3.2. Specielt vedrørende strategier mv. for indsamling af materiale

3.2.1. Baggrund

De nærmere bestemmelser om indsamlingen m.v. er fastsat i kapitel 2 i bekendtgørelse nr. 636 af 13. juni 2005 om pligtaflevering af offentliggjort materiale.

Pligtaflevering af internettet administreres af Det Kongelige Bibliotek og Statsbibliotektet i fællesskab. I den forbindelse har de to biblioteker dannet en organisation, Netarkivet.dk, der ledes af en styregruppe bestående af repræsentanter for de to biblioteker samt den daglige leder af Netarkivet.dk.

Netarkivet begyndte som planlagt at indsamle og bevare den danske del af internettet i juli 2005. Indsamling og bevaring foretages ved hjælp af et system, som er udviklet af de to biblioteker. Dette system benytter flere open source-programmer, og Netarkivet har i 2007 udsendt hele systemet som open source.

Der har været stor international interesse for Netarkivets aktiviteter og systemerne i særdeleshed. Således har flere europæiske nationalbiblioteker haft systemet på prøve, allerede inden det blev frigivet som open source, og flere ønsker at teste, om Netarkivets system vil kunne klare deres forpligtelser vedrørende arkivering af internettet. Netarkivet vil i fremtiden udbygge det internationale samarbejde om videreudviklingen af det danske system.

Netarkivet indsamler det danske internetmateriale efter tre strategier:

- Tværsnitshøstning (4 gange årligt)
- Selektiv høstning (80 udvalgte web-steder høstes hyppigt)
- Begivenhedshøstning (2-3 gange årligt)

Baggrunden for valget af denne trestrengede strategi var, at de tre indsamlingsstrategier skulle komplettere hinanden, og det vurderedes, at de hver især med den angivne høstningsfrekvens var nødvendige for at sikre en så fuldstændig dækning af det danske internet som muligt.

3.2.2. Erfaringer

Tværsnitshøstning

Formålet med denne type høstninger er at få et komplet »øjebliksbillede« af det danske internet ved at høste alt det offentliggjorte materiale fire gange om året. Der er i de første to år foretaget tre høstninger, en i 2005 og to i 2006. I 2007 vil der blive foretaget tre høstninger.

En tværsnitshøstning går ud fra en opdateret liste over domæner inden for topdomænet .dk, som Netarkivet får fra DK Hostmaster. Strategien for de første tre høstninger har været at høste i trin, således at alle web-steder høstes op til en vis størrelsesmæssig grænse. Forskellige tekniske metoder til grænsedragning har været afprøvet

Konklusion vedrørende tværsnitshøstningerne: Det har ikke fra starten været muligt at gennemføre fire årlige høstninger på grund af tekniske udfordringer, som dels betød, at der måtte indsættes ekstra udviklingsressourcer på fejlretning og udvikling af særlige moduler, dels at høstningen i perioder måtte stilles i bero. Hertil kommer at det faktiske volumen for en tværsnitshøstning er langt større end oprindeligt anslået. Det anslåede volumen i 2003 var 3,5 TB, mens det faktiske volumen i 2006 var 11 TB. Dette skyldes bl.a. at det er blevet almindeligt, at folk lægger fotos og videoer på deres hjemmesider. Udviklingen bekræfter, at nettet ikke er statisk, og at det derfor er nødvendigt løbende at tilpasse indsamlingsstrategi, software og økonomi.

Selektiv høstning

Formålet med selektiv høstning er at indsamle websteder, der ud fra en faglig vurdering har værdi for fremtiden, og hvor kun det værdifulde bevares. I dialog med forskere fastsattes i forbindelse med loven et antal på 80 udvalgte web-steder.

Det blev i lovbemærkningerne understreget, at da internettet er et dynamisk medie, der præges af skiftende aktivitetsformer og genrer, vil udvalget af websteder, der høstes, skulle vurderes løbende.

Når et web-sted er udpeget som genstand for selektiv høstning, skal det undersøges, hvor ofte og hvor dybt der skal høstes. En netavis vil typisk have to eller flere forsider, der skifter flere gange i døgnet og derunder mange lag sider, der ikke ændrer sig. Selektiv høstning kræver derfor megen manuel overvågning af nettet og de udvalgte web-steder for at sikre, at der høstes de nødvendige sider og i den nødvendige frekvens.

Konklusion vedrørende selektiv høstning: Det har ikke været muligt fra starten at identificere alle 80 web-steder, dels på grund af internettets vækst i omfang og genrer, dels på grund af behov for forbedring af det tekniske system, som de faglige medarbejdere benytter til opsætning og justering af indsamlingsprofilerne for de enkelte web-steder. Efter tilførsel af ekstra fagkyndigt personale forventes samtlige 80 websteder at blive identificeret i løbet af 2007.