DITTE LAURSEN, DLA@STATSBIBLIOTEKET.DK, SENIORFORSKER, STATSBIBLIOTEKET
PER MØLDRUP-DALUM, PMD@STATSBIBLIOTEKET.DK, SPECIALKONSULENT, STATSBIBLIOTEKET

NETARKIVET 10 ÅR

I Netarkivet kan forskere søge og klikke rundt i arkiverede danske netsteder i deres forskellige versioner over tid. Arkivet er netop fyldt 10 år.

I 2005 blev en ny pligtafleveringslov vedtaget i DK, og arkivering af det danske internet blev obligatorisk. Hensigten med loven var en total dækning af den offentligt tilgængelige danske del af internettet ved hjælp af tre forskellige og komplementære strategier for indsamling: 1) tværsnitshøstning 3-4 gange om året, 2) selektive daglige høstninger af hyppigt opdaterede hjemmesider og 3) begivenhedshøstninger. Med den nye lov blev Danmark et af de første lande i verden med pligtaflevering for dynamisk internetmateriale, og vi indsamler i dag over 1.2 millioner domæner. Arkivet fylder ca. 650 TB og vokser pt. med ca. 130 TB pr. år, og det beskæftiger en række medarbejdere på Statsbiblioteket og Det Kongelige Bibliotek. Opgaven løses af de to institutioner i fællesskab. Vi har tidligere fortalt nærmere om, hvordan vi indsamler materialet, og hvilket slags materiale vi indsamler. I denne artikel vil vi se tilbage på 10 års arkivering af det danske internet. Vi fokuserer på, hvordan vi har håndteret nogle af de mange udfordringer, der opstår, når man vil indsamle og formidle et felt i hastig udvikling.

Udviklingen af antal domæner

I 1987 blev top level-domænet .dk oprettet. Antallet af domæner voksede fra 49 i 1987 til 70 i 1988, og domænerne blev først og fremmest benyttet af firmaer og organisationer i Danmark. I de første mange år var der ikke den store udvikling i antallet af domæner; først fra sluthalvfemserne ser vi en stigning. I dag (23. okt. 2015) er der 1.307.630 .dk-netsteder:

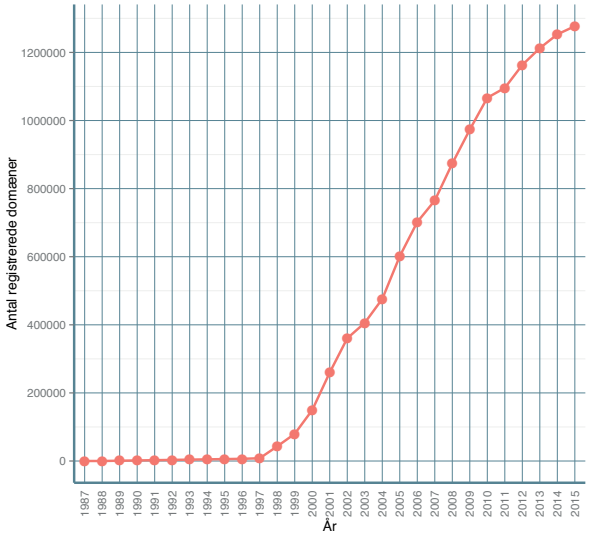

Diagram 1: Udviklingen af .dk domæner.

Med denne udvikling in mente kan man sige, at den danske lovgivning og Netarkivet var sent ude med loven af 2005, hvor der allerede i flere år havde været hundredtusinder af danske domæner. Men faktisk pålagde den tidligere pligtafleveringslov fra 1998 nationalbibliotekerne at indsamle såkaldt offentliggjorte og afsluttede arbejder på nettet. Det var fx monografier, tidsskrifter og pdf-dokumenter, men ikke egentlige netsteder. 32.926 dokumenter blev indsamlet i 1998-2004, og samlingen er åben for offentligheden i dag via Statens Netbibliotek. Men da der kun blev indsamlet manuelt registrerede dokumenter, er masser af netindhold faktisk gået tabt frem til 2005. Netarkivet har dog efterfølgende erhvervet eksempler på materiale fra før 2005, fx gennem et samarbejde med det amerikanske

Internet Archive, som startede med at høste i 1996. Vi kan ikke dække hele hullet frem til 2005, men vi har mange netsteder, der i nogen grad kan eksemplificere udviklingen.

Specialsamlinger

Netarkivet har en række specialsamlinger som både indeholder materiale fra før 2005 og materiale indsamlet parallelt med hovedarkivet. Siden 2009 har vi fx indsamlet e-bøger separat, fordi de er afsluttede værker og typisk ligger bag ved betalingsmure. De indsamles derfor bedst med en målrettet og skræddersyet indsats. I andre tilfælde bygger vi specialsamlinger af nød, fordi materialet ikke kan høstes af vores almindelige høstere. Det gælder fx ved høstning af video på internettet. I 2012 udviklede Netarkivet et værktøj til at indsamle danske videoer fra youtube. Samlingen består i dag af ca. 125.000 videoer og er stadigt voksende. Derudover har Netarkivet lavet videooptagelser (skærmfilmning) af materiale, som ikke er født som video, men som dokumenteres bedst via video, fx interaktive produktioner som Second Life og andre spillignende universer som Lego Universe. Netarkivet har også udviklet et værktøj til skærmfilmning, som kan programmeres på forhånd. Det er fx blevet brugt på valgaftener, hvor der har været live tv-udsendelser på nettet, som ikke var de samme udsendelser, der blev sendt på almindelig broadcast tv. Specialsamlingerne er alle et udtryk for, at Netarkivet gør hvad det kan for at indsamle materiale, som dokumenterer det offentlige danske internet. Imidlertid er der bevaringsudfordringer for specialsamlingerne, som ikke ligger i samme formater som hovedarkivet, ligesom der også er tekniske barrierer for at tilgængeliggøre samlingerne, da systemerne til tilgængeliggørelse pt. ikke er bygget til disse formater.

Skærmfilm af det nu lukkede interaktive Lego Universe (2010) .

Udenlandsk webdanica

Pligtafleveringsloven dækker også sider uden for .dk-domænet, som er skrevet på dansk, skrevet af/om danskere og/eller rettet mod et dansk publikum, fx netsteder som lego.com, skrivunder.net og carlnielsen.org. Gennem årene har vi benyttet forskellige metoder for at lokalisere de relevante netsteder, fx har vi benyttet et geo-ip program, der kan afgøre, om domænet er placeret på en server i Danmark, og derfor kan antages at være dansk. Vi har også på netarkivet.dk en side, hvor alle kan angive netsteder uden for .dk-domænet.

Mængden af danske webmaterialer uden for .dk vokser, og nye analyser viser, at der kan være op mod 100.000 relevante danske domæner uden for .dk ud over de ca. 45.000 domæner, som vi i dag indsamler. I det kommende år vil Netarkivet derfor lave en indsats for at indsamle dette materiale, nu og fremadrettet. En særlig udfordring er domæner, som kun delvist er danske. Eksempelvis er facebook.com ikke dansk, men der er alligevel masser af offentlige danske profiler, som er omfattet af pligtafleveringsloven. Her skal høsterne indstilles på en måde, der afgrænser det danske fra det udenlandske.

Tekniske udfordringer i indsamlingen

Internettet er noget af det mest komplicerede materiale, man kan tænke sig at indsamle. Udviklingen af de bagvedliggende teknologier går hurtigere og hurtigere, og de anvendte høstningsteknologier har svært ved at følge med. Links beregnet i browsere ved hjælp af JavaScript er i dag en af de store udfordringer for vores høstere, som har svært ved at følge og beregne disse dynamisk oprettede links.

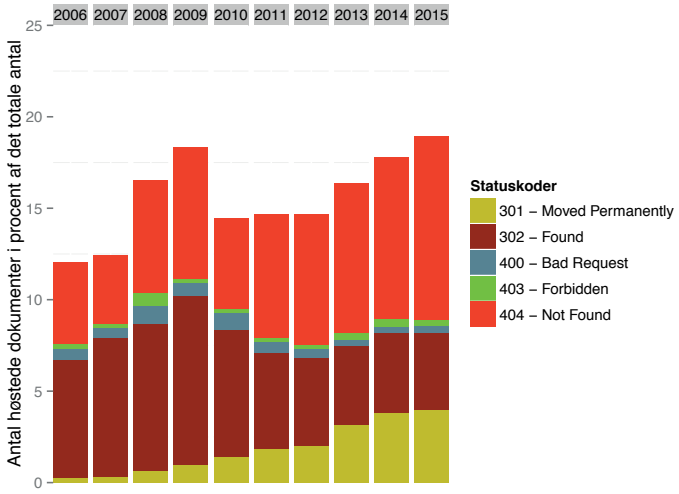

Diagram 2: Diagram over svarkoder.

Svarkoder modtaget fra internettet under høstning kan bruges til at analysere høstningskvaliteten. Hvis en svarkode har værdien *OK*, blev netstedet modtaget som forventet og kan gemmes i arkivet. Det modsatte af *OK* er *Not Found*, der angiver, at det ønskede netsted ikke eksisterer. Disse Svarkoder, som Netarkivet modtager på høstninger, er faktisk vokset over tid og udgør i 2015 ca. 1,5% af den samlende datamængde. Forholdsvist er det ikke meget data, men i én tværsnitshøstning svarer det dog til 85 millioner netsider eller mere end 500 GB data.

Som det ses på diagram 2 findes der en del andre svarkoder, som kan give indsigt i høstningen. Fx ses et fald i *Found* anmodninger. Sammenholdes dette med en stigning i antallet af *Not Found*, kunne en mulig forklaring være, at internettet bliver mere dynamisk, men at den tilgængelige høstningsteknik falder mere og mere bagud i forhold den anvendte teknologi på nettet. For at forbedre høstningsteknologierne udvikler nationalbibliotekerne, med udgangspunkt i høstningsanalyser som denne, løbende på det centrale høstningssystem NetarchiveSuite. Heldigvis bliver langt det meste høstet succesfuldt.

Adgang til Netarkivet

I 1999 kunne alle kun få adgang til de pligtafleverede netdokumenter via en enkelt computer placeret på de to pligtafleringsbiblioteker. Denne maskine blev internt navngivet "munkemaskinen" med en allusion til klosterbiblioteker, hvor bøger var lænket fast til læsepultene. Men da den nye pligtafleveringslov i 2005 blev udvidet til at omfatte alle webmaterialer, blev adgang til Netarkivet også omfattet af persondataloven, og adgangen blev derfor begrænset. For at få adgang til Netarkivet skulle man være forsker og have et forskningsprojekt. Sådan er det også i dag.

Med den nye pligtafleveringslov udviklede vi et nyt softwaresystem til indsamling, NetarchiveSuite, som også havde et adgangsmodul. Kun enkelte forskere havde adgang til dette for dem ikke særlige brugervenlige system. I 2012 blev integration med Wayback-systemet fra Internet Archive muliggjort, og nu kunne forskerne gennemse netsteder, hvis de på forhånd kendte url'en. Det var en helt ny og nem adgangsform, og der kom flere brugere til arkivet. Et endnu større vendepunkt kom i foråret 2015, hvor Net Archive Search blev lanceret som en fødselsdagsgave til Netarkivets 10-års fødselsdag. Denne nye og banebrydende søgemaskine bygger på et stort indeks, der for første gang gør det muligt for forskere at udføre tekstforespørgsler i det komplette Netarkiv. For eksempel kan man bede om pdf-dokumenter skrevet på arabisk og høstet i 2014, eller man kan bede om de HTML-sider fra DR, der linker til TV-2's hjemmeside. De udførte søgninger kan også samles i komplekse søgetermer og derved skabe en kombination mellem kvalitativ og kvantitativ forskning – alt sammen i realtid og på en eksplorativ måde.
En anden nyåbnet mulighed er etableringen af DeIC Nationale Kulturarvscluster på Statsbiblioteket. Denne supercomputer benytter de meste moderne teknologier inden for data science og giver for første gang mulighed for at stille forskningsspørgsmål til Netarkivet som helhed. Hvis man fx vil kende danske domæners størrelse over tid, er det sådan et anlæg, man skal bruge.

Vi vil gerne udvide vores brugergruppe til andre end forskere, og vi undersøger lige nu muligheden for en kombination af automatiseret og manuel screening af en

mindre del af arkivet med henblik på at øge adgangen. Efter screening kan materialet gøres tilgængeligt på læsesal på pligtafleveringsinstitutionerne, og måske på et senere tidspunkt også online for studerende. Nye metoder kan også, på nye måder, åbne adgang. For eksempel kan kontrolleret datamining i princippet åbne op for nye brugergrupper, ved at de får adgang til resultater af dataminingen uden at få adgang til selve arkivet.

Samarbejde internationalt og nationalt

Netarkivet samarbejder med andre nationale arkiver om at udvikle høstningssystemer og udveksle viden og erfaringer. Det sker blandt andet i regi af IIPC (International Internet Preservation Consortium), som Netarkivet var med til at stifte i 2003. Vi har også et internationalt community omkring vores selvudviklede indsamlingssystem NetarchiveSuite, som andre nationale arkiver også benytter og videreudvikler på, fx i Frankrig og Østrig. På nationalt plan samarbejder Netarkivet med forskere om at udvikle forskningsværktøjer og 'workspaces' til forskning i Netarkivet. Det sker blandt andet i regi af DigHumLab, som er en national paraplyorganisation over specialiserede, nationale og forskerdrevne digitale infrastrukturer. Bag DigHumLab står de humanistiske fakulteter på AU, AAU, SDU og KU samt Statsbiblioteket og Det Kongelige Bibliotek.

Fremtiden

Det danske internet vokser fortsat, og vi har store udfordringer i forhold til både indsamling, bevaring og tilgængeliggørelse. Indsamlingsmæssigt skal vores høstningsværktøjer fange mere og mere avanceret indhold, og kun omfattende analyser kan hjælpe os til at blive klogere på vores arkiv og udvikle vores høstere. Lige nu arbejder vi på specialiserede indsamlingssystemer til e-bøger, e-aviser og e-musik, samtidigt med at vi udvikler bedre automatiserede metoder til den generelle indsamling. Ud over høstning skal vi også bevaringsmæssigt udvikle vores systemer, så de kan skalere til mængderne, differentiere samlinger og håndtere data under kontrollerede forhold, fx ved at data ligger på forskellige typer af lagringsmedier i geografisk adskilte kopier. Tilgængeliggørelsesmæssigt stiger antallet af brugere, og de krav, de stiller til at kunne tilgå materialet, vokser. Dette forudsætter solide, sikre og brugervenlige systemer. Vi har også en stor formidlingsopgave i at fortælle brugere om, hvad arkivet indeholder, og hvordan man kan benytte det. Heldigvis – og på trods af de mange udfordringer med et stadig voksende og nær ved ustyrligt og kaotisk arkiv – så har vi et fantastisk materiale, som vi er stolte af at vise frem.

En længere akademisk version findes i Laursen, D. og Møldrup-Dalum, P. in Brügger (ed.) (2017): Web 25: Histories from the first 25 Years of the World Wide Web.

Netarkivet 2005 -

Netarkivet drives af Det Kongelige Bibliotek og Statsbiblioteket i fællesskab med udgangspunkt i pligtafleveringsloven.

Netarkivets formål er at arkivere materiale offentliggjort på den danske del af internettet.

Arkivet starter i juli 2005 og anvender fire forskellige høstningsstrategier:

Tværsnitshøstning af alle danske domæner 4 gange årligt, pt. ca. 1 million domæner.

Selektiv høstning af ca. 80-100 domæner med hyppigere frekvens, typisk nyhedssites (fx dagligt).

Begivenhedshøstning af 2-3 begivenheder årligt (fx folketingsvalg).

Forskere og ph.d.-studerende kan få onlineadgang til arkivet

Andre, som skal bruge arkivindhold til videnskabelige studier, kan benytte arkivet på Det Kongelige Bibliotek og Statsbiblioteket.

Arkivet fylder lige nu 654 TB.

Læs mere om Netarkivet på netarkivet.dk