgerne. Alt i alt endte valget med at fylde 2,2 Tb (2.200 Gb). Valget i 2005 fyldte til sammenligning 293 Gigabytes. En liste over begivenheder, der er høstet, kan også findes på vores websted: http://netarchive.dk/indsamlingsDoku.html.

Det er vigtigt at understrege, at disse tre strategier er indsamlingsstrategier – ikke samlingsopbygningsstrategier. Alt høstet materiale indgår i ét netarkiv, så fremtidige brugere kan surfe rundt på det danske net, som brugeren i dag gør det, men vi laver en liste over det, vi indsamler, og hvad der har været af problemer, (http://netarchive.dk/heritrixVersioner.html), så der er dokumentation for, hvordan Netarkivet er blevet opbygget, og hvad der kan forventes at være i arkivet for fremtidens brugere.

Udfordringer de næste par år

Danmark er sammen med de øvrige nordiske lande, der nu alle indsamler den nationale del af internettet, i front med denne form for pligtaflevering, men det betyder jo ikke, at vi kan hvile på laurbærrene. Vi står over for adskillige udfordringer de næste par år.

En af dem er, at den del af internettet, vi kan indsamle ifølge pligtafleveringsloven, stiger hastigere end vores ressourcer, det gælder både indsamling, udvikling og bevaring. Det er jo en gammel problemstilling, som pligtafleveringsbibliotekerne er inderligt bekendte med, og det forsøger vi løbende at løse.

En anden udfordring er én, vi har haft siden loven blev vedtaget, nemlig den ovenfor nævnte restriktive adgang til Netarkivet. Adgangen er pt. begrænset til forskere og ph.d. studerende og for dem er der kun adgang i forbindelse med forskning eller statistiske undersøgelser. Det skyldes, at arkivet kan indeholde følsomme persondata, og det bevirker, at arkivet falder ind under persondataloven, ifølge Datatilsynets tolkning af EU direktivet om personoplysninger (Europa-Parlamentet og Rådets direktiv 95/46/EF af 24. oktober 1995 om beskyttelse af fysiske personer i forbindelse med behandling af personoplysninger og om fri udveksling af sådanne oplysninger), uanset at de indsamlede data har været offentliggjort og for en dels vedkommende stadig kan ses på nettet.

Det er naturligvis ikke vores ønske, at Netarkivet.dk fortsat skal være så lukket, så vi arbejder i øjeblikket ad to spor, et juridisk og et teknisk, for at finde løsninger, så vi kan give generel adgang til publikum i hvert fald fra pc'er på Det Kongelige Bibliotek og Statsbiblioteket svarende til den generelle adgang alle borgere har til de trykte nationale samlinger på Det Kongelige Bibliotek – og til radio/tv på Statsbiblioteket.

Det juridiske spor går ud på at analysere lovgivningen for at se, hvordan balancen er mellem de to modstridende interesser: beskyttelse af den enkelte borger og offentlighedens behov for at få oplysninger (fx fra valgkampagner) til at beskytte demokratiet og også se på, hvordan andre lande, der indsamler internettet, har tolket EU-direktivet. En jurist fra Aalborg Universitet vil i løbet af efteråret lave et responsum herom, som vi så kan arbejde videre ud fra.

Det tekniske spor går ud på at udvikle programmer til at identificere sider med følsomme persondata, som så kan filtres fra, når der gives generel adgang til arkivet. Det er ikke et spor, vi pt. nærer de største forventninger til, da hidtidige forsøg har vist, at det er endog meget svært at identificere følsomme persondata, fordi konteksten skal tages med. Følsomme persondata er oplysninger om en identificerbar persons racemæssige eller etniske baggrund, politiske, religiøse eller filosofiske overbevisning, fagforeningsmæssige tilhørsforhold samt oplysninger om vedkommendes helbredsmæssige og seksuelle forhold. Hvis neutrale data om en person kan sættes ind i en sammenhæng, der kan give oplysninger om ovenstående, bliver de følsomme. Vi vil fortsætte vores forsøg med at finde tekniske løsninger, når vi har fået det juridiske responsum.

En tredje udfordring, vi står overfor, er registreringen af det indsamlede. Skal vi pille enkelte værker ud og lave en manuel nationalbibliografisk registrering, eller skal vi videreudvikle værktøjer, så man kan søge på tværs i arkivet, når alle borgere i forhåbentlig ikke for fjern fremtid får adgang?

Der vil utvivlsomt dukke flere udfordringer op fremover. Det skal vi nok indvie DF Revys læsere i.

Hvert websted skal løbende vurderes, både om det fortsat skal indsamles selektivt, hvor ofte det skal indsamles og hvor mange sider af websteder, der skal indsamles selektivt, så det er en strategi, der kræver en del manuelt arbejde.