netarkivet.dk

Nyhedsbrev juli 2008

Velkommen til Netarkivets nyhedsbrev for juli 2008

Som Netarkivets afgående driftsleder vil jeg gerne starte med at beklage, at dette nyhedsbrev er blevet forsinket. Det har været en meget travl periode, hvor tiden især er blevet brugt på overdragelse af det daglige ansvar for driften af Netarkivet til en ny projektkoordinator. Det er stadig planen, at nyhedsbrevet skal udkomme ca. fire gange årligt.

Dette bliver dermed det sidste nyhedsbrev fra min hånd, og i den forbindelse vil jeg gerne benytte lejligheden til at takke alle samarbejdspartnere såvel på som uden for de to nationalbiblioteker. Det har været tre meget spændende år, og det er med både stolthed og vemod, jeg nu slipper den daglige drift af et projekt, jeg har været med til at udvikle gennem de sidste mange år.

Ny projektkoordinator for Netarkivet

Netarkivet ansatte pr. 1. maj 2008 en ny projektkoordinator til at afløse Bjarne Andersen, som har været driftsleder gennem de første tre år af Netarkivets levetid. Claus Lomborg startede på 20 timer pr. uge, men takket været en ekstra bevilling fra Kulturministeriet kan stillingen nu opnormeres, så Claus arbejder 30 timer pr. uge fra 1. september 2008. Det burde give tid til at gøre mere ved bl.a. de kommunikerende aktiviteter i projektet.

Claus præsenterer sig selv på følgende måde:
"Jeg hedder Claus Lomborg, er 51 år og bor med min kone og vores tre børn i Solbjerg lige syd for Århus. Jeg har været i it-branchen i snart 20 år og derved prøvet lidt af hvert. Det meste af tiden har jeg været udvikler, men de seneste år har jeg arbejdet mere og mere med andre ting, såsom projektkoordinering, teknisk it-projektledelse, medicinske databaser, it-analyser, og som webmaster og kundevendt konsulent mm."

Jeg vil gerne byde Claus velkommen i projektet. Han har formelt overtaget den daglige styring fra 1. juli 2008.

Jeg fortsætter indtil videre i Technical Committee for IIPC (International Internet Preservation Consortium) og indtræder sandsynligvis i Netarkivets styregruppe i stedet for udviklingschef Birte Christensen-Dalsgaard, der forlader Statsbiblioteket til oktober 2008.

Femte tværsnitshøstning

Netarkivets femte tværsnitshøstning blev startet 17. juli 2007. De to skridt (først op til 10 Mb pr. domæne for at færdiggøre de mindste domæner (næsten 90% af de danske domæner er mindre end 10Mb) og derefter op til 2 Gb pr. domæne for de domæner, der ramte 10 Mb grænsen) blev afsluttet 9. januar 2008. Perioden på næsten seks måneder er klart for lang til at sikre fire årlige tværsnitshøstninger. Derfor er der inden det sidste skridt af den sjette tværsnitshøstning arbejdet med en omkonfigurering af høsteren. Tests har vist, at dette kan give op imod en fordobling af høstningshastigheden, og nu må praksis vise, om de tal holder stik. Samtidig har styregruppen besluttet at tage ekstra høstermaskiner i brug. Netarkivet indkøbte i december 2007 fem nye servere, der skal bruges til adgang til arkivet, men da der ikke har været den nødvendige udviklingstid, er disse maskiner endnu ikke taget i brug. De inddrages

derfor midlertidigt som høstermaskiner, hvilket betyder en fordobling af den rå serverkraft fra fem til ti maskiner.

Den femte tværsnitshøstning fyldte 11.6 Tb efter duplikatreduktion. Mange netsteder (omkring 6000) ramte standardgrænsen på 500 Mb pr. domæne. Disse domæner skal gennemgås manuelt for evt. at hæve grænsen. På grund af de store og voksende mængder har styregruppen besluttet at hæve standardgrænsen til 1 Gb pr. domæne. Det skulle gerne bringe antallet af domæner, der rammer grænsen og dermed skal tjekkes manuelt, ned på under det halve.

Den femte tværsnitshøstning afslørede også en stigende tendens til, at der arkiveres materiale fra andre domæner end de, det var tiltænkt. Det er et tydeligt resultat af Web 2.0 tendensen, der eksploderer i disse år, og som gør, at flere og flere netsteder indlejrer materiale fra andre netsteder (fx videomateriale fra youtube.com). Denne tendens skal følges, og stiger mængden af indlejrede materialer fra "fremmede" domæner for kraftigt, kan det komme på tale at ændre Netarkivets domæne-grænse-logik således, at det indlejrede materiale tæller med i datamængden for det domæne, hvori materialet er indlejret, i modsætning til den nuværende logik, hvor indlejret materiale tæller med under domænet, hvorfra det hentes.

Sjette tværsnitshøstning

Efter at have ligget stille i godt to måneder, mens der blev arbejdet med omkonfigurering af høstersoftwaren, og mens der blev rettet en alvorlig hukommelsesfejl i NetarchiveSuite, er endnu en tværsnitshøstningen sat i gang.

Netarkivets sjette tværsnitshøstning blev igangsat 31. marts 2008. Første skridt (op til 10 Mb pr. domæne) blev afsluttet 20. maj 2008. Der var 99.330 domæner, der ramte 10 Mb grænsen (mod ca. 95.000 domæner ved forrige tværsnitshøstning). For disse domæner er nu igangsat næste skridt af tværsnitshøstningen med en standardgrænse på 1Gb for alle domæner og 2 eller 4 Gb for udvalgte og manuelt godkendte domæner, der tidligere har ramte grænsen på 1 eller 2 Gb

Selektive høstninger

De selektive høstninger har nået det oprindelige mål på 80 netsteder og arkiverer nu løbende (fra seks gange dagligt til månedligt) 82 udvalgte netsteder. Arbejdet koncentreres herefter om løbende kvalitetskontrol af det indsamlede samt fortsat ajourføring af puljen, idet nye interessante netsteder hele tiden ser dagens lys.

Datamængden for de selektive høstninger er svagt stigende. I øjeblikket indsamles med en hastighed på omkring 1 Tb mere pr. år end budgetteret. Styregruppen har besluttet, at den nuværende indsamling fortsætter, idet det vurderes, at arbejdsindsatsen for at reducere den datamængde, der indsamles, vil være dyrere end det, det ekstra forbrug af lagerplads koster.

Begivenhedshøstninger

Krisen omkring tegningerne af Muhammed blussede op igen, hvilket medførte en mindre, ekstra indsamling af en stor del af de netsteder, der blev indsamlet, da krisen oprindeligt var på sit højeste, samt et mindre antal nye netsteder.
Styregruppen har i samråd med Redaktionsgruppen besluttet, at OL 2008 kun tages som begivenhedshøstning, hvis der kommer en mere regulær politisk krise i forbindelse med begivenheden. Indtil videre dækkes OL af de eksisterende medier og Netarkivets selektive høstninger samt tværsnitshøstninger. Der er dog lavet en mindre, ekstraordinær indsamling af en række officielle OL-netsteder, således at forhistorien er med, hvis en krise skulle opstå.

Status på bitarkiverne

Bitarkiverne rummer nu følgende datamængder:

- Tværsnitshøstning: 55988 Gb

- Selektiv høstning: 8825 Gb
- Begivenhedshøstning: 5895 Gb

Totalt har Netarkivet altså indsamlet og arkiveret næsten 71 Tb data.
På Netarkivets lokation på Det Kongelige Bibliotek (KB) består bitarkivet af en stor pulje almindelige pc'er og mindre servere. Dette har været et helt bevidst valg af arkitektur helt fra projektets begyndelse, og det betyder, at der er masser af processorkraft til rådighed til fx indeksering af hele arkivet. I praksis består Netarkivets bitarkiv-installation på KB i øjeblikket af 56 fysiske maskiner. En del af disse maskiner giver problemer, og antallet af maskiner får nu sammen med stigningen i datamængden Netarkivet til at overveje at skifte til en anden teknisk platform. Præmissen for dette platformskifte skal være, at det oprindelige mål med megen processorkraft skal bevares.

IIPC

Netarkivet spiller stadig en meget aktiv rolle i IIPC (International Internet Preservation Consortium). Bjarne Andersen er fortsat medlem af det indholdsmæssigt styrende organ: Technical Committee.

IIPC afholder det næste møde i Århus i forbindelse med ECDL2008 som Statsbiblioteket i år arrangerer. Dette, sammen med den årlige internationale webarkiverings workshop (IWAW), som også afholdes i Århus i samme uge, betyder, at ugen 15. - 20. september kommer til at stå i webarkiveringens tegn, og at stort set alle verdens eksperter på dette felt samles i Danmark.

NetarchiveSuite

NetarchiveSuite, som Netarkivets komplette system til arkivering af internettet blev døbt i forbindelse med udgivelsen af systemet i open source, er nu aktivt i brug på flere europæiske institutioner.

Det østrigske og det skotske nationalbibliotek bruger systemet til daglig indsamling. Flere andre institutioner tester systemet seriøst, og det lader til, at flere af disse institutioner ender med at vælge Netarchive-Suite som teknisk løsning. En af de mest interessante kandidater til et egentligt udviklingssamarbejde er det franske nationalbibliotek (BnF), som efter længere tids test af flere forskellige systemer, ser ud til at ende med at basere deres nationale løsning på NetarchiveSuite.

Redaktionsgruppemøde

Netarkivets redaktion holdt møde 17. januar 2008 på Det Kongelige Bibliotek.
Redaktionsgruppen er, som den har været hele vejen igennem, et meget nyttigt organ til rådgivning og vejledning om Netarkivets aktiviteter.
Redaktionsgruppen har haft fokus på dokumentation af Netarkivets indsamlinger, og det har indtil videre ledt til forskellige oversigter over selektive høstninger og begivenhedshøstninger, der kan ses på http://www.netarkivet.dk
Redaktionsgruppen er formelt nedsat til 2009, hvorefter Kulturministeriet skal se på, hvordan organet bemandes i en ny periode.

Tværsnitshøstning for det Grønlandske Landsbibliotek

Netarkivet har fået en henvendelse fra det Grønlandske Landsbibliotek (Nunatta Atuagaateqarfia) om muligheden for, at Netarkivet kan foretage tværsnitshøstninger af de grønlandske domæner som en service over for Grønland. Netarkivet har taget positivt imod denne henvendelse og arbejder i øjeblikket på en beskrivelse af et pilotprojekt, der kunne løse denne opgave.

Styregruppen har besluttet at Føroya Landsbókasavn skal tilbydes at være med i pilotprojektet

på samme vilkår som det grønlandske bibliotek. Som parallel aktivitet skal pilotprojektet begynde at kigge på udvikling af forretningsmodeller for Netarkivet.

Alvorlig fejl i Netarkivets indekseringsmekanisme

I sidste nyhedsbrev blev omtalt en alvorlig fejl i Netarkivets indekseringsmekanisme. Fejlen er nu løst, og både høstninger og adgang til arkivet skulle igen fungere normalt – også for meget store høstninger (tværsnitshøstninger).

Det var ikke første gang, de meget store datamængder, som genereres ved de store høstninger, har afstedkommet problemer, og det synliggjorde blot endnu engang vigtigheden af at teste softwaren så grundigt som muligt uden faktisk at foretage en høstning i fuld skala, inden softwaren frigives og installeres på Netarkivets systemer.

Bemanding af Netarkivets udviklingsfunktion

Lars Clausen valgte desværre at forlade Statsbiblioteket og dermed projektet pr. 31. marts 2008. Lars var en af pionererne; han havde været med fra 2003 og dermed været med til at udvikle både den nye pligtafleveringslov og i særdeleshed grundstenene i hele NetarchiveSuite systemet.

Lars' fratræden kombineret med almindelig travlhed har betydet, at udviklingsfunktionen i Netarkivet har kørt på lavere blus end normalt og vil fortsætte med det indtil slutningen af september. Herefter skrues op for blusset, og udviklingen kommer tilbage på det vanlige niveau på omkring 1.6 årsværk fordelt på de to biblioteker.

Planer for den kommende periode

1. Afvikling af ECDL, IWAW og IIPC turorial/workshop/møder i Århus i September
2. Resterende overdragelse af den daglige drift til Claus Lomborg

De bedste sommerhilsner og endnu en gang tak
Bjarne Andersen