Netarkivet.dk indsamler data om .dk

Internettet er også del af den kulturarv, der skal dokumenteres for fremtidens generationer. Derfor er Statsbiblioteket og Det Kongelige Bibliotek under det fælles banner Netarkivet.dk gået i gang med at indsamle og bevare den danske del af internettet.

Af Birgit Nordsmark Henriksen bnh@kb.dk, Det Kongelige Bibliotek

Mange har hørt om CERN's (Conseil Européen pour la Recherche Nucléaire) største partikelaccelerator under den fransk-schweiziske grænse ved Genevé. De fleste af os kan blive helt åndeløse, når vi hører om de store datamængder, der årligt produceres på stedet: 15 Pbyte (15.000.000 Gbyte) eller nok til at fylde 1,7 millioner dual-layer DVD'er.

Knap så mange har hørt om de digitale lagre ved nationalbiblioteksfunktionerne på Statsbiblioteket og Det Kongelige Bibliotek. De øges med cirka en halv Pbyte årligt, og data opbevares i de to institutioners egne installationer, hvorfra de gøres tilgængelige.

Mål: Dokumentation af kulturarven

Den nye pligtafleveringslov, som trådte i kraft 1. juli 2005, var startskuddet for den kraftige vækst i omfanget af digitale data. Væksten i data forventes fordoblet over en femårig periode blandt andet på grund af en øget retro-digitalisering af bibliotekernes egne samlinger.

De mange data indsamles for at kunne dokumentere kulturarven for eftertiden. For at sikre at de indsamlede data kan læses både på kort og på langt sigt, har bibliotekerne måttet tilegne sig nye kompetencer inden for "digital bevaring". Begrebet dækker over alle de processer, der skal sikre, at digitalt materiale fortsat er læsbart og forståeligt, når alt omkring dataene har ændret sig.

Tre indsamlingsstrategier

Indsamlingen af offentliggjort materiale fra den danske del af internettet (.dk) samt den del af nettet, som er rettet mod et dansk publikum, sker ved brug af programmer, som indsamler dokumenter ved automatisk at følge links og hente dokumenter ned til Netarkivets datalagre i København og Aarhus. Der indsamles løbende efter tre forskellige strategier: En selektiv, en begivenhedsorienteret og et fuldt tværsnit.

Den *selektive* indsamling består i indsamling fra mellem 80 og 100 websites på daglig, ugentlig eller månedlig basis. Fælles for de udvalgte sites er, at indholdet af dem ændres meget hyppigt, og at websitet adskiller sig ved at være meget unikt eller repræsenterer en typisk dynamisk og meget benyttet site. En rådgivende gruppe støtter Netarkivet med henblik på at vedligeholde listen over sites, der skal indsamles efter denne model.

En *begivenhed* er kendetegnet ved, at den udløser debat i befolkningen, og den forventes at få betydning for dansk historie eller påvirke det danske samfund. Begivenheden genererer en masse nye websteder med kort levetid, og den dækkes intensivt på allerede eksisterende websites. Begivenhedsindsamlinger er de mest arbejdsintensive, fordi det indsamlede materiale ofte er meget flygtigt og skal kvalitetskontrolleres, inden originalmaterialet forsvinder fra internettet. Et eksempel på en gennemført begivenhedsindsamling, som ikke var planlagt, er forløbet omkring Muhammed-tegningerne.

En *tværsnitsarkivering* tager udgangspunkt i den fulde domæneliste af top-domænet .dk suppleret med knap 50.000 domænenavne, som ligger uden for .dk, og som retter sig mod et dansk publikum. En dansk tværsnitsindsamling tager mellem tre og fem måneder. Denne indsamlingstype genererer langt flere data end de to andre til sammen. I 2005 fyldte en enkelt tværsnitsarkivering knap 8 TByte – i dag, fem år efter, fylder den ca. 28 TByte, og antallet af danske domæner er i samme periode steget fra ca. 600.000 til cirka 1,2 millioner. Som udgangspunkt startes en ny tværsnitsarkivering, når den forrige er afsluttet.

Netarkivets største udfordring er, at adgangen pt. er begrænset til forskningsbrug, skønt alt materialet er indsamlet fra den offentlige del af internettet. Dette skyldes, at det endnu ikke er lykkedes at finde fuldautomatiske metoder til at afgøre, om en konkret webside indeholder følsomme persondata.

Open Source og ny lagringsløsning
Indsamlingen til Netarkivet sker ved hjælp af software udviklet af Statsbiblioteket og Det Kongelige Bibliotek i fællesskab. I 2007 besluttede de to biblioteker at udgive softwaren i open source under navnet "Netarchive Suite" i håb om, at flere institutioner ville bidrage til opgaven med at videreudvikle programmet. I 2009 sluttede det franske og det østrigske nationalbibliotek op om Netarchive Suite, og fælles udviklingsplaner for den kommende toårsperiode er netop blevet annonceret.

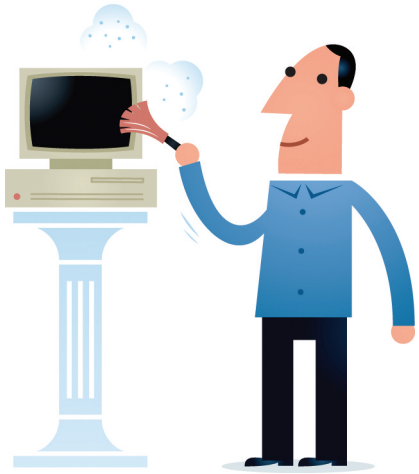
Netarchive Suite indeholder et modul, som sikrer, at alle de indsamlede data lander i København og Aarhus i lagerløsninger, som er organisatorisk, teknisk og geografisk forskellige for at nedsætte risikoen for datatab mest muligt. Netarkivets bitbevaringsløsning – hvor materialets korrekte sekvens af 0'ere og 1-taller bevares over tid – har vi efterfølgende udvidet og hævet op på et højere og mere generelt niveau i et nyt projekt, hvor vi tilrettelægger en IT-arkitektur til et 'fælles bitmagasin'. Gennem egenudviklet software, som skal lægges oven på de medvirkende institutioners forskellige fysiske installationer, vil vi gøre det muligt for en institution at gemme og bevare sine bits på tværs af flere institutioner i det antal kopier og med den sikkerhed, som man ønsker og er villig til at betale for.

Internationalt samarbejde
International Internet Preservation Consortium (IIPC) har til formål at kunne bistå ved indsamling, bevaring og tilgængeliggørelse af internetmateriale for fremtidige generationer. Arbejdet sker gennem aktiv deltagelse i projekter og arbejdsgrupper, som adresserer hver af de tre nævnte hovedaktiviteter (netpreserve.org).

Open Planets Foundation (OPF) har til formål at imødegå udfordringerne ved digital langtidsbevaring med viden om best pratice og adgang til generelt udbredte og anvendte løsninger. Visionen for OPF er at blive organisationen, som binder væsentlige initiativer inden for området sammen (www.openplanetsfoundation.org/).

Danmark er i kraft af Statsbiblioteket og Det Kongelige Bibliotek centralt placeret i ledelsen af såvel OPF som IIPC, og de internationale netværk og projekter er helt afgørende for vores fremdrift på området.

OPF blev dannet i kølvandet af EU-projekt PLANETS, hvor Danmark gennem forskning og udvikling var med til at sikre en koordineret europæisk indsats inden for logisk bevaring af digitalt materiale. PLANETS-projektet fik stor positiv international pressebevågenhed ved dets afslutning, da det deponerede en tidskapsel med et "Digitalt Genom" i de højsikrede datacentre dybt inde i de schweiziske alper. I tidskapslen ligger fem digitale objekter i forskellige dataformater med beskrivelser af, hvordan man om nødvendigt skal genskabe den hardware og software, som skal bruges for at



Kilde: www.digitalbevaring.dk

Digitalbevaring.dk fungerer som et forum for videndeling.

læse og forstå disse filer i fremtiden. Det er planen at tage tidskapslen frem igen om 25 år og se, hvor meget af den gemte information, der er anvendelig.

Fokus nu og i fremtiden
Både Statsbiblioteket og Det Kongelige Bibliotek vil nu og fremover være optaget af en videreudbygning af deres faciliteter til forvaltning – fra indsamling til formidling - af det digitale samlingsmateriale, som de hver især har ansvaret for.

Erfaringer og viden fra PLANETS-projektet og de nationale projekter skal indarbejdes, problemstillinger ved håndtering og processering af de meget store datamængder skal adresseres, og antallet af håndterbare materialetyper skal udvides.

Se mere …
Planer og strategier for bevaring af det digitale materiale er ved at være på plads og kan findes på bibliotekernes websites. Videndeling til andre fagfolk og offentligheden er flerstrenget og sker internationalt på konferencer og gennem forskningspublicering. På nationalt plan sker det især via de tre følgende websites :

www.digitalbevaring.dk indeholder baggrundsartikler, ordforklaringer og nyheder inden for feltet (tekst og illustrationer kan frit benyttes under henvisning til sitet).

www.costmodelfordigitalpreservation.dk indeholder et excel-baseret værktøj til estimering af omkostningerne ved de mange processer, som digital bevaring omfatter.

www.netarkivet.dk indeholder FAQ, artikler og software, og alle kan frit hente og benytte "Netarchive Suite" til arkivering af hjemmesider og nominering af sider, som ønskes arkiveret.