netarkivet.dk

Nyhedsbrev juli 2007

Velkommen til Netarkivets nyhedsbrev for juli 2007.

Netarkivet fejrer to års fødselsdag

4. juli fejrede Netarkivet to års fødselsdag. Det blev festligholdt ved en virtuel sammenkomst via videokonference mellem Statsbiblioteket og Det Kongelige Bibliotek. Eftermiddagen bød på både taler, fremvisninger og kollegial hygge.

Det har været to gode år, der har budt må mange spændende udfordringer. Netarkivet er kommet rigtig godt fra start og kan allerede efter to år bryste sig af at være blevet et af verdens største arkiver af internetmateriale.

Artikel om erfaringerne fra de første to år

Netarkivets erfaringer med indsamling af internetmateriale i de første to år er blevet samlet i en artikel af Grethe Jacobsen fra Det Kongelige Bibliotek. Artiklen blev skrevet til en konference i Spanien og blev i den forbindelse også oversat til spansk. Begge udgaver er tilgængelige via Netarkivets netsted: http://netarchive.dk/publikationer/index-en.php

Indsamlingserfaringerne samt juridiske og økonomiske problemstillinger om Netarkivets virke er ligeledes blevet samlet i et notat til Kulturministeriet. Den mest alvorlige konklusion er, at prisen på lagerplads ikke falder omvendt proportionalt med stigningen i datamængden. Det var ellers Netarkivets oprindelige tese, og den viser sig altså nu ikke at holde vand. Datamængden stiger nogenlunde som forventet, men hardwarepriserne falder ikke tilsvarende. Det betyder, at Netarkivet enten skal bruge flere penge for at opretholde det nuværende niveau af tværsnitshøstninger (fire årlige), eller at der skal udarbejdes en ny indsamlingsstrategi.

Netarkivets system i open source

4. juli blev også dagen, hvor Netarkivets system blev udgivet i open source. Systemet har i den forbindelse fået et navn: NetarchiveSuite og egen hjemmeside: http://netarchive.dk/suite

Systemet blev inden udgivelsen testet af både Internet Archive (USA) og det norske nationalbibliotek, og vi fik i den forbindelse meget rosende feedback.

Netarkivet har planlagt en teknisk workshop om systemet i København i september. Der har allerede nu været vist stor interesse, og de 15 pladser er stort set besat.

Integrering af leverede data

Der har i de seneste par måneder været arbejdet på et projekt støttet af Kulturministeriet om integration af leverede data. Der er i den forbindelse lavet tekniske forsøg om, hvordan man bedst lægger leverede net-data i et format, der gør, at de kan gemmes sammen med de data, Netarkivet selv indsamler.

Vi har fået en del meget blandede data fra Danmarks Radio. Der er udviklet en prototype på et program, der kan tage leverede data og lægge dem i de såkaldte ARC-filer (som er det format, Netarkivet lagrer alle indsamlede data i).

For at teste kvaliteten af konverteringen og de leverede data selv har Netarkivet lavet en test-installation af to open source adgangsværktøjer: NutchWAX og WayBack.

De to værktøjer giver fritekstsøgning og tidsnavigation i netmateriale. Det har været meget positivt at teste værktøjerne, og vi ser frem til at tilbyde disse på netarkivets samlede datamængde til i første omgang forskerne og på lidt længere sigt forhåbentligt også til alle andre interesserede.

Forskeradgang til Netarkivets data

Siden slutningen af marts 2007, hvor den første forsker fik adgang til Netarkivet, er rygtet løbet, og pt. har seks forskere etableret adgang, mens tre andre har fået bevilget adgang via ansøgning men endnu ikke fået adgangen etableret rent teknisk.

Adgangen er stadig meget primitiv. Vi er nødt til at vide præcis, hvilke domæner forskeren ønsker at kigge på, og på baggrund af domænenavne kan vi generere lister over, hvilke indsamlinger vi har af de pågældende domæner. Ved hjælp af disse lister kan forskerne så kigge på forskellige arkiverede udgaver af de valgte domæner.

Med den store forskerinteresse er behovet for mere avancerede slutbrugerværktøjer ikke blevet mindre. Udviklingsplanerne for Netarkivet er pt. således, at integration mellem NetarchiveSuite og de to tidligere omtalte open source værktøjer skulle komme på plads i foråret 2008. Indtil da må vi og forskerne leve med den primitive og lidt administrationstunge adgang.

IIPC møde i Paris

International Internet Preservation Consortium havde stormøde for alle nye og gamle medlemmer i Paris i april. Netarkivet var godt repræsenteret og fik gjort god reklame for NetarchiveSuite og systemets udgivelse i open source.

Bjarne Andersen fra Statsbiblioteket blev desuden valgt til Technical Committee, og Netarkivet har dermed fået en rigtig god plads i et meget spændende forum.

Fjerde tværsnitshøstning

Den fjerde tværsnitshøstning blev igangsat 1. marts 2007. Den inkluderer nu ca. 803.000 domæner, hvoraf det forventes, at ca. 80 % er aktive – det vil sige, at Netarkivet indsamler ca. 640.000 danske hjemmesider.

Tværsnitshøstningen forventes afsluttet en af de nærmeste dage og kommer totalt til at fylde ca. 14.5Tb mod 9.8Tb i den forrige tværsnitshøstning. Den store stigning kan til dels forklares med den ekstraordinært store stigning i antallet af nye domæner, da vi jo for første gang inkluderer næsten 40.000 domæner uden for .dk. Disse forventes desuden at være gennemsnitligt større end netstederne på .dk – men det er naturligvis en tese, der skal afprøves. Der skal laves en samlet analyse, der kan forklare stigningen.

Status på bitarkiverne

Bitarkiverne rummer nu følgende datamængder:

Tværsnitshøstning:	39.618 GBytes i 409.098 filer
Selektiv høstning:	4.415 GBytes i 52.374 filer
Begivenhedshøstning:	3.505 GBytes i 36.977 filer

Der er stadig 100 % konsistens mellem de to online-kopier på henholdsvis Statsbiblioteket og Det Kongelige Bibliotek. Bitarkivet på KB består nu af 36 maskiner, og Netarkivet har efter to år måttet skrotte den første maskine, der voldte så mange problemer, at det var for dyrt at vedligeholde den.

Der er i den seneste periode flyttet en del filer mellem maskiner internt på Statsbiblioteket, og Netarkivets bitbevarings-system har vist sig særdeles nyttigt til hurtigt at tjekke, at alle filer var flyttet sikkert.

Redaktionsgruppemøde 22. marts 2007

Redaktionsgruppen havde møde med Netarkivet 22. marts 2007 på Det Kongelige Bibliotek. Der blev blandt andet diskuteret indsamling af passwordbeskyttede netsteder. Det blev besluttet, at Netarkivet skal lave en redegørelse for, hvor stort omfanget af passwordbeskyttet materiale, der reelt er omfattet af loven, egentlig er.

Redaktionen diskuterede desuden etiske aspekter ved indsamling af fx. dating-sider. Disse er uden tvivl omfattet af loven, men der var flertal for en beslutning om at kontakte udbyderne, inden vi logger ind på disse sites, for ikke at skabe uvenner og for at sikre, at indsamlingen foregår bedst muligt for begge parter.

Desuden blev det besluttet, at redaktionen ved næste redaktionsmøde i september skal kigge på, hvad de øvre grænser for indsamling i tværsnitshøstningerne betyder for arkivet. I øjeblikket opererer Netarkivet med en øvre grænse på 2Gbytes pr. domæne.

DRAMBORA

Statsbiblioteket er partner i det europæiske projekt DPE (Digital Preservation Europe: www.digitalpreservationeurope.eu). DPE har udviklet en model for audit af arkiver i forhold til begrebet *Trusted Repository*. I den forbindelse skal alle DPE-partnere lave en test-audit af et repository for selvfølgelig at teste det valgte repository, men i lige så høj grad for at teste modellen.

Netarkivet var det umiddelbart mest oplagte eksempel, og der har derfor været arbejdet på at samle al tilgængelig information/dokumentation omkring Netarkivet. Det gælder alt fra kravspecifikationer, procedurer og guidelines til juridiske dokumenter som fx. pligtafleveringsloven.

Modellen bygger meget på dokumentation og risikostyring. Den ser lovende ud og vil uden tvivl blive nyttig i det videre arbejde mod en evt. certificering af Netarkivet som Trusted Repository.

Planer for den kommende periode

1. Implementering af DS-484 i forhold til Netarkivets data
2. Workshop omkring NetarchiveSuite i september
3. Start af den femte tværsnitshøstning
4. Support i forbindelse med udgivelsen af NetarchiveSuite