netarkivet.dk

Nyhedsbrev september 2006

Velkommen til Netarkivets nyhedsbrev for september 2006.

Sommerferie

På grund af sommerferieperioden kommer nyhedsbrevet først nu. idet de fleste aktiviteter har ligget stille henover sommeren. De selektive høstninger har naturligvis kørt som planlagt, men tværsnitshøstning og system-udvikling har været næsten sat helt i bero på grund af ferie.

Styregruppemøde

Styregruppen for netarkivet havde styregruppemøde på KB den 26. juni. Mødet omhandlede foruden status på driften generelt (og det går godt), status fra de 6 områder især handlingsplaner for det kommende år for indholdsafdelingerne (pligtafleveringen på KB og nationalområdet på SB) og udviklingsgrupperne.

Udviklingsgrupperne koncentrerer den kommende periodes indsats omkring forbedring af systemets overvågningsmuligheder samt modularisering af systemet som forberedelse til den forestående open source release. Pligt/national har i den forgangne periode arbejdet meget med definition af, hvad en begivenhed er i netarkivets sammenhæng og retningslinjer for hvordan vi skal håndtere netsteder uden for .dk domænet samt retningslinjer for arkivering af netsteder der er større end 500Mb – mere om det sidste senere.

En begivenhed i netarkivets definition kan overordnet karakteriseres ved:

- At begivenheden skaber debat i befolkningen og menes at have en betydning for Danmarks historie eller udvikling
- Begivenheden udløser nye netsteder
- Begivenheden behandles i stort omfang på eksisterende netsteder

Netsteder uden for .dk domænet bør indsamles hvis de opfylder et eller flere af disse kriterier:

- sproget: tekster, vejledninger, etc. er udelukkende eller for mere end 50% vedkommende formuleret på dansk
- registranten af et domænenavn har fast bopæl i Danmark
- registranten af et domænenavn er af dansk oprindelse og har websider, der er rettet mod et publikum i Danmark
- kampagner rettet mod borgere i Danmark
- borgere i Danmark, der anvender andre sprog end dansk

Svenske erfaringer viser af 30 % af de nationale domæner ligger uden for nationens eget domæne. Det betyder op mod 200.000 danske domæner uden for .dk. Der foreligger nu et stort arbejde med indentifikation af disse. Grundet antallet skal der givetvis udviklers systemer, der kan automatisere processen.

Redaktionsgruppemøde

Redaktionsgruppen for netarkivet havde møde på KB d. 14. august. Redaktionen gennemgik handlingsplanerne for det kommende år og havde ikke de store kommentarer. Redaktionen gennemgik ligeledes oplæggene til definition af en begivenhed, retningslinjer for netsteder uden for .dk samt retningslinjer for arkivering af de store danske netsteder. Der var kun mindre kommentarer til disse 3 oplæg der alle blev godkendt af redaktionen.

Mere diskussion kom der op under gennemgangen af en bruttoliste på 80 kandidater til de selektive høstninger som de faglige medarbejdere på SB havde udarbejdet. Redaktionsgruppen kom med deres holdninger til de foreslåede kandidater og kom med nye forslag til en del andre, der kunne være interessante. Konklusionen på dette punkt er, at det stadig er meget arbejde for de faglige medarbejdere på SB, der har ansvaret for disse indsamlinger.

De store websites (større end 500Mb)

Netarkivets sidste tværsnitshøstning blev gennemført med en maksgrænse på 500Mb pr. domæne. Der var 6300 domæner der ramte grænsen, og et stort manuelt arbejde med at gennemgå disse blev gennemført af medarbejderne i pligtafleveringen på KB henover sommeren. Da vi i øjeblikket ikke har hverken lagerplads eller hardware-ressourcer til at høste totalt i bund på hele det danske domæne, blev der udviklet nogle retningslinjer for hvilke domæner, der skulle arkiveres mere fra, og hvilke der ind til videre kunne stoppes på de 500Mb.

Disse retningslinjer foreskriver, at der ind til videre IKKE arkiveres mere end 500Mb fra følgende typer netsteder:

1. Sider hvis indholdet ikke er dansk, men på .dk-server
2. Sider med privat indhold, typisk med mange billeder af privat karakter, såsom billeder / videoklip af børn, familie, rejser, hobbyer, enkelte skoleklassers websider o. lign.
3. Sider, hvis hovedformål er netsalg, der ikke er af egne produkter
4. Sider med traditionel porno

De 6300 store domæner blev gennemgået i forhold til disse retningslinjer og 2368 domæner slap gennem kontrollen og fik således en ny grænse på 1Gb. Efter næste tværsnitshøstning skal de der rammer den nye grænse gennemgås igen med henblik på en eventuel endnu højere grænse.

Duplikatreduktion

Udviklingsgruppen har hen over sommeren integreret et Islandsk modul til heritrix – netarkivets høster – der kan lave automatisk duplikatreduktion. Dette blev sat i drift 28. august og har vist sig at være særdeles effektiv. Desværre viste der sig en fejl i systemet der gjorde, at da der skulle genereres et indeks til at finde duplikater i til en ny tværsnitshøstning, skalerede systemet ikke til den store datamængde, der allerede nu er i produktionssystemet. Det var en tydelig reminder på, at det kan være meget svært at teste software, der skal håndtere så store datamænger, uden faktisk at teste på produktionsdata.

Fejlen er nu rettet og den næste tværsnitshøstning er sat i gang. Erfaringerne med duplikatreduktionen har for de selektive høstninger være meget positive og har betydet besparelser på lagerpladsen for de daglige høstninger på mellem 50 og 70 procent, hvilket må siges at være en betydelig gevinst.

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. Høstningen blev afsluttet 8. august og godt 70.000 domæner ramte grænsen på 10Mb.

Næste tværsnitshøstning

Den næste tværsnitshøstning blev igangsat d. 25. september efter en opgradering af systemet der rettede den tidligere omtalte systemfejl. Efter en generering af et meget stort duplikatindeks, der i sig selv tog tre en halv dag, er den første tværsnitshøstning med duplikatreduktion gået i gang. Det ser meget lovende ud og vi forventer en pladsbesparelse på 30-40 %. Det bliver især spændende at følge, om hastigheden af tværsnitshøstningen kan holdes nogenlunde på niveau med de tidligere høstninger på trods af duplikatreduktionen, der uden tvivl burger en del maskinkraft.

Indsamling af de meget store netsteder

En særlig indsamling af de meget store netsteder: dr.dk og tv2.dk blev igangsat 22. juni med et loft på 100Gb pr. domæne. Denne høstning viste efter få dage, at der er mange udfordringer i at hente så meget data fra et netsted, uden at høsteren falder i de såkaldte crawler-traps – områder på netstedet hvor høsteren går i ring i det uendelige.

Begge domæner blev dog afsluttet 12. juli og resultatet afslører, at de begge nåede grænsen på de 100Gb. Danmarks Radios netsted havde endnu flere millioner objekter, der ikke var hentet mens høstningen af tv2.dk så ud til at være på vej mod at være komplet.

Disse høstninger skal nu gennemgås for de føromtalte crawlertraps inden grænsen hæves og de høstes på ny – denne gang inklusiv duplikatreduktion der helt givet vil spare en masse diskplads.

Selektive høstninger

De selektive høstninger er i øjeblikket i gang med at blive gennemgået og omdefineret. De har nu kørt i over et år med nogenlunde de samme konfigurationer. Der er naturligvis lavet en løbende kvalitetskontrol, men det er nu på tide at få dem gennemgået og gennemtænkt igen. Det er et stort arbejde at overskue de ofte meget store netsteder og udvælge hvilke sektioner der skal arkiveres med hvilken frekvens på de enkelte sites. Dette arbejde forventes derfor at vare i nogle måneder endnu.

IWAW'06

Netarkivet var repræsenteret på "International WebAchiving Workshop – 2006" i Alicante (Spanien) med Lars Clausen fra udviklergruppen og undertegnede. Lars lavede et glimrende indlæg omkring netarkivets system som en forsmag på den kommende open source release. Trods store vanskeligheder med netværk og teknik i konferencelokalet fik vi gennemført en succesfuld reklame for netarkivet og vores system.

Som forventet var det især de nordiske lande der var særligt interesseret, men også Tjekkiet viste stor interesse i vores system, da det er det eneste der i praksis er bygget til og faktisk kan gennemføre tværsnitshøstninger af f.eks. nationale domæner. Der blev præsenteret et par andre systemer på konferencen, men de var begge udviklet specifikt til arkivering i mindre skala.

Status på bitarkiverne

Bitarkiverne rummer nu følgende datamængder:
Tværsnitshøstning: 17538 Gb i 181101 filer
Selektiv høstning: 2767 Gb i 32322 filer
Begivenhedshøstning: 2929 Gb i 30834 filer

Totalt har netarkivet altså nu arkiveret ca. 23Tb hvilket er en anseelig mængde data. Det ligger dog stadig inden for budgettet der rummer plads til 37Tb ved årsskiftet.

Tidens tand begynder også at slide på hardwaren og således er 2 bitarkivmaskiner ud af 22 maskiner på KB i øjeblikket ude af funktion. Den ene er sandsynligvis helt tabt på grund af disknedbrud mens situationen for den anden pt. er ukendt.

Heldigvis har vi jo 2 kopier mere af alle filer og vores bitbevarings-rutiner skal derfor stå deres første rigtige og helt igennem praktiske prøve.

Planer for den kommende periode

1. Netarkivets system testes på Island
2. Styregruppemøde den 8. november
3. Omlægning af de selektive høstninger
4. Implementering af DS-484