netarkivet.dk

Nyhedsbrev december 2007

Velkommen til Netarkivets nyhedsbrev for december 2007.

International workshop om NetarchiveSuite

Netarkivet afholdt i september 2007 en to-dages workshop om Netarkivets system: NetarchiveSuite. Der var 17 eksterne deltagere fra ti forskellige lande, primært europæiske, men også USA, så interessen var stor. Workshoppen rummede oplæg, der spændte fra meget tekniske emner som installation og konfigurering til mere brugerorienterede sessions omkring håndtering af selve indsamlingen af netmateriale.

Efterfølgende har der været livlig aktivitet på systemets officielle hjemmeside, hvor NetarchiveSuite er blevet downloadet mere end 50 gange fra mange forskellige lande, og ligeledes har trafikken på online manualer og guidelines været pæn, hvilket tyder på, at mange har både installeret og afprøvet systemet.

Flere landes nationalbiblioteker overvejer således i øjeblikket, om NetarchiveSuite kunne være det system, de skal bruge på nationalt plan. Netarkivet håber at kunne få flere med, således at fremtidens udvikling og vedligeholdelse af systemet vil kunne klares i fællesskab.

Fjerde tværsnitshøstning

Den fjerde tværsnitshøstning blev igangsat 1. marts 2007. Den inkluderede ca. 850.000 domæner, hvoraf ca. 78 % er aktive – det vil sige, at Netarkivet indsamlede ca. 660.000 danske hjemmesider.

Tværsnitshøstningen blev afsluttet 14. juli og fylder 15.1 Tbytes. Det er væsentlig mere end den forrige tværsnitshøstning, som ”kun” fyldte 9.8 Tbytes. Den store stigning skyldes flere ting:

- Netarkivet høstede for første gang 42.000 nye domæner uden for .dk. Den første høstning af nye domæner fylder altid mere, idet der ikke findes dubletter i forhold til tidligere høstninger.
- Disse nye domæner er gennemsnitligt større end DK-domænerne. Således fylder et DK-domæne gennemsnitligt ca. 16 Mbytes, hvorimod et domæne uden for .dk fylder ca. 46 Mbytes.
- Web 2.0 slår for alvor igennem. Det betyder, at der i denne tværsnitshøstning er hentet 3.5 Tbytes indlejret materiale, altså materiale, der ikke kommer fra domænerne selv, men er blevet indlejret fra andre domæner i den store verden.

Den store mængde indlejret materiale betyder, at det nu skal undersøges, om netarkivets funktionalitet med en øvre grænse per domæne skal lægges om til også at gælde det indlejrede og ikke kun som nu, materiale fra domænet selv.

Femte tværsnitshøstning

Netarkivets femte tværsnitshøstning blev startet 17. juli. Første skridt – op til 10 Mbytes per domæne – er afsluttet. Der var ca. 95.000 domæner, der ramte 10 Mbytes-grænsen og derfor blev inkluderet i næste skridt, som går op til maksimalt 2 Gbytes per domæne.

Der er en klar stigning i antallet af domæner, der er større end 10 Mbytes, siden 2005, hvor kun godt 50.000 domæner var større end 10 Mbytes.

Indsamling af sider med login

Indsamling af netsteder beskyttet af brugernavn/password er nu for alvor ved at komme i gang. Således indsamler Netarkivet nu fem netsteder, hvor høsteren foretager login og høster den beskyttede del af netstedet. Indsamling af flere selektive netsteder med login er undervejs.

At konfigurere høsteren til at bruge login under indsamlingen har vist sig at være en temmelig tidskrævende proces, som ofte kræver analyser og indtil flere test-høstninger, inden det fungerer.

Redaktionsgruppen bad tidligere på året om en redegørelse for fænomenets omfang – altså en undersøgelse af, hvor mange danske netsteder der benytter login. Der findes basalt set to typer logins på internettet: HTTP-logins, som er en lille boks, hvori browseren beder om brugernavn og password, og HTML-logins, hvor login-mekanismen ligger indlejret på en netside som en del af indholdet og derfor kan rumme andre felter end brugernavn og password.

En automatisk gennemgang af Netarkivets tværsnitshøstninger afslørede følgende antal logins på danske netsider:

- HTTP-logins: 3.706.430 logins på 91.976 unikke domæner
- HTML-logins: 1.318.304 unikke login-URL'er fordelt på 119.192 unikke domæner

Anvendelsen af login er – ikke overraskende – meget hyppig. En stikprøve skal nu afsløre, hvor mange af disse logins, der rent faktisk gemmer indhold, der er omfattet af pligtafleveringsloven, idet det umiddelbart ser ud til, at rigtig mange af disse logins beskytter helt privat indhold og administrative dele af netsider.

Status på bitarkiverne

Bitarkiverne rummer nu følgende datamængder:

Tværsnitshøstning:	51.077 Gbytes i 526.931 filer
Selektiv høstning:	5.964 Gbytes i 71.145 filer
Begivenhedshøstning:	5.847 Gbytes i 60.683 filer

Totalt har Netarkivet altså passeret 62 Tbytes.

Der er stadig 100 % konsistens mellem de to online-kopier på henholdsvis Statsbiblioteket og Det Kongelige Bibliotek. Efter flytning af filer fra en ældre til en ny bitarkivmaskine på Det Kongelige Bibliotek har bitbevaringsrutinerne og brugerinterfacet til bitbevaringssystemet for alvor vist sit

værd, idet der for første gang er genetableret en manglende fil, efter at systemet selv identificerede fejlen og efterfølgende kunne rette den op igen.

Redaktionsgruppemøde

Netarkivets redaktion holdt møde 24. september 2007 på Statsbiblioteket. De vigtigste punkter på mødet var:

- gennemgang af status for de selektive høstninger, som meget snart nærmer sig målet på 80 netsteder
- gennemgang af analysen omkring brug af password-beskyttelse på internettet
- indsamlingsdokumentation i form af en systematisk logning af netarkivets indsamlingsrutiner – f.eks. opgradering af høster-softwaren, og hvilken betydning det har for indsamlingen. (f.eks. inkludering af indsamling af nye teknologier)

Folketingsvalg

Netarkivet var på dagen for udskrivelse af folketingsvalget klar til at indsamle. Folketingsvalget var ventet og var derfor blevet forberedt grundigt, så Netarkivet kunne begynde indsamlingen kun fire timer efter, at Anders Fogh havde udskrevet valget.

Valgkampen blev denne gang ført mere på internettet end nogensinde tidligere, og de faglige medarbejdere havde derfor travlt med at identificere relevante netsteder. Valgkampen blev ført på en del nye steder såsom facebook.com og youtube.com, hvorfor også relevante dele af disse blev samlet ind.

Desuden blev der gjort en særlig indsats for at sikre multimediemateriale såsom video, og med hjælp fra Internet Archive til konfigurering af høsteren blev videomateriale fra især youtube.com en meget vigtig del af denne begivenhedshøstning.

Grundet de nye teknologier, mange netsteder og intensiv indsamling kom begivenhedshøstningen til at fylde ca. 2.2 Tbytes data, hvilket er væsentlig mere end tidligere begivenhedshøstninger. Web 2.0 tendensen slog også igennem her, og det skal undersøges nærmere, om høsteren eventuelt henter materiale fra andre domæner, end det egentlig er nødvendigt.

Alvorlig fejl i netarkivets indekseringsmekanisme

Der blev i efteråret fundet en alvorlig fejl i den indekseringsmekanisme Netarkivet benytter til at lave indekser til at foretage dublikatreduktion og til at browse det indsamlede materiale. Fejlen betød at tilfældige objekter ikke blev indekseret. Det har i praksis betydet to ting:

- dublikatreduktionen har ind til nu ikke fundet så mange dublikater som der faktisk har været.
- ved manuel kontrol af det høstede materiale har der ind i mellem set ud til at mangle objekter, selvom de faktisk var høstet.

Fejlen har altså ikke givet tab eller mangler i forhold til indsamlingen og løsningen af fejlen har umiddelbart givet positive resultater idet dublikatreduktionen nu ser ud til at give mere end 40% pladsbesparelse mod de tidligere omkring 30%.

Desværre har løsningen af fejlen også bevirket at indekset for en hel tværsnisthøstning nu er blevet så stort så den applikation netarkivet bruger til at browse i det høstede materiale ikke længere fungerer. Det er et problem der arbejdes på at løse.

Planer for den kommende periode

1. Løsning af problem med for stort indeks til at browse en hel tværsnitshøstning
2. Analyse af den femte tværsnitshøstning
3. IIPC Technical Committee møde i Paris 14.-15. januar 2008
4. Support i forbindelse med udgivelsen af NetarchiveSuite