netarkivet.dk

STATSBIBLIOTEKET Det Kongelige Bibliotek

Nyhedsbrev – oktober 2005

Det er mig en stor glæde og ære at kunne præsentere det første nyhedsbrev for netarkivet.dk Nyhedsbrevet er tænkt som en månedlig status til relevante partnere i organisationen der som bekendt har base i de to biblioteker. Da dette første nyhedsbrev dækker perioden fra 1. juli og frem til oktober 2005 er det sandsynligvis lidt længere end det i fremtiden vil blive.

Den første juli gik netarkivet.dk efter planen i luften med indsamling og bevaring af den danske del af internettet. Et stort udviklingsprojekt er løbet af staben igennem de seneste 18 måneder i et forbilledligt samarbejde mellem SB & KB. Udviklingsprojektet havde frem til 1. juli koncentreret sig om indsamlingsdelen af det samlede system. Denne deadline blev nået og indsamlingen gik i gang på samme tid som den nye revision af pligtafleveringsloven trådte i kraft.

Udviklingsprojektet har i perioden efter 1. juli arbejdet både på den grafiske brugergrænseflade der i det daglige skal betjenes af pligtafleverings-afdelingen på KB og nationalområdet på SB samt på etableringen at den tekniske infrastruktur til bit-bevaring (tjek for og korrektion af korrupte filer).

Udviklingsprojektet slutter officielt 15. oktober og systemet ”afleveres” derefter officielt til drift. Systemet er stort og kompliceret, mere end 20 servere er i spil i det samlede setup. Selvom systemet nu er nået rigtig langt, har det stadig en række både fejl og åbenlyse mangler (blandt andet duplikat-reduktion der forventes at kunne reducere den nødvendige lagerkapacitet med 50%) hvorfor et eller flere nye udviklingsprojekter søsættes i nærmeste fremtid.

Ud over videreudvikling af systemet arbejdes der pt. på planer om at isolere dele af det indsamlede materiale som ikke rummer personfølsomme data og dermed kan gøres tilgængelig for offentligheden på de to biblioteker. Der ligger i øjeblikket en ansøgning hos KUM omkring dette projekt. Desuden er både KB & SB med i EU-ansøgning omkring digital bevaring

Bibliotekerne har hver især udnævnt 2-3 personer i national/pligt-afdelingerne til at varetage den både praktiske og faglige udfordringer i indsamlingsopgaven. Disse personer er siden 1. august blevet oplært i internettets både fortræffeligheder og finurligheder, ligesom systemet i praksis er blevet gennemgået både gennem undervisning og øvelses-opgaver. Sideløbende med det vi kalder produktions-systemet har bibliotekerne et stort set identisk test-system, som ud over at være udviklernes platform også kan bruges til oplæring, øvelser og ”leg” således at de medarbejdere der skal betjene systemet kan få de bedst mulige forudsætninger for at løse opgaven.

Den oprindelige plan for officiel overdragelse af systemet til national-/pligt-afdelingens daglige drift var sammenfaldende med det store udviklingsprojekts deadline 15. oktober. Systemet (herunder især den grafiske grænseflade) har dog ikke været tilgængelig længe nok til at det virker fornuftigt at kaste de biblioteksfaglige medarbejdere ud i arbejdet alene endnu. Planen er derfor nu en glidende overgang over de næste 1 til 2 måneder.

EDB-afdelingen på KB og IT-afdelingen på SB står får den daglige drift af hardware / netværk m.m. Samarbejdet har kørt godt og driftslederen har afholdt møder med begge afdelinger. I nærmere fremtid skal netarkivet til at kigge på krav for ”Trusted Repositories”, dette arbejde vil inddrage EDB- / IT-afdelingerne i det omfang at fx sikkerhedskrav også kræver deres involvering.

Samarbejdet mellem udviklingsafdelingerne har krævet ugentlige telefonmøder samt jævnlige fysiske møder. Samarbejdet mellem de øvrige afdelinger og driftslederen kræver ligeledes hyppig kontakt. Meget kan klares via telefon og e-mail, men der arbejdes pt. med at etablere bedre virtuelle mødefaciliteter på begge biblioteker i form af blandt andet mulighed for delte applikationer på såkaldte smartBoards samt videokonference.

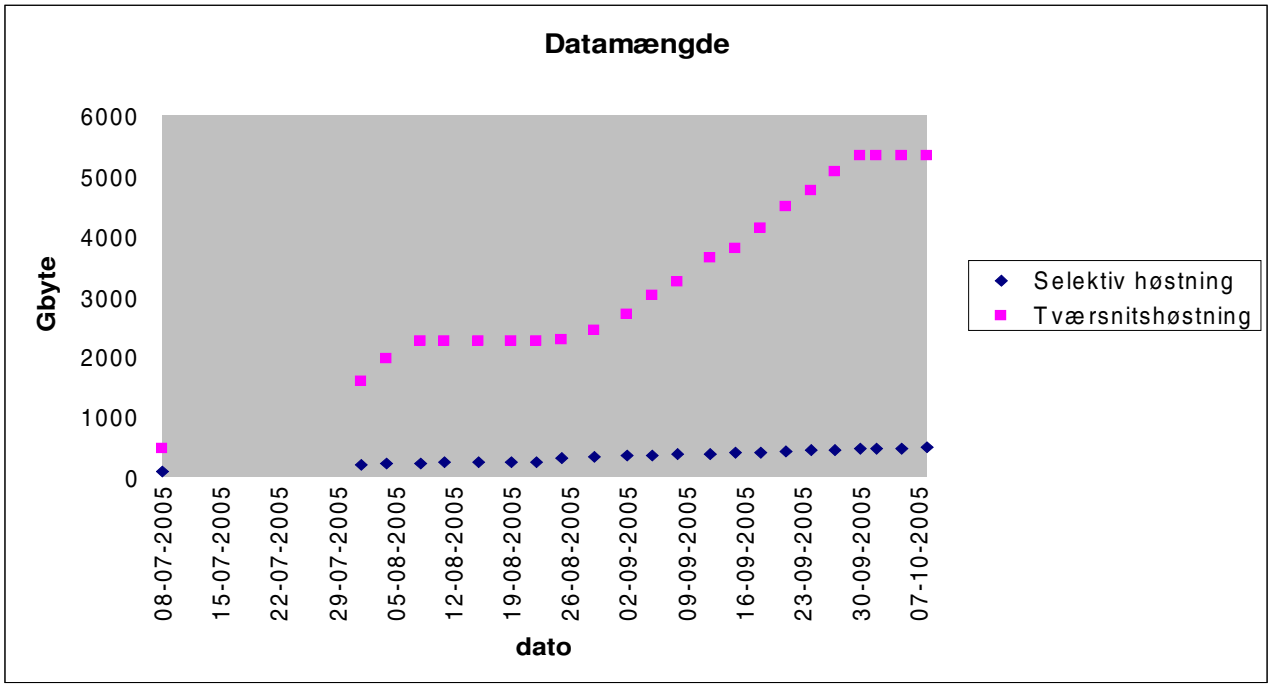
Med hensyn til indsamling af materialet så går det rigtig godt. Den første såkaldte tværsnitshøstning der skal høste alle kendte danske domæner – hvilket pt. kun dækker domæner under .dk, er afsluttet med rimelig succes. For at optimere processen omkring tværsnitshøstningen blev den afviklet i en række iterationer. Vi startede med en høstning der definerede en maksimal grænse på antallet af objekter fra hvert domæne på 10 objekter. Det er naturligvis en meget lille grænse, men det viser sig alligevel, at mere end 350.000 ud af godt 600.000 domæner ikke når den grænse. En stor del (ca. 120.000) svarer slet ikke på DNS-opslag hvilket i praksis betyder at domænenavnet kun er registreret hos DK-hostmaster men ikke taget i brug, resten er bare meget små netsteder.

Næste høstning kørte så forfra med de domæner der faktisk nåede grænsen på 10 objekter (ca. 235.000), denne gang med en grænse på 50 objekter pr. domæne. Igen blev en stor del af domænerne sorteret fra (i praksis nåede de ikke grænsen på 50). Samme proces blev gentaget yderlige 2 gange med grænser på henholdsvis 500 og 5000 objekter pr. domæne. Ved afslutningen af den sidste høstning var der ca. 3500 domæner tilbage der endnu ikke var høstet komplet. Det betyder i praksis at netarkivet således har indsamlet mere end 99% af dk-domænerne så komplet som den anvendte høster-teknologi kan gøre det.

Cirka 40 netsteder er siden 1. juli blevet fuldt tættere med såkaldte selektive høstninger. Netarkivet har indsamlet materiale fra disse helt ned til 6 gange dagligt (fx forsider på nyheds-sites).

Netarkivets første begivenhedssamling koncentrerer sig om det forestående kommunale valg 15. november 2005. Kulturminister Brian Mikkelsen igangsatte officielt denne høstning ved et arrangement på Det Kongelige Bibliotek 4. oktober. Det forløb uden problemer og med nogen pressedækning.

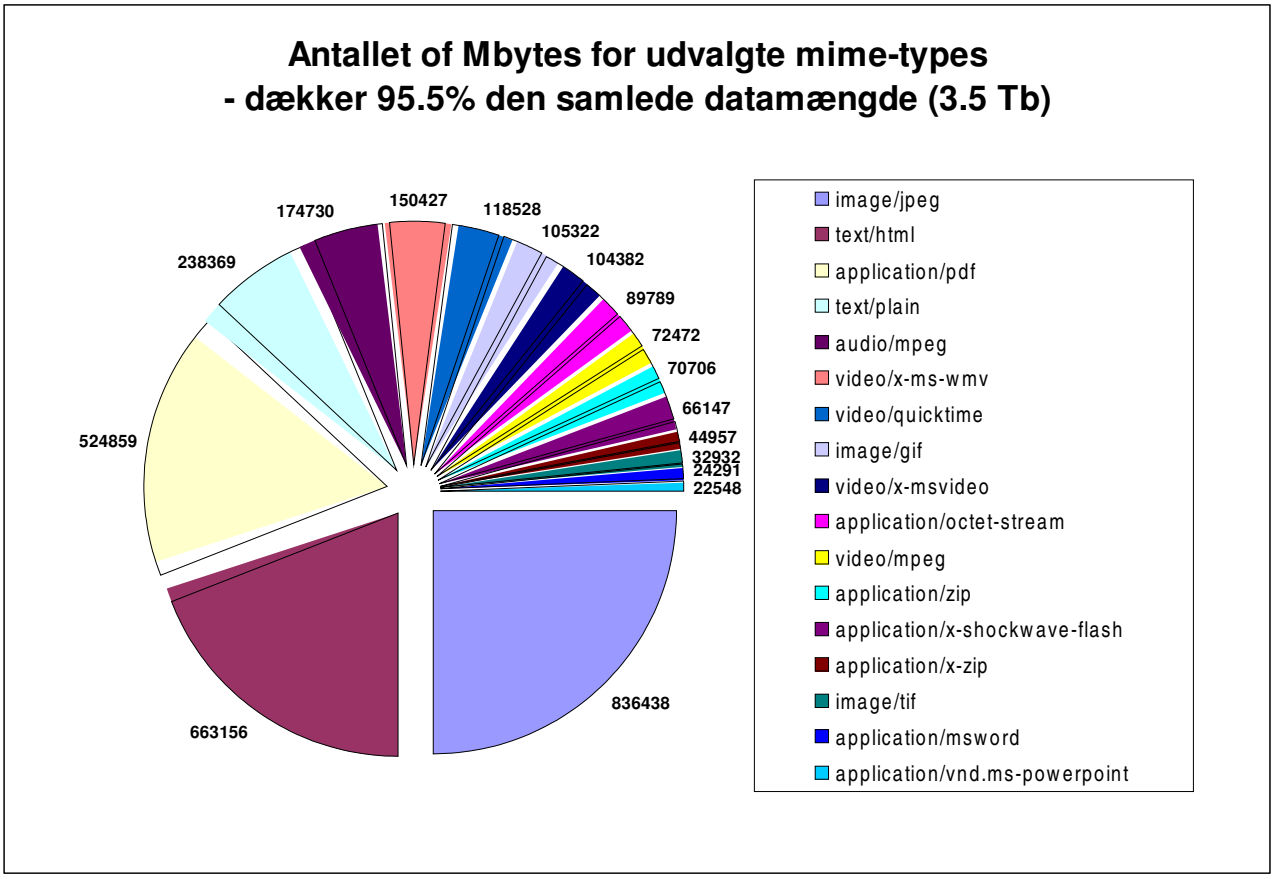
Følgende graf viser den samlede datamængde indsamlet frem til starten af oktober 2005 for henholdsvis tværsnitshøstning og selektiv-høstning.



Arkivet rummer således pt ca. 5.8 Tbyte data. Den mængde repræsenteres af ca. 120 millioner objekter.

I midten af september lavede vi en analyse af de indsamlede data for at undersøge hvilke filtyper der primært var repræsenteret. Undersøgelsen dækker kun data fra tværsnitshøstningen og kan være påvirket af den måde der høstes på i de førnævnte iterationer.
Mest bemærkelsesværdigt er 2 ting. JPEG-billeder ligger på en klar første plads over HTML-filer. Dette er nyt i forhold til tidligere analyser / undersøgelser. Eneste fornuftige forklaring er, at danskerne nu har taget digitalkameraerne til sig og også er begyndt at offentliggøre billederne via internettet. Den anden bemærkelsesværdige ting er, at PDF-filer kommer ind på en 3. plads når man måler på datamængde. Det må dels være et tegn på at PDF-formatet til stadighed vokser i udbredelsen men også, at PDF-filerne bliver stadigt større. Den gennemsnitlige filstørrelse for en PDF-filer ligger pt. på 541 Kb. Dette er en naturlig udvikling eftersom vores internetforbindelser i dag er nogle gange hurtigere (og billigere) end for bare 5 år siden.

Det er også værd at bemærke, at kun 17 mime-types dækker mere end 95% af datamængden.



Planerne for den nærmeste fremtid rummer følgende:

1. gennemførelse af begivenhedshøstningen af det kommunale valg i november
2. igangsættelse af den næste tværsnitshøstning
3. igangsættelse af et eller flere nye udviklingsprojekter
    a. ”færdiggørelse” af systemet
    b. Nyt system til håndtering af ikke-personfølsomme data
    c. EU-projekt(er) omkring digital bevaring
4. etablering af virtuelle mødefaciliteter
5. endelig overdragelse af det daglige arbejde med systemet til national-/pligt-afdelingerne
6. nyt website til www.netarkivet.dk
7. krav til ”Trusted Repositories”

Med venlig hilsen, tak for godt samarbejde i det forgangne og med håb om stadig godt samarbejde fremover.

Bjarne Andersen
driftsleder netarkivet.dk
bja@netarkivet.dk