



Nyhedsbrev Maj 2010

Velkommen til Netarkivets forårs nyhedsbrev.

Status på bitarkiverne

De 12 nye, kraftige HP bitarkiv-maskiner på Det Kongelige Bibliotek giver som forventet en stabil drift. Data er migreret fra de 56 gamle maskiner på Det Kongelige Bibliotek og efterfølgende kontrol fortæller, at der ikke mangler filer hverken på Det kongelige Bibliotek eller på Statsbiblioteket.

Det allersidste tjek, den såkaldte checksum kørsel, har været ramt af en memoryfejl. Men nu er det ved at blive kørt på både kopien på Det Kongelige Bibliotek og på kopien på Statsbiblioteket. Kørslen tager et par uger og er snart færdig. Indtil videre ser det problemfrit ud. Dermed vil den store migreringsopgave være helt afsluttet.

Datamængden i Netarkivet er nu nået op på samlet 146 TB. Hver tværsnitshøstning fylder ca. 26 TB og dette tal er stigende med ca. 10 % per tværsnitshøstning. Derudover høstes hvert år ca. 6 TB for selektive høstninger og begivenhedshøstninger tilsammen. Endelig arbejdes på at konvertere gamle samlinger fra før år 2005 ind i Netarkivet.

Hardware status

Den centrale administrationsmaskine i Netarkivet, "kb-prod-adm-001", er nu blevet udskiftet med en ny kraftig efterfølger. Den fungerer fint fra dag ét og har givet meget bedre svartider.

Nu er Netarkivets hardware pænt opdateret på bitarkiv delen og på den administrative maskine.

Niende tværsnitshøstning

Denne tværsnitshøstning er startet d. 3. feb. 2010 efter nogle drillerier med opgradering af NetarchiveSuite til version 3.10.0., nogle konfigurationsproblemer og den central maskine (kb-prod-adm-001) som bare ikke kunne mere og måtte udskiftes. Herefter kører det tværsnitshøstningen pænt. Tværsnitshøstningen er planlagt til at køre over cirka 4 måneder.

Høstningskapaciteten på netværket på Det kongelige Bibliotek er sat op med yderligere 2 tværsnitshøster applikationer. Men samtidigt er der begyndt at vise sig problemer med 3 af de ældste høstermaskiner, hvor høstningerne går i fejl. Der viste sig derfor nødvendigt at slukke for nogle høsterapplikationer, så en stor del af den forøgede høstningskapacitet som de ekstra opstartede høsterapplikationer gav, er væk igen.

En Apache Derby databasefejl har drillet og været meget svær at få ryddet af vejen. Men det er endeligt lykkes nu. Det har medført et ønske om at skifte til en anden database, fx PostgreSQL, som Netarkivets udviklings samarbejdspartnere bruger.

Selektive høstninger

De selektive høstninger af udvalgte netsteder kører planmæssigt. Som netop nævnt har det på det seneste vist sig at de ældste høstermaskiner på Statsbiblioteket ikke længere er tilstrækkelige til at

trække både tværsnitshøstninger og selektive høstninger på samme tid. Der er nu ændret i konfigurationen, så tingene igen kører tilfredsstillende.

Svært høstbart webindhold er nu blevet en fast, prioriteret opgave på linje med udviklernes øvrige opgaver. Udviklergruppen assisterer samlingsfolkene med at høste de komplekse websites med for eksempel login adgang eller med video indhold.

Begivenhedshøstninger

Begivenhedshøstninger i 2010 er endnu ikke aftalt, da der ikke har været interessante kandidater. Seneste begivenhedshøstninger var klimatopmøde i København og Vinter OL. Jordskælvet i Haiti er der dog lavet høstning på.

Det er relativt små datamængder, der høstes.

Internationalt samarbejde og konferencer

Netarkivet deltog i 2010 IIPC (International Internet Preservation Consortium) General Assembly i Singapore i 3 – 7. maj.

Blandt de mange nye initiativer her var de måske mest interessante emner:

- Heritrix version 3 som Netarkivet på et tidspunkt kommer til at skulle skifte til
- Automatisk kvalitetskontrol af høstninger som sikrer en højre kvalitet af det indhøstede materiale.
- Rich Media Crawler til høstning af komplicerede websites baseret på fx Flash

Desuden er der aftalt en workshop for NetarchiveSuite udviklingen som nævnt i næste afsnit.

Open source NetarchiveSuite

Det internationale samarbejde omkring udvikling af NetarchiveSuite er nu blevet hverdag. Der er regelmæssigt telefonmøder med deltagelse af udviklere/kuratorer fra Österreichischen Nationalbibliothek (ONB) og La Bibliothèque nationale de France (BnF). Vores udenlandske samarbejdspartnere skaber nu nyttig kode, test og reviews. Dette samarbejde forventes udbygget.

20 - 22. september er der aftalt en 2½-dages workshop omkring NetarchiveSuite i Wien i Østrig med deltagelse af de aktive bidragsydere af kode – det vil sige BnF, ONB og Netarkivet. Workshopen vil indeholde en teknisk del og en mindre teknisk del.

NetarchiveSuite har med succes testet release 3.12.0. Testmiljøet er på vej til at blive opgraderet til denne, og der skal laves accepttest. Næste trin er derefter at opgradere produktionsmiljøet også. Dette forventes at ske mellem to tværsnitshøstninger i juni.

Redaktionsgruppemøde

Kulturministeriets nye redaktionsgruppe har haft sit første møde d. 11. januar.

Redaktionsgruppen, der er udpeget for de næste 2 år, består af:

- Lektor Niels Brügger, Medievidenskab, Aarhus Universitet
- Projektchef Peter Djurup, Foreningen af Danske Interaktive Medier
- Webmanager Tina Skærbæk, IT Universitetet i København
- Forlægger Henrik Borbjerg, Den Danske Forlægger Forening

Netarkivet er glade for interessante vurderinger og kommentarer ved de 2 årlige møder.

Personale

Den relativt nye netarkiv udvikler Henrik Kirk er kommet på et andet projekt på Statsbiblioteket. Derfor har Statsbiblioteket ansat en ny udvikler Mikis Seth Sørensen, der er startet d. 1. maj. Netarkivet siger tak til Henrik for et godt samarbejde og velkommen til Mikis.

Claus Lomborg
Daglig leder, Netarkivet
clo@netarkivet.dk