netarkivet.dk

Nyhedsbrev december 2006

Velkommen til Netarkivets nyhedsbrev for december 2006.

Access-systemer

Til styregruppemødet for netarkivet d. 8. november havde en lille arbejdsgruppe bestående af driftslederen, en repræsentant for udviklingsgruppen og en repræsentant fra KB-IT lavet et lille udredningsarbejde for at afdække, hvad det vil kræve at implementere en egentlig slutbrugeradgang til netarkivet. Notatet omhandler såvel hardware som overslag for udviklerressourcer og driftsomkostninger. Hardwaremæssigt kunne opgaven med den nuværende datamængde løses for omkring 75.000, men grundet en del pres på især udviklingsgruppen udskydes beslutningen om denne investering til medio 2007.

Desuden skal der igangsættes et koordineringsarbejde mellem netarkivet og SB/KB mht. implementering af de love og regler, der gælder for adgang til netarkivets data. Dette igangsættes primo 2007 hvor såvel SB som KB har afsluttet første fase af den generelle implementering af DS-484, som bibliotekerne i forvejen arbejder intenst med. Adgangen til netarkivet følge i høj grad DS-484 med et ekstra sikkerhedslag på grund af de personfølsomme data i arkivet.

Danske domæner uden for .dk

Der har i de sidste måneder været et intenst arbejde i gang omkring indsamling af danske domæner uden for .dk. Indsamlingen har baseret sig på flere forskellige strategier:

- Søgning på danske stednavne via google (URL'er der ikke rummer .dk)
- Gennemgang af .dk-forsider der øjeblikkeligt sender brugeren uden for .dk-domænet
- Søgning efter links i alt det høstede materiale (dvs. alle høstede sider på .dk)

Alle 3 strategier gav store puljer af potentielle domæner – fx. gav søgning efter links fra alle .dk-sider godt 3.3 millioner unikke domæner

Den totale pulje af domæner blev derefter kørt gennem en såkaldt IP-lokalisator, der med rimelig sikkerhed kan afgøre, i hvilket land et givent navn hører til (hvor den server der har navnet fysisk er placeret).

Dette bragte antallet af domæner der placeres i Danmark ned på 46609. Disse fordeler sig således på top-10:

Top Level Domæne	*Antal*
.com	22750
.se	8058
.net	4913
.org	2381
.nu	1473
.de	1108

Top Level Domæne	*Antal*
.no	1106
.info	1092
.eu	488
.biz	474

IP-lokaliseringen har vist sig at være ret præcis og det er derfor besluttet, at alle de fundne domæner lægges i systemet inden næste tværsnitshøstning.

Mest overraskende var det nok, at der er fundet mere end 8000 .se (svenske) domæner og også ganske mange tyske (.de) og norske (.no). Vi overvejer i øjeblikket om de skal inkluderes. Vi ved, at i hvert fald svenskerne selv arkiverer alle .se netsteder, og Norge er på vej til at gøre det samme. Det ville derfor give god mening at undlade dem i netarkivets høstninger.

Duplikatreduktion

Duplikatreduktionen har nu været i produktion i nogle måneder og resultaterne har været rigtig gode. For den netop afsluttede tværsnitshøstning har denne mekanisme givet en reduktion i den data vi faktisk gemmer i arkivet på godt 31 % - det betyder for denne ene høstning mere end 3TB pladsbesparelse.

Vi havde oprindeligt kalkuleret med en besparelse på op mod 50 % og dette mål er altså ikke helt nået endnu. Dette skyldes primært, at der mellem de 2 sidste høstninger gik næsten 5 måneder og det betyder i praksis, at mere har ændret sig på nettet, end hvis der kun var gået 3 måneder, som er vores mål (4 årlige tværsnitshøstninger).

For de selektive høstninger er resultatet en besparelse på mellem 50 og 70 % - det er naturligvis på en noget mindre datamængde, men stadig ikke uden betydning. Det betyder for disse høstninger, at der bliver "råd" til at lade høstningerne gå lidt længere/dybere, end hvis systemet ikke havde kunnet sortere dubletter fra.

Tredje tværsnitshøstning

Netarkivets 3. tværsnitshøstning blev færdig den 18. december efter at have været godt 6 måneder under vejs med de 2 skridt høstningen er foretaget i (først op til 10MB pr. netsted – derefter op til 500MB / 1GB). Fra de 6 måneder skal trækkes en pause på ca. to en halv måned hvor vi ikke tværsnitshøstede pga. implementering af duplikatreduktions-mekanismen, så vi kan selv i 2 skridt foretage en tværsnitshøstning på ca. 3½ måned.

Høstningen fylder 9811 GBytes mod den forrige høstnings 10171 GBytes. Så selvom nettet uden tvivl er vokset er høstningen blevet mindre på grund af flere faktorer:

- Duplikatreduktionen, der har sparet ca. 3 Tbytes
- en lang række alias'er (forskellige navne for samme netsted) er blevet undgået
- en række generelle såkaldte crawlertraps (fx. kalendere) er blevet undgået

Denne tværsnitshøstning havde for 2300 domæners vedkommende fået en grænse på 1GBytes per domæne mod standard-loftet på 500MBytes. Der var dog en mindre fejl i systemet der gjorde, at de jobs der skulle indsamle netop disse ikke kunne køres helt færdige og derfor ikke blev indsamlet så komplet, som det kunne ønskes. Antallet af domæner der ramte loftet (på enten 500MB eller 1GB) var 4751. Af disse ramte 4478 domæner 500MB og 263 domæner ramte 1GB.

Disse skal – for de der ikke blev tjekket efter sidste 500MB høstning – nu tjekkes og have grænsen vurderet igen.

Ved de tidligere tværsnitshøstninger har vi tydeligt kunne mærke, når en ny tværsnitshøstning blev igangsat, idet antallet af henvendelser per telefon eller e-mail steg mærkbart. Denne gang har der kun været 4 henvendelser omkring tværsnitshøstningen af mere end 700.000 domæner, så det må siges at være et meget lille antal. Noget tyder på, at de netsteder der har følt sig generet nu har fået løst problemerne, og at vi derfor fremover forhåbentlig kun vil opleve et meget lille antal negative henvendelser omkring vores virke.

Næste tværsnitshøstning

Den næste tværsnitshøstning igangsættes i starten af 2007. Den kommer for første gang til at gå langt ud over .dk idet de nyligt identificerede domæner uden for .dk bliver inkluderet.

Som noget nyt planlægger vi også at køre indsamlingen i 1 skridt – dvs. alle netsteder på en gang, og ikke som tidligere først alle de små (under 10MB) og derefter resten.

Det skal desuden blive spændende at følge hvor meget duplikatreduktionen tager på to høstninger med forholdsvis kort tid imellem.

Netarkivet dokumenterer strukturreformen

I forbindelse med strukturreformens ikrafttrædelse den 1. januar sker der en masse omlægninger ikke bare fysisk blandt kommune, amt og stat men også virtuelt på de involveredes netsteder.

Mange af de gamle kommuner og deres netsteder nedlægges og nye opstår. Nogle af de nye kommuner har allerede nye domænenavne (fx. www.nyskanderborg.dk) og nogle beholder de gamle men lægger nyt indhold på i tiden omkring 1. januar. Mange af de nye er de sidste mange måneder blevet brugt som informationskanaler omkring arbejdet med sammenlægningen (fx. www.nyhjoerring.dk) og disse midlertidige netsteder vil uden tvivl blive afløst af kommunernes officielle netsted, og i den forbindelse vil megen information sandsynligvis gå tabt.

Amterne nedlægges som bekendt helt, og deres netsteder følger efter al sandsynlighed med. De nye regioner bidrager med 5 nye netsteder, der alle har fungeret i længere tid, men som med mange af de nye kommuner ind til nu, som mere midlertidige opslagstavler med information om opbygningsarbejdet.

Netarkivet ser en meget vigtig opgave i denne forbindelse og gør derfor i øjeblikket en ekstra indsats for i første omgang, at identificere netsteder, der ændres radikalt eller nedlægges efter 1. januar 2007. Disse vil så blive indsamlet ekstraordinært i slutningen af december måned for at have et så komplet billede som muligt inden overgangen.

Status på bitarkiverne

Bitarkiverne rummer nu følgende datamængder:
Tværsnitshøstning: 25.538 GBytes i 263.559 filer
Selektiv høstning: 3.068 GBytes i 36.018 filer
Begivenhedshøstning: 2.929 GBytes i 30.834 filer

Totalt har netarkivet altså nu arkiveret knap 32Tb hvilket er en anseelig mængde data. Det ligger dog stadig inden for budgettet der rummer plads til 37Tb ved årsskiftet 2006/2007.

Den manglende udfyldelse af lager-budgettet skal primært findes i det faktum, at netarkivet i 2006 kun har gennemført 2 tværsnitshøstninger mod reelt planlagt 4. Den ene af disse var uden duplikatreduktion og fylder derfor naturligt mere end planlagt.

Som skrevet i sidste nyhedsbrev er tidens tand også begyndt at påvirke bitarkivet på KB. 3 såkaldte bitarkivmaskiner har været gået ned og de 2 af disse er endnu ikke kommet i live igen pga. manglende leverance af reservedele. Desuden har netarkivets interface til håndtering af fejl som denne vist sig ikke at kunne håndtere så store datamængder på fornuftig vis. Det er derfor planlagt at 2-3 nye maskiner sendes til Århus for at blive fyldt med de manglende data igen. Videreudvikling af interfacet er på udviklings-programmet for foråret 2007, og da praksis nu har vist, at disk-crash nok er det mest hyppigt forekommende, vil understøttelsen af netop det blive prioriteret højt.

Netarkivet har som mange ved, 3 kopier af al data (2 kopier på disk og 1 kopi på bånd). Den anden kopi på disk er senest blevet kontrolleret i begyndelsen af december, og der var stadig ikke en eneste fejl, så al data er stadig 100 % intakt.

Planer for den kommende periode

1. Indrulning af de identificerede danske domæner uden for .dk
2. Netarkivets system testes på Island
3. Styregruppemøde den 29. januar
4. Fortsat omlægning af de selektive høstninger
5. Implementering af DS-484 i forhold til netarkivets data

Det var alt for denne gang – det er jo også sidst på året, så jeg vil gerne benytte lejligheden til at sige tak til alle, der har medvirket til et godt samarbejde mellem SB og KB omkring netarkivet i det forgangne år.

God jul og godt nytår.