netarkivet.dk

Nyhedsbrev juni 2006

Velkommen til Netarkivets nyhedsbrev juni 2006.

Status år 1

Netarkivet er nu rigtig godt på vej mod sin etårs fødselsdag, og det er tid at gøre status over året, der er gået. Indsamlingsmæssigt har vi nået rigtig meget men der er stadigt også et stykke vej, før vi kører på fuld skrue. Netarkivet har gennemført 3 tværsnitshøstninger hvoraf den ene er en mindre høstninger op til 10Mb pr. domæne hvilket dækker mere end 85 % af domænerne men naturligvis ikke kommer ret langt på de større domæner. Ingen af de 3 tværsnitshøstninger er kommet helt i bund på de allerstørste sites idet der ikke endnu er høstet forbi 500Mb pr. domæne.

De selektive høstninger har kørt stabilt på 40 udvalgte sites og de kommende måneder skulle meget gerne bringe antallet op på de planlagte 80.

Der har været gennemført to begivenhedshøstninger omkring folketingsvalget og muhammed-krisen. Disse har vist hvor vigtigt det er at være parat, når begivenheden opstår og konstant have fingeren på pulsen.

Der har naturligvis været udfordringer undervejs. Problemer hos producenterne med for stort trafikload på deres servere, for mange varer i den elektroniske indkøbskurv og crawlertraps der resulterede i arkiveringen af ligegyldig data har været nogle af de opgaver, der har skullet løses undervejs. Alle udfordringer er blevet løst til stor tilfredshed for såvel netarkivet som langt de fleste producenter. Antallet af henvendelse omkring problemer har været under 100 og i betragtningen af, at der pt. er registreret mere end 700.000 .dk-domæner, må det siges at være et meget lille antal.

Et hvert system har børnesygdomme og således led netarkivets system i begyndelsen en del under hastighedsproblemer i det administrative brugerinterface. Disse er nu løst, og det har gjort hverdagen for det biblioteksfaglige personale meget lettere.

Der er kommet gang i arbejdet med den nedsatte redaktionsgruppe. Der har været afholdt et enkelt møde, og det viste stor interesse og entusiasme omkring netarkivets aktiviteter. Redaktionsgruppen mødes igen i august.

Konklusionen på netarkivet år-1 må være, at vi har været igennem en indkøringsperiode, der trods små problemer undervejs har kørt ganske fornuftigt. Den danske del af internettet er blevet indsamlet i stor stil efter de tre forskellige strategier og kvaliteten af det indsamlede ser rimeligt fornuftig ud. Hovedtemaet for år-2 bliver konsolidering og udrulning af netarkivets system i Open Source – mere herom senere i nyhedsbrevet.

Nyt på netarkivet.dk

Netarkivets netsted har for nylig fået en formular, hvor den almindelige danske kan hjælpe med at indrapportere relevante netsteder uden for .dk-domænet til tværsnitshøstning samt kandidater til netsteder der burde indsamles både gennem de selektive høstninger og begivenhedshøstningerne. Der kommer meget snart en pressemeddelelse fra netarkivet, og den skulle meget gerne lede

opmærksomheden lidt på denne formular såvel som hjælpe på den almindelige danskers kendskab til netarkivet.

www.netarkivet.dk har stadig en jævn trafik og det tydeligt, at vi har opmærksomhed udefra – også internationalt. For nyligt blev en artikel skrevet af Niels H. Christensen på KB omkring netarkivets bitbevarings-system lagt på nettet. Artiklen blev annonceret på en international mailing-liste for webarkiverings-interesserede, og allerede 48 timer senere var artiklen hentet knap 50 gange.

Anden tværsnitshøstning

Netarkivets anden tværsnitshøstning blev afsluttet primo maj 2006.
Høstningen fyldte 8431 Gbytes, hvilket var noget mere end forventet. Det satte fokus på nødvendigheden af duplikatreduktion, der skal sørge for, at vi ikke gemmer det samme objekt mere end en gang. Udviklingsressourcerne er således blevet omprioriteret og næste release af netarkivets system er planlagt til primo august. Dette release inkluderer duplikatreduktion og erfaringerne viser, at denne mekanisme kan betyde helt op til 50 % besparelse på diskpladsen

Ca. 6300 domæner ramte grænsen på 500Mbytes – disse kigges der nu manuelt på, og de der findes relevante får tildelt en højere grænse. Det anslås at ca. 1500 domæner skal indsamles mere i dybden end 500Mbytes. De domæner der ind til videre kun indsamles op til 500Mbytes er fx private fotoalbums og netsteder for firmaer, der ikke handler med egne produkter.

Arbejdet med at gennemgå de mange domæner har også afsløret en del såkaldte alias'er, hvor samme netsted kan tilgås under flere forskellige navne. Denne information kan lægges i systemet, der derefter selv finder ud af kun at indsamle en version af netstedet. Ligeledes identificeres en del crawlertraps (især kalender-applikationer). Begge dele er med til at nedbringe pladsforbruget.

Tredie tværsnitshøstning

Netarkivets tredje tværsnitshøstning blev igangsat den 8. juni med en grænse på maksimalt 10Mbytes pr. domæne. Denne høstning er startet helt forfra på de nu mere end 700.000 danske domæner der er registreret hos DK-hostmaster.

Anden begivenhedshøstning

Netarkivets anden begivenhedshøstning, der omhandlede krisen som følge af muhammedtegningerne, er nu afsluttet. De faglige medarbejdere vurderer, at krisen nu har nået et niveau, hvor den dækkes af de daglige nyhedshøstninger. Skulle krisen blusse op igen er definitionerne i systemet klar til, at der bare kan trykkes på knappen, der starter indsamlingen af de udvalgte netsteder igen.

Indsamling af de meget store netsteder

Det er blevet besluttet at håndteringen af de meget store (og vigtige) netsteder skal håndteres uden for den almindelige tværsnitshøstning idet enkelte sites er så store, at de alene sandsynligvis er større end resten tilsammen. Derfor er en specialindsamling af dr.dk og tv2.dk blevet igangsat med en grænse på 100Gbytes pr. domæne. Forventningen er, at begge netsteder vil ramme grænsen, da de både er store og sandsynligvis rummer crawlertraps der skal håndteres for at komme helt i bund. Næste gang de indsamles, kan grænsen derfor sættes op om nødvendigt efter en analyse af de første 100Gbytes.

I fremtiden vil andre meget store sites kunne overflyttes fra den almindelige tværsnitshøstning til denne særlige høstning for meget store netsteder.

Netarkivets system i Open Source

Netarkivets udviklingsgrupper planlægger i øjeblikket, hvordan det udviklede system kan lægges ud i Open Source. Der har gennem længere tid været stor international opmærksomhed og interesse omkring systemet – specielt fra de nordiske lande.

Planen er derfor, at netarkivets system modulariseres lidt mere end det er i dag og derefter lægges ud i Open Source i etaper i 2007, således at det første modul bliver frigivet i foråret 2007, og det sidste modul gerne skulle blive tilgængeligt ultimo 2007

Et frugtbart samarbejde med Island omkring duplikatreduktion har gjort Islændingene til beta-testere af systemet, og efter planen skal en test-installation være klar på Island ultimo 2006

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 godt og vel 21Tbytes, der fordeler sig på følgende måde:

Tværsnitshøstning: 16374 GBytes
Selektiv høstning: 2214 GBytes
Begivenhedshøstning: 2929 Gbytes

Den aktive bitbevaring har givet 100 % konsistens i begge bitarkiver der nu rummer næsten 225.000 filer uden en eneste fejl. Integriteten tjekkes løbende og eventuelle fejl kan via brugergrænsefladen rettes op.

Budgettet for pladsforbruget siger ind til videre 37Tbytes til 31/12 og det er vores forventning, at dette budget holder.

Netarkivet afslører sikkerhedshul

Netarkivet fik i maj en henvendelse fra en producent, der havde oplevet, at netarkivets høster havde været inde i producentens netsteds administrations-modul og slette artikler. Det var naturligvis en kedelig sag, men de tabte data blev pr. konduite og til producentens store tilfredshed leveret tilbage fra netarkivet og i tilgift fik producenten lukket det sikkerhedshul høsteren havde afsløret.

Planer for den kommende periode

1. Udvælgelse af store sites der skal indsamles mere end 500Mbytes fra
2. Implementering af duplikatreduktion
3. Planlægning af udrulningen af netarkivets system til Open Source
4. Redaktionsgruppemøde den 14. august