Bevaring av digitalt skapt arkiv – metode gitt i OAIS og DIAS

Av Solveig Heløe Olsen, rådgiver Arkiv Troms

Tekst til styreseminar om digitalt skapt arkivmateriale, 07.09.18.

Det ligger flere utfordringer knyttet til langtidsbevaring av elektronisk skapt arkivmateriale, og enkelte risikofaktorer.

Arkiv Troms mottar uttrekk fra sak- og arkivsystemer og fagsystemer fra kommunene for bevaring og langtidslagring.

Først og fremst gjelder utfordringene aspekter rundt lagringssikkerhet, lesbarhet over tid, og opprettholdelse av integritet og autensitet. Arkivmaterialet må bevares med et uendret logisk informasjonsinnhold. Og Arkiv Troms som depotinstitusjon må kunne bevise at informasjonen i uttrekket er uendret.

Langtidsbevaring av digitalt skapt arkiv kan ikke gjennomføres uten fysisk endring av hvordan arkivet framstår. I motsetning til papir vil opprettholdelse av lesbarhet framtvinge transformasjon, og bevaringsmetoden går fra tidligere å være statiske papirdeponeringer til dynamisk endring av produsentspesifikt framstilt digitalt arkiv.

Det er en kjensgjerning at informasjon i elektroniske systemer er enkelt tilgjengelig i produksjonsøyeblikket og mens det befinner seg i sitt opprinnelsessystem, det vil si sak- og arkivsystemet eller fagsystemet. Idet systemet er utdatert og byttes ut er det store muligheter for at informasjon endres, korrumperes eller mistes helt.

Når et arkivuttrekk mottas av oss kan det ha vært gjennom flere tekniske prosesser som har gitt endringer. Deponering av et elektroniske uttrekk kan utgjøre et arkiv som dekker en tidsperiode på 10-15 år. Tidsperioden omfatter gjerne et arkiv som har vært satt opp i NOARK 3 eller 4 struktur, og senere importert inn i et nyinnkjøpt system til en NOARK 5 struktur.

De tre arkivstandardene er ulike, og har en rekke elementer og metadatakrav som ikke samsvarer med hverandre. Informasjon fra den eldre delen av arkivet tvinges inn i sett med datakrav og tabeller som ikke fantes i det opprinnelige gamle systemet, og det er en fare for at innhold endres eller legges inn på feil sted. Når uttrekket mottas hos oss, må vi identifisere slike endringer og teste om strukturene er etter standarden. Og vi må vise at informasjon som framstår med feil oppsett i strukturene ikke har blitt sånn hos oss.

Sak- og arkivsystemene som kommunene kjøper og benytter har en begrenset levetid. De produseres av profesjonelle aktører, og vil av proprietære årsaker inneholde teknisk informasjon som ikke vil være synlig i uttrekk. I dette tilfelle vil det innebære at de ikke vil vise sin kildekode, og vi kan derfor ikke se hvorfor enkelte deler av informasjonen har endret

seg eller lagt seg feil. Vi må forholde oss til arkivet slik det ankommer, og vise at vi ikke har lagt til eller trukket fra informasjon i arkivet slik det framstår for oss.

OAIS-standarden er en modell som nettopp skal sikre at arkivdepoet tar vare på informasjonen/arkivet slik det er mottatt, og kunne vise at det er autentisk. Som arkivdepot skal modellen gi oss et rammeverk som hjelper oss å dokumenterer at vi ikke har endret informasjonsinnholdet i uttrekket vi har mottatt.

Den norske strategien gitt av Riksarkivet, og gjennom lover og forskrifter for bevaring av elektroniske registre og databaser, er migrering. Data flyttes/migreres fra en produsentspesifikk teknologi gjennom konvertering til standardisert lagring. Det opprinnelige systemet eller databasen bevares ikke, kun informasjonen som er trukket ut av basen og en strukturbeskrivelse av det opprinnelige systemet.

Vi tar altså ikke vare på datasystemene som kommunene har benyttet for å føre arkiv, men innholdet i systemene.

Det innebærer en drastisk endring av arkivets framtoning når det flyttes mellom teknologiske plattformer; data "demonteres" og gjøres nøytrale gjennom "rene" sekvensielle tekstfiler.

Valgt filstruktur i Norge er XML-formatet. Struktur- og innholdsbeskrivelsen blir nøkkelen til all gjenskaping og senere bruk av informasjonen. Uten en fullstendig og eksakt beskrivelse av databasestrukturen vil de data som bevares være uhåndterlig og uten praktisk verdi.

XML er et format som kan brukes til deling av strukturerte data mellom informasjonssystemer. XML brukes også til koding av dokumenter og som kommunikasjonsmiddel mellom ulike informasjonssystemer. Filformatet .xml organiserer data i en hierarkisk struktur. Formatet er et vanlig tekstformat, leselig for mennesker, der merker, eller tagger, gir informasjon om hva innholdet er. Kort oppsummert er XML:

- Markeringsspråk egnet for datautveksling
- Lesbart både maskinelt og som tekst
- Organiserer data i hierarkisk struktur
- Logisk verifiserbart og basert på åpne standarder
- Plattformuavhengig
- Benytter Unicode – representerer alle nåværende og kjente historiske tegnsystemer

Migrasjon av datainnhold til et standardformat for å opprettholde lesbarhet gir utfordringer når arkivmaterialet skal benyttes igjen. Det vil være utfordrende for depoet å skulle vise at informasjon i uttrekk er uendret og autentisk.

Arkivstandarden OAIS – *Reference Model for an* ***Open Archival Information System*** (ISO 14721:2003) gir en standard som skal ivareta troverdighet og pålitelighet for det elektroniske arkivet. Standarden gir prinsipper, terminologi og de funksjonelle beskrivelsene som utgjør modellen for integritet gjennom operasjoner i et arkivdepot.

OAIS gir et rammeverk for å innlemme, administrere og benytte arkivmateriale i et arkivdepot. Bevaringsobjektet – det være seg et enkelt dokument eller en hel database – skal framstå som en autonom og selvdokumenterende arkivpakke.

OAIS definerer krav som må implementeres for at et arkiv skal være konformt med standarden. Sentralt i modellen er **bevaring av informasjonsinnhold**. Standarden introduserer begrepet Information Package – **informasjonspakke**. En *informasjonspakke*, slik det benyttes i OAIS, beskriver en konseptuell pakke som inneholder to typer informasjon – Content Information og Preservation Desccription Information (PDI). På norsk kan vi kalle disse henholdsvis ***informasjonsinnhold*** og ***bevaringsmetadata*** (eller bevaringsbeskrivende metadata). Disse to grunnelementene i arkivpakken bindes sammen gjennom OAIS Descriptive Information – Informasjon som beskriver arkivpakkens innhold.

Informasjonsinnhold for en arkivpakke vil typisk være datafilene og tekniske metadata som identifiserer filene.

Bevaringsmetadata er gjerne referanser (eks. UUID), proveniens, kontekst, identifikator og integritetssikring (eks. sjekksum) og rettigheter for bruk og kassasjon.

Et arkivdepot vil i utgangspunktet forholde seg til tre ulike varianter av arkivpakker som identifiseres i OAIS:

SIP – Submission Information Package
Dette er arkivpakken som overføres fra arkivskaper til depot. Den skal omfatte informasjonsinnhold, bevaringsmetadata og en informasjonsfil fra arkivskaper. SIP omtales gjerne som avleveringspakken.

AIP – Archival Information Package
Bevaringspakke. Arkivpakke som bevares i arkivdepot med informasjonsinnhold og metadata. Denne vil forekomme i flere generasjoner pakker.

DIP – Dissemination Information Package
Brukspakke. Arkivpakke klargjort i bruksversjon med mulighet for visning.

Den funksjonelle modellen for OAIS fokuserer først og fremst på oppgaver og forpliktelser med tanke på depotstyring og arkivbehandling. Modellen viser seks funksjonelle enheter som må være til stede i arkivdepoet, og deres relasjoner til hverandre.

Den sentrale funksjonen er administrasjon. Administrasjonsenheten tar del i alle funksjoner i arkivdepoet, og gir strukturen for å opprettholde standarden for arkivet som OAIS krever. Enheten gir krav til, og avtaler deponering av SIP med arkivskaper. I tillegg vil administrasjon delta i alle funksjoner innenfor OAIS.

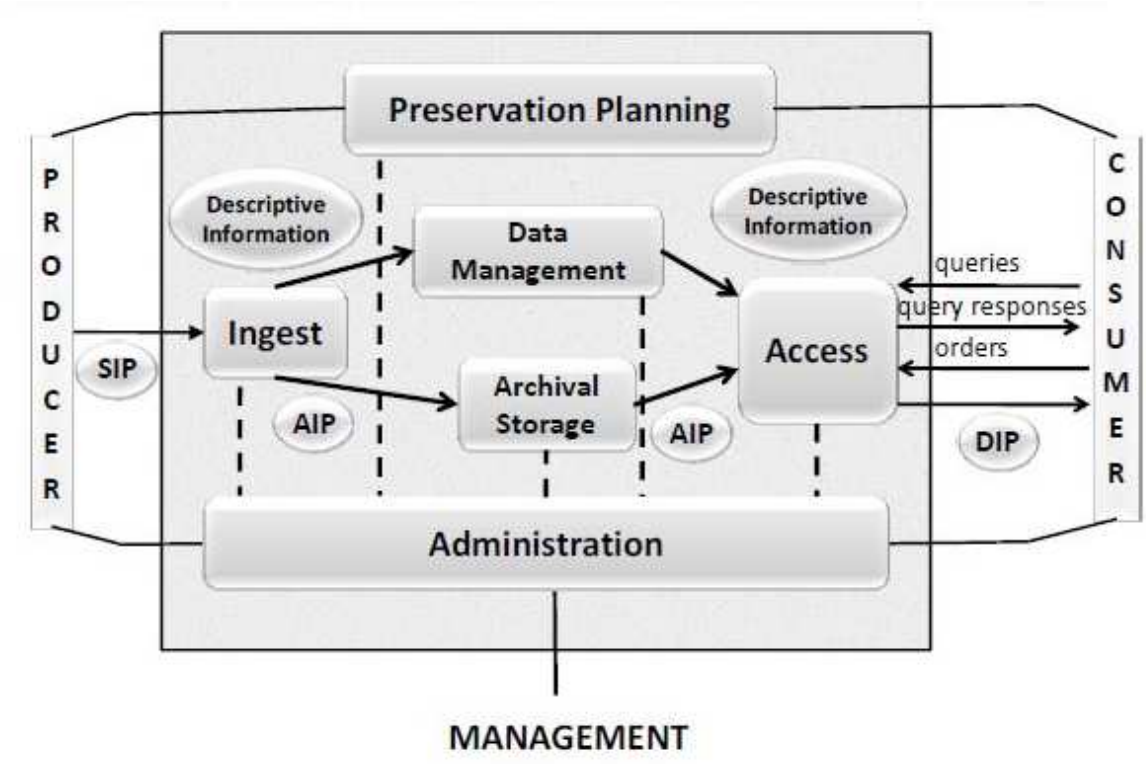
Ingest – mottak og kontroll. Funksjonens oppgave i modellen er å motta SIP, teste denne, og evt. godkjenne deponeringen. Deretter genereres AIP i tråd med godkjent standard på formatene (struktur, dokumenter, foto etc.). Informasjonsinnhold og metadata for bevaring legges til AIP, og overføres til Data Management og Archival Storage.

Archival Storage – funksjonen skal stå for prosessene med langtidslagring, bevaring (eks. bytte av medier, nytt format) og innhente nye generasjoner/tilvekster av AIP til deponeringen.

Data Management – har vedlikehold og dokumentasjon av oppdateringer og endringer av basen som funksjon. Her kontrolleres integritet gjennom listeføring av utførte operasjoner på data og sporbare sjekksummer ved oppdatering av programvare, systemer etc.

Preservation Planning – gir dynamisk planverk for langtidslagring. Her planlegges migrering til nye formater dersom dette kreves pga. lesbarhet eller tilgjengelighet i et langtidsperspektiv. Dette omfatter å evaluere standarder, være oppdatert på lovverk og formater, og påse at beste praksis til enhver tid følges for datauttrekkene. Risikoanalyser og innsikt i endringer av teknologisk miljø er en del av denne funksjonen.

Access – tilgjengelighet. Funksjonen gir opplysninger om at arkivpakker eksisterer, beskrivelse av innhold, hvor man finner opplysningene og hvorvidt de er tilgjengelig.



(Illustrasjon hentet fra Referanse Model For Open Archival Information System, ISO-14721)

OAIS standarden går i dybden på de ulike prosessene som arkivdepoet skal følge for å sikre autensitet og pålitelighet for innholdet i arkivene. Modellen gir detaljerte beskrivelser av oppgavene for de ulike samhandlende funksjonene i OAIS, og de prosesser som må utføres av hver enkelt funksjon.

Modellen setter en rekke krav til selve arkivdepoet:

- Skal sette krav til og akseptere SIP

- Skaffe tilstrekkelig kontroll over informasjonsinnholdet til å sikre langtidslagring
- Identifisere hvem som skal benytte informasjonsinnholdet, og gjøre det tilgjengelig for disse
- Påse at informasjonsinnholdet er forståelig for de som ønsker å benytte det
- Påse at informasjonsinnholdet er sikret og ivaretatt
- Gi troverdige og pålitelige "kopier" av opprinnelig informasjon
- Vise til en prosess som sikrer informasjonsinnholdet som troverdig
- Opptre i tråd med gjeldende lover og reguleringer
- Ha fullmakt til å inngå nødvendige samarbeidsavtaler for å sikre informasjonsinnholdet, f. eks. benytte KDRS sine tjenester som digitalt depot
- Sikre at informasjonsinnholdet kan benyttes og forstås over tid, uavhengig av brukergruppe
- Ha rutiner og strategi for bevaring av arkivpakker for framtiden

OAIS er en referanse- og begrepsmodell med metadata-kategorier. Implementering i depot og konkrete forslag gis i oppfølgingsstandarder. TRAC er en slik standard, utviklet av det amerikanske National Archives og The Research Libraries Group.

Da OAIS først ble en standard i 2002, begynte flere amerikanske kulturhistoriske institusjoner å erklære at de fulgte OAIS og dermed hadde økt troverdighet og høyere pålitelighet enn andre. Likevel fantes det ingen etablert forståelse av hvilke konkrete tiltak dette skulle innebære for depoet. Det var ingen måte å beregne i hvilken grad man fulgte OAIS standarden eller ikke.

TRAC – Trustworthy Respository Audit and Certification – gir en sjekkliste på 90 konkrete kriterier som må oppfylles dersom et depot skal anses som troverdig. Sjekklisten er delt i tre seksjoner:

- Organisasjonens infrastruktur
- Administrasjon av digitale objekter
- Teknologi, teknisk infrastruktur og sikkerhet

Sjekklisten gjør depoets troverdighet om til en målbar størrelse, og kan definere om et depoet følger TRAC og OAIS i større eller mindre grad.

I hovedsak omhandler punktene at depoet driver forebyggende vedlikehold at arkivpakkene, beskytter arkivene mot uautorisert tilgang, dokumenterer rutiner og loggfører alle operasjoner som gjøres på uttrekkene. Alle prosesser som utføres må loggføres og dokumenteres, slik at tilbakeføring til tidligere versjoner hele tiden lar seg gjennomføre.

OAIS gir en terminologi og begrepsavklaring med tanke på de oppgaver som ligger til et elektronisk depot. Denne terminologien er tatt opp i TRAC, og har vært med på å etablere begrepene SIP, AIP, DIP, Ingest etc. i vokabularet til depotinstitusjonene.

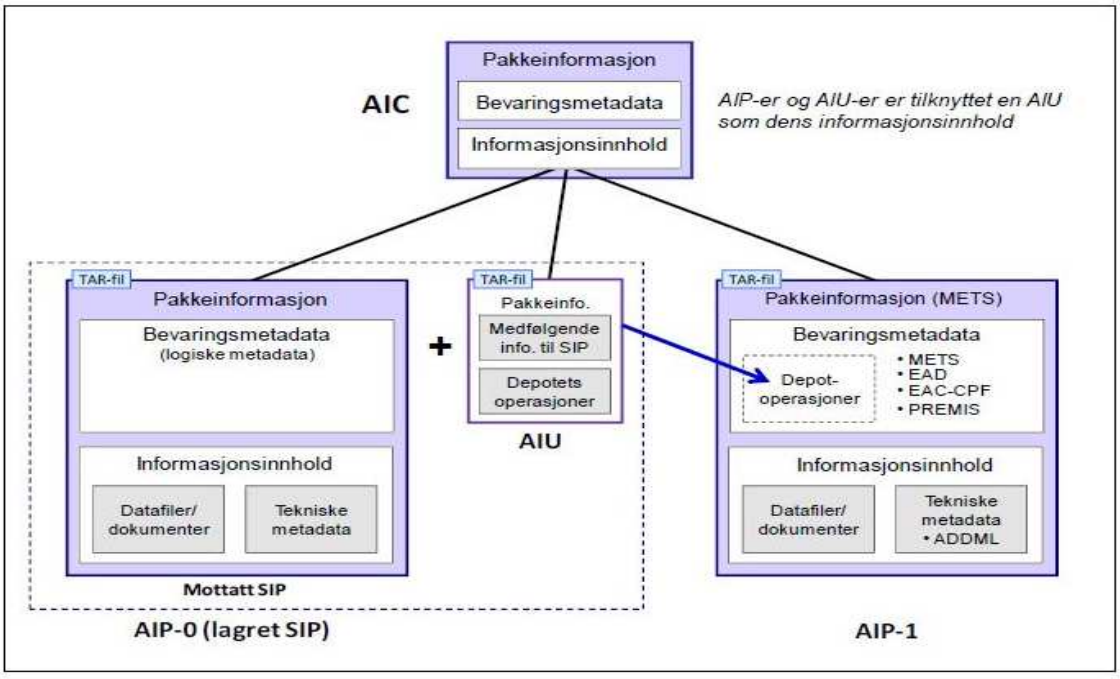
TDR – Trusted Digital Repositories ble publisert i 2011, og ble ISO-standard 16363 Audit and Certifiaction of Trustworty Digital Repositories i 2012. Standarden definerer en praksis for depot som gir en målbar gradering av troverdighet. Standarden baserer seg på kriteriene gitt i TRAC, og gir grunnlag for sertifisering av depot.

Sentralt i TRAC står depoets rutiner for behandling av elektronisk arkiv. Depoet må kunne vise at bevart informasjonsinnhold fortsatt samsvarer med opprinnelig mottatt innhold. Alle operasjoner som innebærer transformasjon av arkivpakker, må være ettersporbare. Depoet må kunne vise at informasjonsinnhold er bevart uendret fra og med mottak for å framstå som troverdig og pålitelig. Det garanterer ikke at innholdet er autentisk i seg selv, men det garanterer at depoet ikke har endret det.

TRAC ble utviklet med tanke på at arkivinstitusjonene blir pålagt en slags omvendt bevisbyrde. Standarden kom bl. annet som en konsekvens av O. J. Simpson-saken i 1995, hvor fingeravtrykk ble underkjent av retten kun fordi behandlingsrutinene i politietaten kunne gitt politiet *mulighet* for manipulasjon. Konklusjonen er at den som forvalter informasjon må selv kunne bekrefte at informasjonen er autentisk og ikke tuklet med. Tanken er at arkivdepot skal selv kunne produsere dokumentasjon som verifiserer ekthet, og fjerne grunnlaget for tvil eller spekulasjoner om arkivets troverdighet.

Metoden for å implementere OAIS-standarden i Norge ble gitt gjennom DIAS-prosjektet (Digital arkivpakkestruktur) publisert av Riksarkivaren i 2012. DIAS har spesifisert en arkivpakke- og prosessmodell som en definert måte å bruke OAIS. TRAC-kravene er et sentralt element i DIAS, og spesifikasjonene er satt etter seksjonen av TRAC som omhandler Administrasjon av digitale objekter.

Denne DIAS-prosessmodellen er illustrert i figuren nedenfor.



Figuren viser alternativet hvor en mottatt SIP ble lagret som AIP-0 uten samtidig generering av en DIAS-organisert AIP-1. I dette tilfellet kreves en AIU i tillegg for å unngå noen form for endring av AIP-0 og dens samlede pakkesjekksum. Ved en senere generering av AIP-1 inn-arbeides AIU i denne. Behovet for en AIU bortfaller når AIP-0 og AIP-1 genereres samtidig. Når/hvis metadata senere endres for AIP-1, kan de på tilsvarende måte lagres i en tilknyttet AIU – og da for å bevare AIP-1 uendret, og unngå eller utsette generering av en full AIP-2.

(Illustrasjon hentet fra DIAS sluttrapport, Riksarkivaren 2012)

DIAS arkivpakker gir en overordnet organisering med implementering av OAIS modell gjennom følgende standarder (internasjonale standarder forkortet med):

- METS – beskriver arkivpakkens indre struktur og beholderen som omslutter den. Dette spesifiserer de overordnede elementene i arkivpakken: organisasjon, struktur og innholdsoversikt.
- PREMIS – for bevaringsmetadata og annen forståelsesinformasjon
- EAD – for logisk arkivbeskrivelse
- EAC-SPF – for aktørbeskrivelse (kommunen som arkivskaper)
- ADDML – for fil- og postbeskrivelse (tekniske metadata)

Hver av disse har definerte XML-skjemaer basert på standardene.

DIAS-modellen omfatter 5 ulike informasjonspakker/arkivpakker:

- SIP – leveringspakke
- AIP – bevaringspakke
- AIC – samlepakke (inneholder flere generasjoner av en AIP)
- AIU – tilleggspakke til AIP, inneholder metadata
- DIP – brukspakke

DIAS gir en standard bygget på OAIS, men også en prosessmodell for å benytte standarden til å sikre integriteten til uttrekket. Når en levering av et uttrekk – en SIP – kommer til arkivdepot skal den bevares uendret – for alltid.

SIP som kommer til depot skal være pakket som en TAR-fil. Med SIP følger det en informasjonsfil om uttrekket fra kommunen. Her skal det stå informasjon om selve uttrekket og systemet det er hentet fra, samt en sjekksum på TAR-fila. Denne sjekksummen holder integriteten for uttrekket; den verifiserer at uttrekket sendt fra kommunen er det samme som uttrekket vi bevarer. SIP blir AIP generasjon 0.

Det er ikke pålagt fra Riksarkivets side å følge OAIS standard gjennom DIAS prosessmodell for offentlige norske arkivinstitusjoner, men det er anbefalt. Som vi ser det, vil det være sterkt ønskelig at modellen er implementert i vårt elektroniske depot.

Når Arkiv Troms begynner å teste SIP med loggføring av sjekksum, innhold, evt. migrering til XML, etc. vil dette dokumenteres utenfor AIP-0. Denne tilleggsinformasjonen lagres enten i AIU tilknyttet via AIC eller i en ny generasjon av AIP-1. I AIP-1 vil beskrivelsene i METS, PREMIS, EAD og EAC-CPF tilkomme. Senere kan det bli nødvendig å lage flere generasjoner av AIP/bevaringspakken.

DIAS gir et oppsett for at operasjonene i arkivdepoet blir logget gjennom PREMIS Events, Rights og Agents. Her er det gitt hendelser for mottak, innsjekking til depotsystemet, testing av SIP, skape AIP/AIC, vedlikeholde arkivpakken og skape DIP.

Forvaltningen av DIAS-arkivpakker skal gjøres gjennom KDRS depot. Her benyttes systemet ESSArch fra ES Solutions AB. ESSArch er tilpasset DIAS og oppfyller samtlige punkter i standarden, samt genererer og vedlikeholder digitale arkivpakker med integritetssikring.

Arkiv Troms har ikke tatt i bruk ESSArch og KDRS depot per i dag. I år har vi satt opp 2 arbeidsstasjoner i sikker sone, og satser på at vi i løpet av neste år begynner å overføre digitale deponeringer til sikringsdepot.

OAIS og DIAS definerer forpliktelser og oppgaver i depotstyringen som skal bevare autensitet og integritet for det overførte elektroniske arkivet. Dette forutsetter at SIP – arkivpakken som overføres fra kommunen – møter de kriterier som ligger til grunn for en arkivpakke. Per i dag har vi mottatt få uttrekk som følger formatkrav i lovverk og standarder, noe som gjør at vi ikke har prioritert arbeid med arkivpakker til digitalt depot.

I praksis vil dette bety for oss at vi mottar et uttrekk/SIP fra en kommune som inneholder:

- uttrekksfiler/dump/backup avtalt i henhold til hvilket system og Noark-standard som skal deponeres – pakkes i TAR-format
- info-fil om uttrekket med sjekksum som integritetsbærer
- håndbøker eller annen informasjon om hvordan programmet har vært brukt

Arkivdepoet må kunne bevise og demonstrere at bevart uttrekk/informasjonsinnhold er i samsvar med opprinnelig mottatt innhold. De tekniske operasjoner som foretas i depot må loggføres og være ettersporbare tilbake til den originale arkivpakken. I praksis vil dette legge restriksjoner på de operasjoner depoet kan utføre på uttrekket – informasjonsinnholdet skal være i samsvar med mottatt innhold. Dump fra baser må migreres og samtidig transformeres til fastsatt XML-format, men det vil være uheldig om arkivdepoet endrer på innhold i uttrekket, ordner, rydder, retter opp feil etc.

I Riksarkivarens brev av 15.05.09 vedr. distribusjon av testverktøyet ArkN4 påpekes det at manipulering av uttrekket kan oppfattes som dokumentfalsk og rammes av straffeloven. Retningslinjene for bruk av verktøyet sier at det ikke er greit å lage korrigeringer i selve uttrekket i den hensikt å få dette godkjent. Det skal være fullt samsvar mellom uttrekket til depot og opphavsbasen på uttrekkstidspunktet.

Feil og avvik i uttrekket vil bli avdekket gjennom testing i depot, og vi vil be arkivskaper/kommunen se om feilene også ligger i basen og helst forklare hvordan de har oppstått. Arkivskaper kan rette opp feilene i basen og deretter utføre et nytt uttrekk, eller dokumentere feilene slik de framstår i opphavsbasen.

I denne prosessen er det særdeles viktig at kommunene tar eierskap over innholdet i uttrekket, og gjør bevisste valg med hensyn til kvaliteten på informasjonsinnholdet som skal bevares. Kvaliteten på informasjonen som hentes fra uttrekk vil stå i direkte relasjon til de ressurser arkivskaper er villig til å bruke på kontroll av sak- og arkivsystemet før uttrekk.

Kommunene har mulighet til å deponere et uttrekk med feil som en del av innholdet, så lenge de er innforstått med at feilene vil bevares med uttrekket for framtiden. Depoet kan bevare uttrekket med avvik så lenge de er dokumentert av arkivskaper som en del en den historiske basen.

Dersom avvik i uttrekket ikke samsvarer med opphavsbasen og framstår som mangelfullt, er det trolig feil i uttrekksprosessen. Da må uttrekksverktøy korrigeres før nytt uttrekk produserer og oversendes depot.

De erfaringer vi har gjort til nå har resultert i at vi ønsker et databasedump av uttrekksbasen, uavsett hvilken NOARK-standard uttrekket produseres fra. Databasedumpet vil gjøre det mulig for oss å sammenligne uttrekket med NOARK-versjonen av innholdet, og dermed være i stand til å rådgi kommunen hvorvidt feil og avvik skyldes innføringer og bruk av basen, eller om de har oppstått i uttrekksprosessen. Å skille disse to årsaksforholdene vil være avgjørende for hvordan vi og kommunen kan følge opp kvaliteten på uttrekket.

Så langt vi kan trekke erfaringer i dag, ser vi at både verktøyssiden for testing av uttrekk, men også selve uttrekkene vi mottar fra kommunene, har stort forbedringspotensial.

For å benytte KDRS digitalt depot og innlemme uttrekk fra kommunene, må vi godkjenne uttrekkene i henhold til NOARK-standardene de er levert i, og i henhold til krav til format som gitt i lover og forskrifter.

Ved deponeringer opplever vi at format på dokumenter ikke er i henhold til lover og forskrifter, men også at det «fylles inn» informasjonsinnhold i uttrekkene. Det siste skjer sannsynligvis for at XML-filene som deponeres skal validere mot de ulike standardene sine skjemaer for innhold i strukturene. Eksempler på dette er at arkivmapper ser ut til å være opprettet i år 1, at endringer i føringer er utført i 1899, at tid oppgis i hundretusendels sekunder etc. Da kan man anta at script er kjørt fra uttrekksleverandørens side uten hensyn til de konsekvenser dette får for arkivet, eller troverdigheten til innholdet. Slike endringer opplyses det sjelden om når uttrekket deponeres, og denne type «feil» skaper mye merarbeid i depot.

OAIS og DIAS gir rammer for hvordan et depot skal opptre og ivareta elektroniske arkiver. Lister over hendelser som skal dokumenteres i depot, og oppsett av et forvaltningssystem for arkivuttrekk gjennom modellene gir et godt grunnlag for gjennomføring av troverdig arkivbehandling i vårt depot.