DK-domænet i ord og tal

Af Bjarne Andersen

Den 1. juli 2005 trådte en ny revision af pligtafleveringsloven i kraft i Danmark. Det betød at nationalbibliotekerne i Danmark – Det Kongelige Bibliotek og Statsbiblioteket – fik pligt og lovhjemmel til at indsamle og bevare den danske del af internettet.

Bjarne Andersen
Driftsleder
netarkivet.dk
bja@netarkivet.dk
netarkivet.dk

Indsamlingen begyndte i juli 2005 og varetages af et virtuelt center: **netarkivet.dk**, der drives i et tæt samarbejde mellem de to biblioteker.

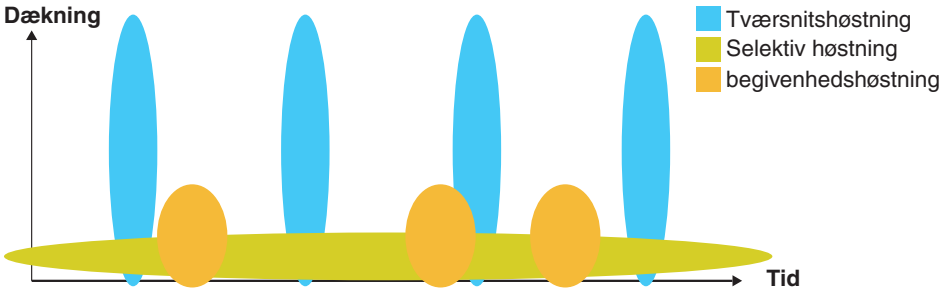
Indsamlingen gennemføres med traditionelt webcrawler-software der foretager automatiseret indhentning – såkaldt høstning - efter en 3-strenget strategi:

1. Tværsnitshøstning af alle relevante domæner 4 gange årligt
2. Selektiv høstning af ca. 80 domæner med hyppigere frekvens (fx dagligt)
3. Begivenhedshøstning af 2-3 begivenheder årligt

Den samlede strategi kan illustreres som vist på figur 1.

Tværsnitshøstning giver et bredt billede af den danske del af internettet. Alle kendte domæner hentes med få ristrektioner komplet. De **selektive høstninger** dækker udvalgte netsteder (fx nyhedsmedier) med højere frekvens – potentielt ned til 1 gang i timen og giver et kontinuerligt billede af et lille antal særligt betydningsfulde og dynamiske netsteder, mens **begivenhedshøstningerne** kombinerer de 2 andre strategier og indsamler et højere antal – i størrelsesordenen 2-4000 – med en hyppigere frekvens fx dagligt. Netarkivet har i oktober/november gennemført den første begivenhedshøstning omkring kommunalvalget 2005.

Denne artikel omhandler netarkivets første erfaringer med tværsnitshøstning som blev gennemført i juli-oktober 2005. Denne tværsnitshøstning dækkede kun DK-domæner, hvorfor erfaringerne fra den giver et meget godt billede af netop DK-domænet. Det er så vidt vides den første så dækkende karakteristik af det danske domæne.



Figur 1: *Høstning - efter en 3-strenget strategi*

Konceptuelle præmisser

Tværsnitshøstningen er gennemført under en række præmisser:

1. Høsterne blev igangsat på en komplet liste fra DK-hostmaster fra juni 2005. Listen rummede ca. 607.000 domænenavne
2. Høsterne kunne på det tidspunkt ikke håndtere domænenavne, der rummede danske karakterer (æ, ø og å), hvorfor det reelle antal domæner der forsøgtes høstet var ca. 579.000
3. Høsterne respekterede ikke robots.txt standarden (læs længere nede hvorfor)
4. Indsamlingen blev gennemført med en grænse på maksimalt 5000 objekter pr. domæne for at undgå blandt andet overlast på de besøgte web-servere og for at undgå de såkaldte crawlertraps (læs mere om dem senere)

Tekniske forhold

Teknisk blev høstningen gennemført underfølgende forhold:

1. På 2 maskiner med hver 2 CPU'er og 4 GB RAM. Hver maskine kørte 2 instanser af webcrawlersoftwaren
2. Med open source høsteren heritrix[1]
3. Gennem nationalbibliotekernes 100 Mbit netforbindelse på forskningsnettet.
4. Med en båndbreddebegrænsning på maksimalt 3 Mb/sek pr. høster-maskine og maksimumgrænse på 500 Kb/sek pr. domæne.
5. Tog ca. 3 uger i effektiv høstnings-tid fordelt over perioden juli-oktober 2005 – altså et effektivt båndbreddeforbrug på 2.9 Mb/s hvilket svarer til ca. 25 Mbit/s
6. Hentede 138.796.750 objekter som fylder ca. 5.3 Tbytes (5.300 Gbytes)

Begrænsninger

Da dette var den første totale indsamling af det danske domæne og da netarkivet ønskede at gå varsomt frem, blev det besluttet at sætte en maksimal grænse på antallet af objekter pr. domæne. Grænsen blev sat til 5000 objekter pr. domæne ud fra en tese om, at dette ville sikre hovedparten af de danske netsteder komplet og stadig give et dækkende billede af de meget store sites.

Grænsen blev primært sat af 2 grunde. For det første ønskede vi ikke at overbelaste de danske web-servere mere end højst nødvendigt. Vi var klar over denne indsamling for mange netsteder ville generere en langt højere trafik

end normalt og ønskede naturligvis til stadighed at holde os gode venner med ejerne af de berørte netsteder – de er trods alt vores vigtigste samarbejdspartnere, om end de i princippet ikke selv skal foretage sig noget.

Indsamlingen med webcrawler-software fungerer konceptuelt på den måde at softwaren fodres med en række startsteder (her den komplette liste af DK-domænernes forsider). Disse sider hentes, softwaren finder links til nye sider og sådan kører processen, ind til der ikke er flere sider tilbage på de domæner, der blev startet på, eller til grænsen på de 5000 objekter per domæne er nået.

For det andet fungerer grænsen også som et praktisk værn mod de såkaldte crawlertraps. Crawlertraps er steder på nettet hvor web-crawleren som navnet antyder 'fanges' i den virtuelle verden, forstået på den måde, at web-crawleren kommer til at hente i princippet uendelige mange sider, hvis den ikke bremses på den ene eller anden måde. En typisk og ofte forekommende crawlertrap er en kalenderapplikation, hvor man på et netsted kan klikke sig rundt i kalenderen – fx med links til næste dag, næste måned og næste år. Her kan webcrawleren bliver ved med at finde nye links og hente nye sider helt uden reelt indhold, da de færreste kalendere rummer indgange mange år ud i fremtiden (eller fortiden).

Det at anvende en maksimal grænse giver naturligvis en række usikkerheder på den efterfølgende statistik. Især påvirkes statistikkerne over hvor store de danske netsteder er, forstået på den måde, at de netsteder, der nåede grænsen på de 5000 objekter, jo reelt er større end 5000 – vi ved bare ikke hvor meget. Antallet af sites, der helt praktisk nåede grænsen, var dog ikke så stort, som den efterfølgende statistik vil vise, hvorfor usikkerheden reelt kun omfatter et mindre antal sites.

Forekomsten af crawlertraps trækker også statistikken over de danske netsteder størrelse lidt op, idet crawlertraps får sites til at se større ud end de reelt er. Vi har ikke nogen brugbar statistik over hyppigheden af crawlertraps, primært fordi de er næsten umulige at finde maskinelt.

Robots.txt

Indsamlingen af den danske del af internettet respekterer ikke de såkaldte robots.txt direktiver. Undersøgelser fra 2003-2004 viste, at rigtig mange af de særligt vigtige netsteder (fx nyhedsmedier, politiske partier) havde ret strenge robots.txt direktiver, der gjorde, at hvis de blev efterlevet, ville der ikke bliver arkiveret noget som helst. Robots.txt er derfor eksplicit nævnt i bemærkningerne til loven, ikke fordi robots.txt er en juridisk standard (snarere en slags gentlemen agreement), men netop fordi de parter, der blev inviteret i høringsfasen (professionelle aktører inden for internetbranchen), havde mulighed for at forholde sig kritisk til den tilgangsvinkel.

Statistikken for den første tværsnitshøstning viste, at godt 35.000 netsteder havde robots.txt direktiver og dermed havde mere eller mindre strenge regler for, hvad web-crawlere må hente. Netarkivet har ikke ressourcer til manuelt at kigge på indholdet af disse mange forekomster, hvorfor udgangspunktet med at ignorerer robots.txt standarden med arkivets formål in mente synes naturligt.

Netarkivet er naturligvis klar over, at robots.txt også er opfundet for at hindre web-crawlere i at lave forespørgsler på URL'er, der potentielt kunne skabe problemer på de besøgte netsteder (fx sender indlæg til debatfora, sender mail til webmaster og lignende). Vi står altid til rådighed til at finde en hurtig og effektiv løsning, såfremt det skulle vise sig af netarkivets web-crawlere, har opført sig uhensigtsmæssigt, og vi har da også haft et mindre antal henvendelser af den art. En oplagt løsning er, at overholde robots.txt på de sites, hvor det måtte skabe gener at lade være, og den model er da også blevet brugt i et par tilfælde. En anden mindre begrænsende løsning er, at netarkivet i sit system kan indbygge nogle regler for bestemte URL'er (eller URL-synktakser), der ikke må hentes.

Nogle få henvendelser har også peget på robots.txt direktivernes anvendelse til at undlade materiale, som producenten finder irrelevant (fx private fotos m.m.). Det er hverken netarkivets eller folks egen opgave at definere, hvad der i 2005 er relevant eller ej. Fremtidens forskere kunne måske netop afsløre interessante ting på baggrund af materiale, der måske i dag umiddelbart vurderes som uinteressant, men om 10 eller 50 år måske er vigtige kilder.

Da netarkivets formål er at bevare den danske del af internettet i al evighed (i princippet), skal vi ikke begrænses af hvad nogle i dag måtte finde relevant eller irrelevant, hvilket også er hovedårsagen til at tværsnitshøstningen overhovedet gennemføres.

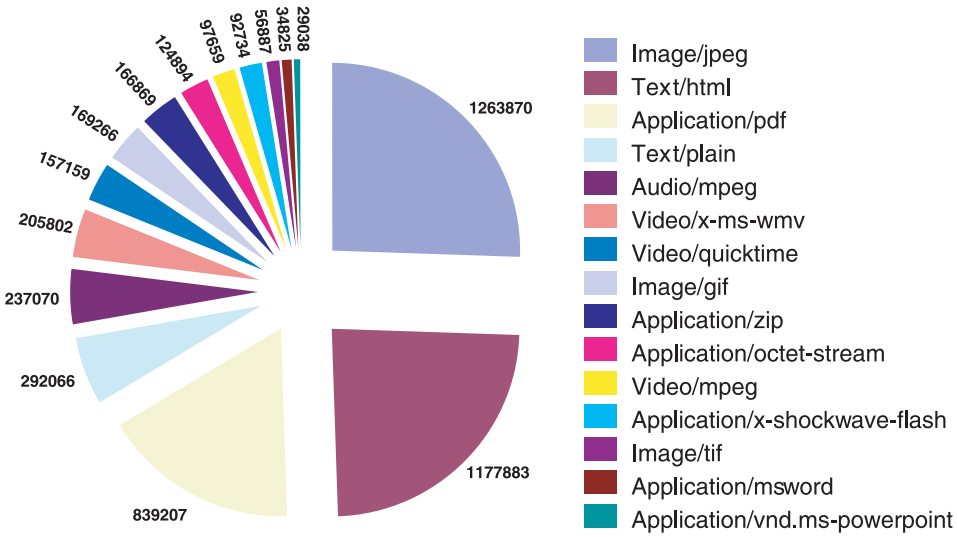
I andre lande (fx Australien) har man ind til for ganske nyligt kun indsamlet efter en selektiv strategi, men også her har man nu indset (og fået en lovgivning der matcher dette), at vi ikke i dag kan sige, hvad der er interessant i fremtiden, hvorfor det findes mest sikkert at prøve at bevare alting. Prisen på lagerplads falder stadig, hvorfor dette i praksis også i dag er en overkommelig opgave sammenholdt med for bare 5-10 år siden, hvor datamængder af denne størrelse skulle lægges på bånd for at holde internetarkiveringsprojekter på et rimeligt budget.

Filtyper

Indsamlingen af hele det danske domæne giver flere interessante statistikker. Den følgende statistik viser fordelingen af filtyper på det indsamlede. Den kan blandt andet være med til at sige noget om, hvilke formater der er de mest populære – om danskerne fx foretrækker Word eller PDF, når der skal publiceres dokumenter på internettet.

Figur 2 viser datamængden i Mbytes for de 15 filtyper der fylder mest i indsamlingen. Mest overraskende er det nok, at JPG-billeder ligger som topscorer. Vi tolker dette som et vidnesbyrd om, at digitalkameraer efterhånden er blevet ret almindelige, og at danskerne i stadig stigende grad lægger det private fotoalbum ud på den offentlige del af internettet. Stikprøver har vist, at vi fra en del netsteder har hentet private fotoalbums med 1000-vis af billeder, flere af disse albums med billeder i endog meget høj opløsning.

En mindre overraskelse har det været, at PDF-filer ligger på 3.-pladsen med antal Mbytes. Det er et klart tegn på, at PDF-filer i 2005 er langt det mest foretrukne format til publicering af dokumenter (bortset fra dokumenter i HTML-format). Det bliver interessant at følge den statistik i de kommende år.



Figur 2: *Datamængden i Mbytes for de 15 filtyper der fylder mest i indsamlingen.*

Ikke overraskende fylder AV-materialet (audio/* & video/*) også meget i statistikken, fordi disse formater i sig selv fylder meget.

Sammenlagt dækker de 15 filtyper mere end 93 % af den samlede datamængde. Ud af en samlet liste på 613 unikke mimetypes, hvor af en pæn del ikke er officielt registrerede mimetypes, er det således en stor andel. Langtidsbevaringen af filer i netarkivet vil sandsynligvis ikke kunne holde liv i samtlige filtyper, men denne statistik viser, at vi kan bevare langt størstedelen af det samlede arkiv, hvis 'bare' vi kan bevare et mindre antal forskellige filtyper. Flere af filtyperne på top-15 listen rummer dog også endog meget store udfordringer, når det kommer til bevaring, men det er en problemstilling der slet ikke er plads til i denne artikel.

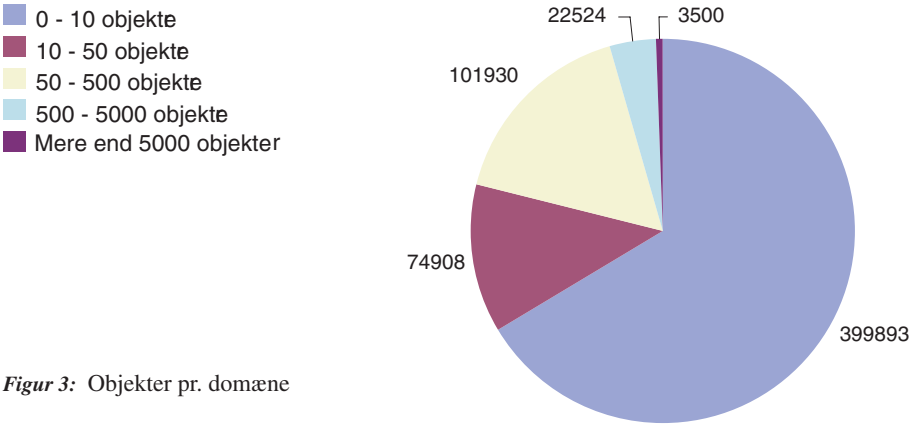
Statistikken over de 15 filtyper der fylder mest målt i antal Mbytes vises i tabel 1. Tabellen rummer også information om antallet af objekter, der er hentet af de pågældende typer, samt den gennemsnitlige filstørrelse for filtyperne.

Målt i antal er det ikke overraskende HTML-filer, der er langt de mest udbredte – over 50 %. Derefter kommer de 2 mest udbredte billedformater, JPG og GIF. Sammenlagt dækker de 3 filtyper over mere end 97 % af antallet af objekter, men altså kun 52 % af mængden målt i Mbytes.

Målt i antal er det også klart, at PDF er langt mere udbredt end Microsoft Word, da der er mere end 8 gange så mange PDF-filer end Word-filer.

Også publicering af videoklip på internettet ser ud til at være en stadig mere almindelig praksis. Således har den første tværsnitshøstning indsamlet mere end 88.000 videoklip.

Den gennemsnitlige filstørrelse for samtlige filer, der er indsamlet, ligger på ca. 40 Kb. Dette er en pæn stigning i forhold til en undersøgelse de to biblioteker lavede i 2001. Her var den gennemsnitlige filstørrelse ca. 34 Kb. Stigningen på 17 % vidner om stadig større båndbredde i de danske hjem og mere plads på webhotellerne rundt om i det danske land.



Figur 3: Objekter pr. domæne

Danske netsteders størrelse

Den første tværsnitshøstning blev reelt igangsat på ca. 579.000 domæner pga. problemer med domænenavne, der indeholder æ, ø og å. Statistikken kan derfor være en smule forvredet i de tilfælde, hvor domæner, der ikke kunne hentes størrelsesmæssigt, fordeler sig som andre domæner. Vi har ingen fornuftig grund til ikke at tro, at de gør, hvorfor statistikken er rimelig præcis.

Antallet af registrerede domæner er stadigt stigende. Således var der i oktober registreret 6495102 domænenavne mod de godt 607.000 der var på listen fra juni 2005 som indsamlingen tog udgangspunkt i. Antallet af domæner med de danske tegn er stort set uændret i samme periode – fra 28.250 til 28.426[3]. Danskerne kan åbenbart stadig blive ved med at finde ord og bogstavkombinationer, der endnu ikke er registreret hos DK-hostmaster.

Statistikken i figur 3 viser at næsten 70% af alle domænenavne dækker over netsteder, der rummer mellem 0 og 10 objekter – i praksis kun en 'forside' måske med et par billeder.

Kun 4,4 % af domænenavnene (26.024) rummer mere end 500 objekter og kun 0,6 % (3.500) rummer mere end 5000 objekter. Det sidste tal er der en vis usikkerhed på pga. de tidligere beskrevne begrænsninger. Et gennemsnitligt dansk netsted rummer således ca. 275 objekter.

Det meget store antal registrerede domæner har vist sig at ligge væsentligt over det reelle antal netsteder. En stor del af de registrerede domænenavne svarer nemlig enten slet ikke på DNS-opslag eller web-serveren, der har fået navnet, svarer ikke. Disse tilfælde dækker over lidt mere end 100.000 domænenavne eller godt 17 % af alle registrerede navne. Reelt er der altså kun arkiveret materiale fra 479.000 'levende' danske domæner.

Tabel 1

Mimetype	Mbytes	% af alle	Antal	% af alle	Filstørrelse
Imtex/html	1263870	25,56%	36914322	28,33%	35
Text / html	1177883	23,82%	68634852	52,68%	18
Application / pdf	839207	16,97%	68634852	1,20%	551
Tekst / plain	292066	5,91%	1559645	0,65%	255
Audio / mpeg	237070	4,79%	841932	0,07%	2846
Viodeo / x-ms-wmv	205802	4,16%	41164	0,03%	5120
Video / quicktime	157159	3,18%	26927	0,02%	5977
Image / gif	169266	3,42%	21013245	16.13%	8
Application / zip	166869	3,37%	79093	0,06%	2160
Application / octet-stream	124894	2,53%	229051	0,18%	558
Video / mpeg	97659	1,97%	20005	0,02%	4999
Application / x-shockwave-flash	92734	1,88%	632402	0,49%	150
Image / tif	56887	1,15%	11552	0,01%	5043
application / msword	34825	0,70%	190267	0,15%	187
Application / vnd.me-powerpoint	29038	0,59%	16800	0,01%	1770

Figur 4: Fordeling af Mbytes på danske danske netsteder

Dette antal modsvarer ikke antallet af 'unikke' netsteder, idet mange producenter (primært de kommercielle) har registreret flere domænenavne til samme netsted – såkaldte aliasser. Den første tværsnitshøstning har ikke forholdt sig til denne problematik, men bare hentet alle de netsteder, der 'var liv i' ud fra den samlede liste. Der arbejdes i øjeblikket på funktionalitet, der skal være med til automatisk at identificere disse aliasser for dermed at undgå at arkivere helt identiske netsteder under flere forskellige navne. Dette både for at spare plads i arkivet, men også for at sætte belastningen på de berørte web-servere dramatisk ned.

Målt på antal Mbytes fordeler de danske netsteder sig som vist i figur 4. Igen er der en vis usikkerhed på de største netsteder, men diagrammet viser med al tydelighed, at langt de fleste sites rummer under 100Mb data (98,2 %), hvilket svarer til at lige under 10.000 danske sites rummer mere end 100Mb. Et gennemsnitligt dansk netsted fylder således knap 12 Mb.

Statistikken kan også vise noget om, hvor udbredt anvendelsen af forskellige host-navne er på det danske domæne (sporten.tv2.dk / nyhederne.tv2.dk / politik.tv2.dk). Langt hovedparten af domænerne har kun www.domænenavn.dk. Kun 34.636 (7,3 %) af de besøgte domæner har mere end et host-navn defineret (og besøgt af netarkivets høster). Gennemsnitligt har disse domæner 5.14 host-navne. Langt de fleste har præcis to forskellige host-navne (de fleste bruger både www.domænenavn.dk og domænenavn.dk)

Af disse har 5829 mere end to host-navne. I denne gruppe har domænerne gennemsnitligt mere end 21 forskellige host-navne defineret, hvilket peger på, at når et domæne først har taget mere end 2 navne i brug, stiger chancen for at få endnu flere dramatisk. Det gennemsnitlige antal påvirkes i opadgående retning af et mindre antal domæner med rigtig mange host-navne. Topscoren ligger på mere end 20.000 host-navne til samme domæne. Her er ord i en ordbogslignende applikation brugt som host-navn, hvorfor tallet bliver unormalt stort.

Eksterne netsteder

Web-crawlersoftwaren er indrettet således, at der kun findes links og køres videre på de domæner, der blev inkluderet i start-listen. Dog arkiveres også 'eksterne' filer, der skal bruges for at genskabe websiderne – f.eks. billeder og lignende. Det har resulteret i, at netarkivets høstere har hentet materiale fra i alt 155.208 unikke domæner uden for DK-domænet. Her kan man hurtigt konkludere, at det er meget almindeligt at linke til fx billeder uden for sit eget netsted.

Antallet af objekter der er hentet fra servere uden for DK-domæner er dog 'kun' 7.934.537, hvilket svarer til lige godt 6 % af det samlede antal. Fra eksterne servere er der altså hentet gennemsnitligt 51 objekter pr. domæne.

Konklusion

Den første tværsnitshøstning af det samlede danske DK-domæne har vist en række interessante ting omkring det danske domænes størrelse og indhold:

1. Der var i perioden juli-oktober 2005 omkring 479.000 DK-domæner, der var i live.
2. Et gennemsnitligt netsted rummer 275 objekter og fylder knap 12 Mbytes
3. 98,2 % af alle domæner rummer mindre end 100Mb data
4. JPG-billeder er den filtype, der sammenlagt fylder mest
5. HTML-filer er den filtype, der sammenlagt er flest af
6. PDF-filer er langt det mest anvendte til publicering af dokumenter, der ikke ligger i HTML-format.
7. Den gennemsnitlige filstørrelse vokser i takt med båndbredde / serverplads
8. Brugen af flere host-navne på et domænenavn er forholdsvis udbredt og tager man først mere end 2 navne i brug, stiger chancen for at flere kommer til kraftigt.

Erfaringerne fra denne indsamling har vist, at både aliasser og crawlertraps er et reelt problem. Netarkivet arbejder derfor på at finde metoder der kan automatisere identifikationen af disse. På grund af en meget varierende filstørrelse fra det ene netsted til det andet har den øvre grænse på antallet af objekter vist sig at være uhensigtsmæssig, hvorfor der nu opereres med en øvre grænse på antal bytes i stedet (fx kan et netsted med 5000 videofilm fylde rigtig meget).

Netarkivet undersøger samtidig muligheden for duplikatreduktion, idet stikprøver har vist, at mindst 50 % af materialet ikke ændrer sig hvorfor efterfølgende tværsnitshøstninger vil hente og arkivere rigtig meget redundant data. Duplikatreduktionen kan foretages enten efter download, hvilket ikke giver mindre belastning for producenterne / netværket, eller under nedtagningen ved ikke at genhente statiske data, fx JPG, GIF, PDF der ikke har ændret sig siden sidste gang.

Planen er at generere statistikker efter hver tværsnitshøstning, så der efterhånden bliver et godt sammenligningsgrundlag. Flere af tallene er interessante at følge – fx udbredelsen / brugen af bestemte formater, der kan falde / stige eller den gennemsnitlige størrelse på både enkelt objekter og netsteder som et hele.

For yderligere information besøg: http://netarkivet.dk

Noter

[1] Udviklet primært af Internet Archive (http://crawler.archive.org) men i samarbejde blandt andet med de nordiske nationalbiblioteker gennem IIPC-samarbejdet (http://www.netpreserve.org)

[2] www.dk-hostmaster.dk/index.php?id=235

[3] www.dk-hostmaster.dk/index.php?id=116