netarkivet.dk

Nyhedsbrev april 2006

Velkommen til Netarkivets nyhedsbrev april 2006.

Nyt design til netarkivet.dk

Netarkivet har fået det nye design implementeret på http://netarkivet.dk. Alle reminiscenser af tidligere udgaver af netarkivet.dk er fjernet, og al relevant information er flyttet til det nye netsted. Det nye design er blevet annonceret på en international mail-liste for webarkiveringsinteresserede sammen med en artikel omkring netarkivets første tværsnitshøstning, og det er blevet godt modtaget. Vi har fået flere henvendelser fra interesserede, der vil linke til netarkivets netsted. Desuden er artiklen blevet hentet 502 gange (robotter ikke talt med) mellem 28. marts og 14. april, heraf 464 gange på engelsk – så der er en ret stor international interesse for det vi går og laver.

Der arbejdes i øjeblikket med en elektronisk blanket på netarkivet.dk hvor den almindelige dansker kan hjælpe netarkivet med forslag til netsteder der skal samles ind under alle 3 indsamlings-strategier.

Omprioritering af høster-ressourcer

Netarkivet benytter nu heritrix version 1.7.1 med meget få modifikationer. Denne version har vist sig at kræve flere computerkræfter end tidligere versioner, især efter indførelsen af grænserne per domæne idet høsteren før hentning af alle URL'er skal tjekke, om grænsen for det pågældende domæne er nået.

Derfor af netarkivet valgt at omprioritere høster-ressourcer, således at hvor der tidligere "kun" blev tværsnitshøstet fra KB og selektiv/begivenhedshøstet fra SB, så kombineres nu alle 4 høster-maskiner, således at de alle bidrager til alle typer høstninger.

Anden tværsnitshøstning

Andet skridt af netarkivets anden tværsnitshøstning med en grænse på 500 Mbytes per domæne blev sat i gang i starten af februar på de ca. 61.000 domæner, der ramte grænsen på 10 Mbytes. Denne høstning forventes færdig i slutningen af april, og den samlede høstningstid for anden tværsnits-høstning når dermed op på 4½ måned. Høstningen har ligget stille i nogle få kortere perioder, men den kvikke læser kan hurtigt regne ud, at de 3 måneder, der egentlig var afsat til en enkelt tværsnits-høstning er overskredet. Netarkivet må således overveje, om der trods ovennævnte omkonfigurering evt. skal flere høstermaskiner til for at klare opgaven.

Datamængden ved en færdiggørelse på ca. 80 % af den nuværende høstning udgør lidt mere end 6TBytes, så forventningen er, at dette skridt kommer op mellem 7 og 8TBytes. Det betyder, at den anden tværsnitshøstning efter dette skridt totalt kommer op i nærheden af 10Tbytes.

Vi ved endnu ikke, hvor mange domæner der rammer grænsen på 500Mbytes, men forventningen er at ca. 4000 domæner er større end den grænse. Disse skal manuelt gennemgås, og for udvalgte, fx officielle (folketinget, partierne, ministerier, kommuner, amter…) og store "vigtige" netsteder (fx dr.dk, tv2.dk) skal maxgrænsen hæves, og endnu et skridt af anden tværsnitshøstning skal køres.

Forventningen er, at omkring 1000 netsteder skal høstes med en højere grænse (fx 5Gbytes - og for få udvalgte måske mere). Derfor kommer dette sidste skridt sandsynligvis til at bruge mindst 2Tbytes lagerplads.

Det bringer en samlet anden tværsnitshøstning op på i nærheden af 12Tbytes. Det er mere end forventet – i 2003 estimerede netarkivets pilotprojekt en samlet tværsnitshøstning til ca. 5Tbytes. Det betyder i praksis, at behovet for duplikatreduktion rykker meget nærmere. Uden duplikat-reduktion vil 4 årlige høstninger således fylde i nærheden af 50Tbytes. Da duplikat-reduktion reducerer det samlede lagerbehov med mere end 50 %, vil sådanne mekanismer hjælpe meget på situationen.

Netarkivet bliver samtidig nødt til at overveje, om 4 årlige tværsnitshøstninger stadigt er nødvendigt – eller om vi stadigt lagermæssigt har "råd" til 4 årlige tværsnitshøstninger. For at understøtte den beslutning har styregruppen besluttet, at der ultimo 2006 skal laves en analyse af resultaterne af de tværsnitshøstninger, der har været gennemført til den tid, for at vurdere, om frekvensen kan ændres eller eventuelt differentieres.

Anden begivenhedshøstning

Som skrevet i sidste nyhedsbrev startede netarkivet i starten af februar en begivenhedshøstning omkring Muhammed-krisen. Det er en lidt speciel begivenhedshøstning, idet det kan være svært at sige, hvornår krisen slutter. I slutningen af marts / starten af april blev høstningerne neddroslet fra daglige til ugentlige for langt de fleste identificerede sites.

Perioden på næsten 2 måneder med intensive daglige høstninger har også betydet, at denne indsamling lagermæssigt fylder mere end estimeret for begivenhedshøstninger. Således er der pt. indsamlet lidt mere end 2Tbytes data, hvor estimatet hidtil har sagt ca. 1 Tbytes per begivenheds-høstning.

Neddroslingen har betydet, at der nu indsamles ca. 20Gbytes pr. uge mod tidligere mere end 30Gbytes dagligt i den intensive periode.

Planen er at fortsætte den neddroslede indsamling nogle uger endnu, således at vi har indsamlet dokumentation for mellemperioden, såfremt krisen skulle blusse op igen. På et tidspunkt må netarkivet selvfølgelig erklære krisen for overstået for netarkivets vedkommende.

Udviklingsopgaver

Udviklingsprojektet omkring netarkivet har i den seneste 2-måneders periode koncentreret sig omkring færdiggørelse af domænegrænse-funktionaliteten og understøttelse af kvalitetssikring, således at den nu inden længe forhåbentligt bliver muligt at kigge på de høstede data stort set umiddelbart efter, de er høstet – og ikke først efter uger eller måneder, når en total indeksering af hele arkivet er blevet gennemført.

Status på bitarkiverne

Datamængden er i de seneste 2 måneder steget dramatisk, da tværsnitshøstningen for alvor er kommet i gang igen, og da begivenhedshøstningen har krævet en del plads. Således er den samlede datamængde vokset fra 9.6Tbytes til næsten 17Tbytes, der fordeler sig på følgende måde:

Tværsnitshøstning: 12.920 GBytes
Selektiv høstning: 1841 GBytes
Begivenhedshøstning: 2617 GBytes

Der har været lidt knas med uploadmekanismen i systemet, så vi kan nu konstatere, at der mangler omkring 100 filer i enten det ene eller andet bitarkiv – det er ikke filer, der er forsvundet, men filer der på grund af upload-problemerne kun er kommet i det ene af de to bitarkiver. Systemets logik gør, at de høstede data først fjernes fra høstermaskinerne når de er sikret på begge bitarkiver, så der er, så vidt vides, stadigt 100 % konsistens i netarkivets data trods de nu knap 17Tbytes data.

Redaktionsgruppe

Netarkivet havde 13. marts det første møde med den nedsatte redaktionsgruppe. Mødet blev afholdt på Statsbiblioteket, og gruppens medlemmer samt de biblioteksfaglige fra både KB og SB havde gode diskussioner om problemstillinger og muligheder. Gruppens medlemmer virker til at have fingeren meget på pulsen omkring, hvad der rør sig på internettet, og hvordan man hurtigst muligt opfanger, når noget nyt opstår – fx gennem samarbejde med de danske søgemaskiner (fx http://jubii.dk) og statistik-services som fx http://chart.dk.

Redaktionsgruppen mødes igen i august.

Styregruppe

Styregruppen for netarkivet havde møde på Statsbiblioteket den 8. marts, og på dagsordenen var ud over driftslederens beretning især diskussion om begivenhedshøstninger samt planer og ønsker for både kommende kvartal samt prioritering af opgaver 2006.

Overordnet vil samlingsafdelingerne koncentrere sig om det daglige arbejde med systemet – især konfiguration af de selektive høstninger samt analyse af tværsnitshøstningerne og indsamling af udvalgte af de danske netsteder, der overstiger 500Mbytes i størrelse. Desuden skal der arbejdes med udarbejdelsen af retningslinjer for selektion af netsteder og valg af begivenheder.

Udviklingsmæssigt koncentrerer projektet sig i 2006 om stabilisering, forbedring af kvalitets-kontrolmulighederne og overvågningsværktøjerne.

Bitarkiverne lever allerede et nogenlunde stabilt liv, vokser støt og roligt, så 2006 vil for deres vedkommende handle meget om "krav til trusted repositories" og netarkivets honorering af disse.

Styregruppen har næste møde på KB den 26. juni.

IIPC møde på KB

29-31. marts var der IIPC møde i København. Bjarne præsenterede netarkivets system og bruger-grænseflade i en delvist engelsk oversat udgave. Det blev særdeles godt modtaget, og især de nordiske lande og England er meget interesserede i vores system. Muligheden for samarbejds-projekter er helt sikkert til stede og vil blive på dagsordenen til næste styregruppemøde.

Planer for den kommende periode

1. Løbende overdragelse af det daglige arbejde med systemet til national-/pligtafleverings-afdelingerne
2. Udvidelse af antallet af selektive netsteder fra 40 og op imod de planlagte 80
3. Afslutning af den anden tværsnitshøstning
4. Elektronisk blanket til forslag fra den almindelige dansker
5. Krav til "Trusted Repositories"