netarkivet.dk

Nyhedsbrev marts 2011

Velkommen til Netarkivets forårs nyhedsbrev.

Status på hardware og drift

Hardwaren på Det Kongelige Bibliotek opfører sig forbilledligt. To nye høstermaskiner og en ny bitarkivmaskine er idriftsat fuldstændigt uproblematisk. En accessmaskine, der også har nået pensionsalderen, bliver udskiftet med nyindkøbt maskine en af de nærmeste uger.

Driften af tilsvarende bitarkivmaskiner på Statsbiblioteket er også tilfredsstillende. Dog var 3 af de 5 høstermaskiner blevet så gamle (5-6 år), at de nu er udskiftet med 2 nye og meget kraftigere maskiner. Alle disse maskiner er opgraderet til nyeste version af operativsystem, java og ftp server.

Den samlede datamængde er nu ca. 180 TB. Yderligere lagerplads til resten af året er budgetmæssigt på plads.

Gamle data

Der arbejdes fortsat med konvertering af gamle samlinger fra før år 2005, så de kan blive indlemmet i arkivet.

Afprøvning af procedurer

Den centrale administrationsmaskine i Netarkivet, kb-prod-adm-001, er blevet udskiftet med ny maskine helt fra scratch ud fra en båndkopi af det software, der lå på den. I andet forsøg lykkes det helt som ønsket og proceduren, herunder dokumentationen, er nu opdateret. Skulle katastrofen indtræffe hvor denne centrale maskine bryder sammen, vil vi kunne være kørende igen på under en dag.

Status på Wayback adgang til arkivet

Løsningen, der skal køre på en server på Det Kongelige Bibliotek, er nu så langt at der kan vises bitarkiv indhold herfra. Der er etableret løsning gennem en VPN (Virtual Private Network) server, så sikkerheden er hermed på plads. Der er stadig nogle få hurdler inden Netarkivets testere kan sig OK til at give vores eksterne brugere adgang. Blandt andet skal de løbende indekseringer af nyt bitarkiv indhold afvikles automatisk og flettes ind i det eksisterende indeks. Der forventes adgang til eksterne brugere i foråret.

Tværsnitshøstning

Den 10. tværsnitshøstning sluttede d. 31. januar 2011. Straks herefter påbegyndtes produktions opgradering af NetarchiveSuite til stable release 3.14. Dette har imidlertid drillet. I denne release er inkluderet det første store bidrag fra BNF (Bibliothèque nationale de France), og der skulle justeres på forskellige opsætningsparametre og rettes en enkelt fejl, før driften blev stabiliseret.

Desuden har vi sloges med en fejl, hvor bit applikationer gik i stå efter ftp kommunikation til SB. Det er nu lykkes os at slippe af med fejlen, efter at SB's høstermaskiner er opgraderet til nyeste operativsystem version, java version og ftp server.

Tværsnitshøstning nr. 11 er startet d. 9. marts. Denne tværsnitshøstning et sat op til at køre med generel respekt for robots.txt. Robots.txt er en fil under hvert domæne med instruktioner til forbipasserende webcrawlere, fx http://www.statsbiblioteket.dk/robots.txt. Her kan det enkelte netsteds administrator give maskinlæsbare instrukser om hvad, der er tilladt at crawle, og hvad der er

forbudt. Netarkivet har ikke tidligere respekteret tobots.txt, da testhøstninger viste, at vi i så fald fik høstet for lidt til vores formål. Men nettet modnes, og det er nu blevet tid til at prøve dette på en tværsnitshøstning og bagefter evaluere, om vi stadigt høster tilstrækkeligt og relevant. Der har været en del henvendelser fra webejere, der gerne vil have, at vi respekterer dette. Det forventes en betydeligt reduceret høstningsperiode og også reducerede datamængder. Forhåbentligt viser de manglende datamængder sig at være ikke-relevant materiale.

Selektive høstninger

De selektive høstninger har nogle udfordringer, da webstederne ændrer sig hele tiden og de gentagne høstninger ikke nødvendigvis følger med. Derfor blev kvaliteten af nogle af de selektive høstninger ikke tilfredsstillende. Udfordringerne med svært høstbart webindhold er nu en fast regelmæssig opgave for udviklerne sammen med samlingsmedarbejderne. Dvs. at der nu opbygges såkaldt webcrawler engineer-kompetence i projektet (fx høstning med login, video, lyd, nye teknologier). Bagsiden af medaljen er, at det tager tid fra andre opgaver.

Begivenhedshøstninger

Der er ikke indtruffet så mange begivenheder i 2010/2011, der har været tilstrækkeligt interessante og relevante til begivenhedshøstninger. Af tidligere erfaring ved vi, at begivenheder ofte kommer i klumper, så forhåbentligt kommer der gode kandidater i år. Under alle omstændigheder kommer der et folketingsvalg.

Når en begivenhed er ved at indtræffe, iværksætter samlingsmedarbejderne på Det kongelige Bibliotek og Statsbiblioteket indsamling af relevante URL'er og igangsætter derefter gentagne høstninger af dem, indtil begivenheden er slut.

Internationalt samarbejde og open source NetarchiveSuite

Som nævnt har BNF bidraget med en betydelig mængde ny funktionalitet. Det bekræfter formålet med at lægge NetarchiveSuite ud som open source projekt og finde partnerskaber hos andre nationalbiblioteker. Nu begynder gevinsten af samarbejdet at komme. Den franske kode er også nyttig for danske brugere.

Redaktionsgruppemøde

Redaktionsgruppe har sit tredje møde med de nuværende medlemmer d. 22. november 2010. Næste møde skal afholdes på Det kongelige Bibliotek d. 25. marts 2011.

Netarkivet er glade for gruppens råd og anvisninger som så vidt muligt følges.

Organisationsændring i Netarkivet

Efter sparerunde ændres Netarkivets organisation. Christen Hedegaard på KB giver sine projektleder opgaver til udvikler Mikis Sørensen på SB. Mikis overtager også Tue Larsens releasetest opgaver. Fremover deles driftslederens opgaver ud på Tue Larsen, der med de tekniske opgaver får rolle som driftskoordinator, mens samlingsafdelingerne overtager de administrative og udadvendte funktioner. Det samlede antal årsværk reduceres fra 5,0 til 4,8.

Det betyder således, at undertegnede forlader projektet og jobbet pr. 30. april 2011. Jeg vil derfor gerne sige tak for 3 interessante og også udfordrende år i Netarkivet. Det har været et meget bredt sammensat job, som det har taget tid at komme ind i. Derfor en helt speciel tak til Bjarne Andersen, der har bidraget med meget sidemandsoplæring og råd. Også tak for et godt samarbejde til de af jer, som jeg ikke får sagt personligt farvel til.

Claus Lomborg, Netarkivet
clo@netarkivet.dk