netarkivet.dk

Nyhedsbrev december 2008

Velkommen til Netarkivets nyhedsbrev.

Dette er mit første nyhedsbrev som Netarkivets nye projektkoordinator. På grund af travlhed med overdragelse af det daglige ansvar for Netarkivet og på grund af uregelmæssig drift af arkivet, har nyhedsbrevet været forsinket. Men nu er det her.

Tak til Bjarne Andersen for tålmodig overdragelse. Projektkoordinator/projektleder-jobbet er mere omfangsrigt og sammensat end først antaget. Arbejdet omfatter alt fra driftsovervågning med kørsel af diverse scripts på linuxmaskinerne – over koordinering af PLIGT-, drifts- og udviklingsafdelinger på både Det Kongelige Bibliotek og på Statsbiblioteket – til daglige henvendelser udefra.

Styregruppen for Netarkivet

Bjarne fortsætter i Technical Committee for IIPC (International Internet Preservation Consortium) og er indtrådt i Netarkivets styregruppe i stedet for udviklingschef Birte Christensen-Dalsgaard, der forlod Statsbiblioteket til oktober 2008 til fordel for en stilling som vicedirektør for IT-området på Det Kongelige Bibliotek.
Arne Sørensen, Statsbiblioteket, har overladt sin plads i styregruppen til sektionsleder for IT-drift på Statsbiblioteket Klaus Kjærgaard.

Sjette tværsnitshøstning

Netarkivets sjette tværsnitshøstning blev igangsat 31. marts 2008. For disse domæner blev næste skridt af tværsnitshøstningen startet 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.

Denne tværsnitshøstning forventes at være færdig omkring nytår. Hermed er den noget forsinket. Det skyldes omkonfigurering af høstersoftwaren, hukommelsesfejl i NetarchiveSuite og driftsforhold omkring PC-farmen. De første to problemer er løst. Det sidste problem arbejde driftsafdelingen på Det Kongelige Bibliotek på at løse ved en gennemgribende ny hardwarearkitektur. Arkitekturen på Statsbiblioteket bliver ikke ændret, da den ikke har samme problemstillinger. Resultatet er dog at vi i år ikke kommer i nærheden af de 4 årlige tværsnitshøstninger som er vores mål. Men når vi har etableret den ønskede arkitektur på Det Kongelige Bibliotek vil mulighederne være bedre fremover og ikke mindst vil driftstabiliteten forhåbentligt være meget højere.

Syvende tværsnitshøstning

Næste tværsnitshøstning ventes påbegyndt straks den nuværende er færdig - dvs i januar 2009. Her vil DR formentligt være med og lave et TV-indslag om Netarkivet.

Selektive høstninger

De selektive høstninger 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

Der er startet en særskilt høstning omkring den finansielle/økonomiske krise. Emnet er allerede dækket af nyhedsmedierne men det blev vurderet at de væsentligste finansielle websites skulle høstes. Det er netsteder som banker, kreditforeninger, aktiehandel-sites, investerings-sites, institutionelle sites (fx finansrådet, finanstilsynet), egentlige finansnyheder osv. Desuden er der et specielt fokus på virksomheder, der er ved at dreje nøglen om i forbindelse med krisen. Det blev også undersøgt, om der var specielt væsentlige steder omkring de usikre islandske investeringer i Danmark. Der bliver udvalgt nye domæner i takt med at den finansielle krise udvikler sig til en økonomisk krise.

På et tidspunkt kom der også rystelser i Det radikale Venstre som dog ikke udviklede sig til nogen stor politisk ting. Det, der var, blev høstet.

Som altid gælder at alle, der er involveret i Netarkivet, bidrager med nye ideer til begivenhedshøstninger, fordi det er så væsentligt at indfange begivenhederne tidligst muligt.

Status på bitarkiverne

Bitarkiverne rummer nu (tal fra 13/11/2008) følgende datamængder:

- Tværsnitshøstning: 63417 Gb
- Selektiv høstning: 9766 Gb
- Begivenhedshøstning: 6044 Gb

Totalt har Netarkivet altså indsamlet og arkiveret 79,2 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 hardware-problemer, og antallet af maskiner har sammen med stigningen i datamængden fået Netarkivet til at vælge at skifte teknisk platform. Præmissen for dette platformskifte er, at det oprindelige mål med megen processorkraft skal bevares. Der arbejdes pt. på det, og der laves mange test på nogle nyindkøbte maskiner. Specielt har det vist sig at harddiske fås i forskellige kvaliteter – og priser. Fremover vælges en bedre kvalitet, da pengene spares ind på færre driftsforstyrrelser.

PINDAR, et søster projekt

Colin Samuel Rosenthal, der er trådt ind i gruppen af udviklere på Netarkivet på Statsbiblioteket, har den foregående tid arbejdet på PINDAR projektet.

PINDAR er et fællesprojekt for KB og SB på linje med Netarkivet. Formålet er at indsamle forskningsmateriale fra landets universiteter og andre forskningssteder. Materialet (artikler, papers mv.) er afleveringspligtigt, og datamængderne er små - typisk PDF-filer. PINDAR er færdigudviklet og i releasetest. PINDAR består af en høstningsdel a la Netarkivet plus en simpel webdel hvormed der kan igangsættes og skeduleres høstninger af universiteternes institutional repositories. Der er ikke nogen direkte sammenhæng mellem PINDAR og Netarkivet, men logisk, driftsmæssigt og indholdsmæssigt er der lighedspunkter. PINDAR bruger fx samme arkivmodul (del af NetarchiveSuite) som Netarkivet. Netarkivets projektkoordinator er således også koordinator for dette projekt når det sættes i endelig drift inden jul.

IIPC

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

Afvikling af ECDL, IWAW og IIPC turorial/workshop/møder i Århus i September

IIPC blev afholdt i Århus i forbindelse med ECDL2008, http://www.ecdl2008.eu/, som Statsbiblioteket som arrangør. Samtidigt blev den årlige internationale webarkiverings workshop (IWAW) afholdt samme sted. Ugen 15. - 20. september stod således i webarkiveringens tegn, da verdens eksperter på dette felt var samlet.

Den næste ECDL-konference afholdes på Korfu, Grækenland fra 27.sept. til 2. okt. 2009. Yderligere information på http://www.ecdl2009.eu/

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. 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 3. oktober 2008 på Statsbiblioteket.

Redaktionsgruppen pointerer at få dokumenteret den viden vi har, i arkivet. Det er vigtigt for fremtidige brugere at kende manglerne i arkivet, og den viden går tabt, hvis vi ikke nedfælder den. For tværsnitshøstningernes vedkommende kan kendte mangler fx dokumenteres i forbindelse med, der er blevet foretaget en tværsnitshøstning. For de selektive siders vedkommende kan det være relevant at dokumentere på domæne-niveau. Det er vigtigt at få overblik over, hvad vi har af dokumentation og at få dokumentationen sat i system og samlet på én eller så få steder som muligt.

Oversigter over selektive høstninger og begivenhedshøstninger kan ses på http://netarkivet.dk/indsamlingsDoku.html

Kulturministeriet skal bemande redaktionsgruppen på ny i 2009.

Andre mulige tværsnitshøstninger

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 og sætte en pris.

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.

Slovenien har spurgt til muligheden for webarkiveringsassistance. Det er i styregruppen blevet besluttet ikke at gå videre med det på Netarkiv-niveau. Det kongelige Bibliotek overtager sagen alene.

Bemanding af Netarkivets udviklingsfunktion

Kåre Fiedler Christiansen er på barsel i 4½ måned og vender derefter tilbage.

I mellemtiden er udvikler Colin Rosendahl blevet tilknyttet Netarkivet hvor han blandt andet skal arbejde med Wayback. Colin erstatter Lars Clausen som forlod Statsbiblioteket i foråret

Det kongelige Bibliotek har pr. 1.12.2008 ansat en ny softwareudvikler. Jonas er 24 år og nyuddannet. Vi ser frem til et godt samarbejde ham.

Planer for den kommende periode

1. Etablering af ny hardwarearkitektur på Det kongelige Bibliotek.
2. Implementering af Wayback access-modul
3. Begivenhedshøstninger: minihøstning af OL og klimatopmøde 2009

Mange julehilsner
Claus Lomborg, projektkoordinator