Netarkivet - internettets mejetærsker

28.05.2014
af Anette Lerche

Internettet er en guldgrube af informationer. Nogle mere lødige end andre, men i hvert fald er de et billede på tendenser i vores nutid. Derfor skal de bevares, og det sikrer det danske nationale webarkiv - Netarkivet.

Netarkivet høster det danske internet, så forskere i dag og fremover vil kunne forske i vores nutid. Det betyder i praksis, at Netarkivet indsamler data fra alle netsider med en *.dk-adresse*, danske profiler på de sociale medier og en lang række virksomheders sider.

Informationsmængderne er enorme. Printede man det materiale ud, som Netarkivet hidtil har indsamlet, ville papirrækken nå ud til planeten Venus. Men praksis har ikke altid været at indsamle så omfattende data. I sine første leveår fra 1998 til 2004 indsamlede Netarkivet udelukkende onlinematerialer, der kom ind i arkivet via den almindelige pligtaflevering. Det kunne eksempelvis være offentlige myndigheders publikationer.

Men den metode var uholdbar, hvilket en studerende fra Aarhus var med til at sætte fokus på. Hun havde bygget hele sit speciale op omkring Røde Kors' hjemmeside, men tre dage før afleveringen fandt sted, relancerede Røde Kors sin hjemmeside, og væk var alt det, opgaven byggede på. (Dog var der til alt held for den studerende nogle prints af hjemmesiden). Kollegerne på Center for Internetforskning blev noget alarmerede og bidrog blandt andet til en kulturministeriel redegørelse om, hvordan man bevarer den digitale kulturarv.

Den uheldige studerende fik altså sat fokus på en meget relevant problemstilling: *Hvordan bevarer vi de oplysninger, der er gemt på de danske websites?*

- Starten på vores nuværende praksis var altså, at nogen opdagede, at informationerne er væk, når et website forsvinder, siger Jakob Moesgaard fra Det Kongelige Bibliotek.

Han er samlingsansvarlig for Netarkivet, der er et samarbejde mellem Statsbiblioteket og Det Kongelige Bibliotek. Han holdt et oplæg om Netarkivet på Statsgruppens og Privatgruppens temadag Vild med data den 21. marts.

Tre strategier
Siden 2005 har arkivet derfor høstet informationer efter tre strategier.
Den første er en tværsnitshøstning, hvor en *crawler* (et program, der metodisk skanner internetsider, red.) fire gange årligt indsamler alt, der er hosted på en *.dk-adresse*. Men da nogle websites har en kortere levetid end et par måneder, så har man samtidig identificeret omkring 80 websites af særlig interesse, der høstes flere gange dagligt. Det er eksempelvis nyhedsmediernes sider, men også dba.dk (Den blå avis).

Den tredje strategi, der gerne skulle sikre, at arkivet har relevante informationer for de forskere, der en dag skal studere, hvad der foregik i vores nutid, er en begivenhedshøstning. Altså hvor en aktuel begivenhed medfører en række websites, der hurtigt lukker igen. Det kan eksempelvis være kommunalvalg, de olympiske lege eller det internationale

- Men nettet har udviklet sig, siden de strategier blev fastlagt, og man vil ikke altid kunne forudse, hvad der bliver en begivenhed. For det, nogle ville kalde en begivenhed, gør andre ikke. Man kunne eksempelvis diskutere, om vi skulle have høstet alt materialet om sagen med giraffen Marius, hvilket vi ikke gjorde, siger Jakob Moesgaard.

Jakob Moesgaard glæder sig over, at man i Danmark i høj grad bruger .dk-adressen, der gør det forholdsvis let at identificere de fleste danske sites. Men der er også virksomheder, der bruger en .com eller .eu eller noget helt tredje, hvilket gør det mere vanskeligt at identificere alle de sider, man bør høste. Fra dette område tværsnitshøster man knap 50.000 ikke-danske webadresser. Twitter, Facebook og Wordpress har ikke *.dk-adresser*, men rummer i høj grad materiale, der vil være relevant at gemme, og derfor høster Netarkivet også herfra.

Beskyttet af lovgivningen

Trods det faktum, at alle de informationer, som Netarkivet høster er eller var offentligt tilgængelige, så er Netarkivet det ikke.

- Hvilket jo er lidt sjovt, for langt det meste af det, vi har i arkivet, er jo noget, der stammer fra offentligt tilgængelige sider - vi høster jo ikke fra det, der er beskyttet med et password, siger Jakob Moesgaard.

Det er for det første ophavsretsloven, der gør, at Netarkivet hverken må fremsende eller kopiere til folk, der ønsker at bruge Netarkivet. Og den juridiske løsning på det problem, nemlig at møde fysisk op, ligesom man kan gøre i de fysiske læsesale, når man vil bruge digitalt materiale, er ikke mulig i dette tilfælde, da der kan være tale om personfølsomme oplysninger, som en offentlig myndighed ikke må offentliggøre.

Tilbage er kun, at videnskaben kan få fri adgang. Og det har den da også: Eksempelvis planlægges i øjeblikket et projekt, hvor sprogforskere vil se på sprogbrugen på Twitter for at følge, hvordan det danske sprog udvikler sig.

NB
Netarkivet åbnede i maj for adgangen til centraladministrationens digitale dokumenter. Både journalister og den brede offentlighed kan få adgang via Statens Netbibliotek.