netarkivet.dk

Strategi for langtidsbevaring af materiale indsamlet til Netarkivet ved Det Kongelige Bibliotek og Statsbiblioteket 2014

Introduktion

Dette dokument udgør en strategi for langtidsbevaring af materiale indsamlet til Netarkivet ved Det Kongelige Bibliotek og Statsbiblioteket.

Dokumentet opdateres og godkendes af direktionerne på Det Kongelige Bibliotek og Statsbiblioteket hvert tredje år i forbindelse med opdatering af *Politik for langtidsbevaring af digitalt materiale indsamlet til Netarkivet ved Statsbiblioteket og Det Kongelige Bibliotek*.

Dokumentet er offentligt og publiceres på Netarkivets hjemmeside (www.netarkivet.dk).

Dokumentet indledes med en beskrivelse af formålet med bevaringsstrategien. Herefter gennemgås rammerne for og kravene til bevaringsindsatsen.

Formål

Bevaringsstrategien indeholder en række strategiske modsvar på de punkter, der opstilles i *Politik for langtidsbevaring af digitalt materiale indsamlet til Netarkivet ved Statsbiblioteket og Det Kongelige Bibliotek.*

Det overordnede formål med bevaringsstrategien er at redegøre for principperne og prioriteringerne i arbejdet med langtidsbevaring af Netarkivets materialer, således at der kan skabes et beslutningsgrundlag for planlægningen af bevaringsarbejdet.

- *Indsamling* af materiale til Netarkivet behandles ikke i dette dokument. Her henvises til Netarkivets *Politik for indsamling og kassation af materiale til Netarkivet*

Adgang til materiale i Netarkivet behandles ikke i dette dokument. Her henvises til Netarkivets *Politik for adgang til materiale i Netarkivet* [under udarbejdelse (okt. 2014)].

Rammer for bevaring

Ansvar og roller

Netarkivets ledelse udgøres af en styregruppe med en repræsentant for hvert af de tre faglige områder på de to biblioteker samt driftslederen. Styregruppen refererer til direktørerne på de respektive biblioteker.

Det fælles ansvar for drift af Netarkivet er nedfældet i en samarbejdsaftale, som revideres hvert 3. år.

Videndeling og kompetenceudvikling

Netarkivet tilstræber, at personalet til stadighed har opdateret, faglig viden. Det sker på følgende måder:

- Deltagelse i internationale fora: Netarkivets personale deltager i internationale projekter, konferencer og samarbejder, bl.a. i regi af workshops afholdt af International Internet Preservation Consortium (IIPC) og deltagelse i relevante internationale konferencer, der har til formål at udvikle og videndele på området.

Understøttelse af forskning i digital bevaring.Netarkivet tilstræber desuden, at den viden, der opbygges om digital bevaring, formidles til fagfæller, kolleger og offentligheden gennem bidrag til digitalbevaring.dk.

Trustworthy Digital Repository
Netarkivet skal inden 2020 have status som et *Trustworthy Digital Repository***,** der lever op til standarden ISO 16363:2012 - Space data and information transfer systems -- Audit and certification of trustworthy digital repositories. Dette skal forsøges opnået gennem tæt samarbejde med en partnerorganisation, der beskæftiger sig med samme område, således at Netarkivet og partneren kan hjælpe til gensidigt at sikre institutionernes status som trustworthy repository.

Standarder
Netarkivet anvender i videst muligt omfang internationale standarder i løsningen og evalueringen af bevaringsrelaterede opgaver. Netarkivet fortsætter arbejdet med at implementere følgende standarder:

- ISO 14721: *Reference Model for an Open Archival Information System* (OAIS) som referencemodel i forbindelse med beskrivelse og opbygning af bevaringsløsninger til arkivet.

- ISO 14873: *Information and documentation – Statistics and Quality Indicators for Web Archiving* i forbindelse med beskrivelse og statistik på arkivets indhold. Statistik er bl.a. vigtig i forbindelse med estimering af bevaringsomkostninger og i valget af funktionel bevaringsstrategi.

- ISO 28500: *Information and documentation -- WARC File Format* . I denne sammenhæng tages der stilling til, hvorvidt Netarkivets samlinger i ARC-format skal migreres til WARC.

- ISO 16363:2012 - Space data and information transfer systems -- Audit and certification of trustworthy digital repositories med henblik på at få status som et troværdigt digital arkiv.

- DS/ISO/IEC 27 001 – Informationsteknologi – Sikkerhedsteknikker – Ledelsessystemer for informationssikkerhed – Krav med henblik på informationssikkerhed,.

Netarkivet søger i videst muligt omfang at udvikle og anvende åben, ikke-proprietær software (open source):

- Netarkivet anvender og bidrager aktivt til udviklingen af NetArchive Suite

- Netarkivet bidrager gennem IIPC til udvikling af relevante værktøjer til alle dele af webarkiveringsopgaven, herunder karakterisering og validering af WARC, som indgår i den løbende bevaringsaktivitet.

Dataformater
Netarkivets indhold afspejler de på ethvert indsamlingstidspunkt eksisterende data på internettet. Dette medfører, at alle former for datatyper, formater og versioner indsamles.

Netarkivet tilstræber at overholde følgende generelle retningslinjer:

- Data lagres i originale formater.

- Netarkivet tilstræber at bevare sine samlinger af data i originale formater i pakkeformater, der er egnede til digital bevaring.

- Netarkivet følger i videst muligt omfang internationale standarder på dette område.

- Data lagres i udgangspunktet ukomprimeret. Hvor det af praktiske eller økonomiske grunde besluttes at anvende komprimering af data, anvendes ikke-tabsgivende komprimeringsalgoritmer.Data lagres ukrypteret af hensyn til bevaringssikkerheden. I de få tilfælde hvor adgangssikkerheden vejer tungere end bevaringssikkerheden, tages der stilling til om kryptering skal anvendes. Ved datatransmission af fortrolige samlingsdata krypteres disse, hvor det er muligt.

Lovgivning
Netarkivet udfører bitbevaring af data med udgangspunkt i Ophavsretslovens § 16 (LBK nr. 202 af 27/2/2010), der tillader eksemplarfremstilling (kopiering) af data til bevaringsformål.

Omkostninger
Netarkivet arbejder løbende med at belyse omkostningerne ved digital bevaring for at sikre de nødvendige ressourcer til opgaven.

Hvis der er diskrepans mellem de aktiviteter, Netarkivet vil udføre, og dem, der faktisk kan gennemføres inden for de økonomiske rammer, som er til rådighed, adviserer Netarkivet skriftligt bibliotekernes direktioner herom.

Risikovurdering
Netarkivet er underlagt de risikovurderinger, de to pligtafleveringsinstitutioner gennemfører på deres øvrige digitale samlinger.

I forbindelse hermed vil Netarkivet gennemføre risikovurdering for den samlede driftinstallation. Risikoanalyserne fremlægges skriftligt for bibliotekernes direktioner.

Bitbevaring
Den mest basale form for bevaring af digitalt materiale er *bitbevaring*, en metode hvorigennem det tilstræbes at sikre de oprindeligt indsamlede data mod ødelæggelse. Netarkivet udfører aktiv

bitbevaring på hele arkivet. For at imødekomme de givne politiske retningslinjer samt internationalt anerkendt *best practice* omkring dataredundans, lagres Netarkivets data på følgende måde:

- Dataplacering: 2 replika samt 1 kopi i form af en backup
    - De to replika er *read-only*-arkiver, hvorfra materiale ikke kan slettes. En konsekvens heraf er, at indsamlede dokumenter, der indeholder virus eller anden malware, ikke slettes.
    - Netarkivet udfører ikke virus- eller malwaretjek på det indsamlede materiale. De to replika er onlinekopier i uafhængige miljøer på hhv. Det Kongelige Bibliotek og Statsbiblioteket. Disse miljøer udgøres af software- og hardwaremæssigt forskellige opsætninger og er geografisk og organisatorisk adskilt med ca. 300 kms afstand.
    - Backup-kopien findes på magnetbånd på Statsbiblioteket.
- Integritetscheck:
    - Der udføres løbende integritetstjek for at sikre validiteten af data i de to replika. Integritetstjekket er baseret på samkøring af checksummer på al data fra de to replika.

Netarkivets bitbevaring overvåges automatisk, p.t. af den bitbevaringssoftware, der er inkluderet i NetarchiveSuite.
Al data i Netarkivet skal bitbevares med Det Nationale Bitmagasin software efter de til enhver tid gældende forskrifter og i det antal kopier som er almindelig praksis i de to biblioteker. Der skal fortsat gennemføres regelmæssige integritetstjek af samlingen. Frekvensen fastsættes i Netarkivets årsplan.

Funktionel bevaring
Funktionel bevaring dækker over en række forskellige modeller til permanent sikring af *tilgængeligheden* af arkivets indhold over tid.

Netarkivet har ikke lagt sig fast på en enkelt model, men arbejder løbende med disse to primære typer, der også kan kombineres:

- Migrering. Ved migrering forstås en bevarings- og tilgængeliggørelsesstrategi, der sikrer permanent tilgængelighed af arkivets indhold ved løbende at transformere dette fra oprindelige til aktuelle dataformater.
- Emulering/virtualisering. Denne tilgang tager udgangspunkt i de oprindeligt indsamlede data og tilstræber at stille disse til rådighed ved ad teknisk vej at genskabe de platforme og programafhængigheder, der kræves for at afvikle eller fremvise de oprindelige data. Emulering kan blive særligt nødvendig for Netarkivet, da der her er tale om et usædvanligt stort antal forskellige filformater i en meget kompleks datamodel.

    Emuleringsstrategien fordrer en sekundær indsamling af programmel m.v., som kan sikre afviklingen af de indsamlede data på de emulerede platforme.

I relation til funktionel bevaring af webmateriale søger Netarkivet at sikre, at følgende strategiske mål er opfyldte:

- Skalerbarhed. Alle dele af Netarkivets driftsinstallation skal kunne skaleres, således at bevaringsaktiviteterne i praksis kan gennemføres.

- Deduplikering. Netarkivet deduplikerer visse data. Dette sker primært af hensyn til omkostningerne ved digital lagerplads. Deduplikering må ikke forhindre, at et website kan "samles" til formidling, selvom materialet kan være indsamlet over forskellige høstninger.

- Dataudtræk. Det skal være muligt at udtrække, bearbejde og bevare en delmængde af arkivet og stille denne til rådighed for eksterne brugere.

- Datamining. Det skal være muligt helt eller delvist at stille Netarkivets materiale til rådighed for datamining-projekter.

Netarkivet skal sikre indsamlingen af sekundær software til brug for emulering, evt. gennem internationalt samarbejde.

Metadata og dokumentation
Netarkivet indekseres på følgende parametre:

- URL
- Tidsstempel for indsamling
- Indhold (URL-browsing og fritekstindeks)

Netarkivet dokumenterer arkivmaterialet på en række forskellige planer:

- Kuratorskabt dokumentation af opbygningen af arkivet, herunder dokumentation af indsamlingspraksis (begivenheds- og specialhøstninger mm.), registrering af væsentlige kuratorbeslutninger omkring inkludering og fravalg af materiale.

- Automatisk genereret teknisk dokumentation. Denne dokumentation består bl.a. af generering og lagring af logfiler fra de væsentligste indsamlingsværktøjer, fx Heritrix-logfiler, NAS logfiler, mm. Denne dokumentation lagres sammen med de indsamlede data.Dokumentationen er tilgængelig for alle Netarkivets medarbejdere på Wiki-platforme eller gennem de programmer, der anvendes til indsamlingen af materiale.

Netarkivets manuelle og automatiske dokumentationsprocesser skal løbende revideres i forhold til Netarkivets og dets brugeres behov, samt i forhold til evt. international udvikling, anbefalinger og standarder på området.

Der skal udarbejdes en bevaringsplan for de dele af dokumentationen, der ikke automatisk lagres i WARC-filerne.

I de begrænsede tilfælde, hvor Netarkivet anvender persistente identifikatorer (PID) til dele af samlingerne, skal arkivet løbende vedligeholde disse persistente referencer.

Kvalitetskontrol og Bevaringsplanlægning
Netarkivets kuratorer er ansvarlige for udførelsen af løbende kvalitetskontrol på det materiale, der er indsamlet til arkivet.

Indsamlingsstrategierne for Netarkivet er beskrevet i *Strategi for indsamling af materiale til Netarkivet ved Statsbiblioteket og Det Kongelige Bibliotek.* Procedurerne for kvalitetskontrol af de tre indsamlingstyper er forskellige, grundet de meget store forskelle i omfang mellem fx de selektive høstninger og tværsnitshøstningerne.

Data i Netarkivet indsamles og lagres på URL-niveau i de oprindeligt offentliggjorte dataformater. Data lagres således, at muligheden for at følge links i det indsamlede materiale bevares.

Følgende parametre dokumenteres systematisk for materiale indsamlet gennem de selektive høstninger og ad hoc for de store datamængder i tværsnitsarkiveringerne:

- Funktionalitet af den enkelte webside: Er det muligt at se/afvikle/tilgå tekst, billede, video, audio og interaktive elementer på siden.

- Funktionalitet af arkivets sammenkædning af data: Er det muligt at bevæge sig fra webside til webside i arkivet ved at følge links o. lign. på de indsamlede websider.

- Funktionalitet i relation til *typen* af website: Statiske, dynamiske, sociale, osv. Formålet er at sikre, at der ikke er hele kategorier af websites, der ikke længere kan tilgås.

Udvælgelse af materiale, indsamlingsfrekvens, indsamlingsprocedurer samt dokumentation og kvalitetskontrol af disse findes beskrevet i *Politik* hhv. *Strategi for indsamling af materiale til Netarkivet ved Statsbiblioteket og Det Kongelige Bibliotek*.

Teknologiovervågning og bevaringsplaner
Netarkivet vil gennemføre teknologiovervågning ved at studere internationale watchrapporter og deltage i konferencer og IIPC-aktiviteter på området.

Netarkivet skal tage stilling til, hvorledes arkivet fremover vil arbejde med bevaringsplanlægning.

Ressourcepersoner
Netarkivet udfører sin bevaringsindsats i løbende dialog med de to pligtafleveringsinstitutioners øvrige personale, der beskæftiger sig med digital bevaring. Herigennem sikrer Netarkivet, at de på de respektive institutioner beslutninger og aktuelle tiltag videreføres i bevaringen af arkivet.
Tæt kontakt med relevante indholdsbrugere (forskere mv.) for at få løbende information om samlingens anvendelighed, herunder metadata.

Internationalt samarbejde
Netarkivet vil fortsat indtage en central placering i *International Internet Preservation Consortium* (IIPC) og *Open Planets Foundation* (OPF).

Netarkivet tilstræber at deltage i udviklings- og forskningsprojekter af international relevanssamt at opretholde en tilstedeværelse på konferencer, seminarer, mm.

Litteratur

http://www.digitalpreservation.gov/formats/content/webarch_quality.shtml
http://www.dpconline.org/advice/technology-watch-reports her fra kan man nå
Technology Watch Report 13-01: Web-Archiving [874KB] by Maureen Pennock