Notat
2020-05-29
Utarbeidet i produktteam mottak og bevaring, Arkivverket

Klargjøring rundt utpakking av arkivuttrekk i Mottak

Bakgrunn

Per i dag har vi i så stor grad som mulig forsøkt å la arkivuttrekkene ligge inne i sine tar-pakker når vi behandler dem. Dette har gjort behandlingen av digitalt skapt materiale analog med behandlingen av papir-baserte uttrekk. Samtidig er det en enkel modell som det er lett å resonnere rundt. Underveis i utviklingen viste det seg at to av produktene vi lener oss på, Arkade og DECOM, ikke greier å operere på pakkede arkiv, men trenger tilgang til utpakkede arkiv. I den forbindelse har produktteamet sett nærmere på ideen om å behandle og bevare utpakkede arkiver. Her følger to konkrete forslag for hvordan vi kan behandle arkivuttrekk i utpakket form. Prosjektet anbefaler strategi 1.

I dokumentet er det lagt til grunn at inkrementelle uttrekk er i bruk. Hver generasjon av arkivpakken vil kun inneholde tillegg fra den foregående generasjonen, samt metadata. Disse vil kunne kombineres til et helhetlig arkivuttrekk, eller benyttes hver for seg. Dette dokumenter bygger videre på denne ideen, men er ingen forutsetning for å bevare et uttrekk utpakket. Inkrementelle uttrekk er behandlet i et eget notat.

Strategi 1: Behandle og bevare uttrekk i utpakket form

I dag og i den nærmeste tiden vil uttrekk bli overført til Arkivverket i form av DIAS-pakker (tar-filer). Forslaget her er at vi i tar imot et uttrekk, sjekksum-kontrollerer det, kontrollerer det for virus og pakker det ut i en mappe eller i et objektlager. Utover i dokumentet brukes objektlager, men dette kan ansees som synonymt med mappe.
Objektlageret med uttrekket i bør være dedikert til dette uttrekket. Objektlageret navngis på bakgrunn av UUID med «.0» tilføyd til slutten for å indikere at dette er generasjon 0 av uttrekket. Når uttrekket er pakket ut og vi er sikre på at dette har skjedd uten tap av informasjon så gjøres dette objektlageret i sin helhet *uforanderlig* og DIAS-pakken slettes. For å sikre integritet lagres sjekksum av indeks-filen (dias-mets.xml) eksternt. Her vil vi trenge å gjøre endringer i Arkade slik at Arkade sender over denne sjekksummen utenfor tar-filen sammen med sjekksummen til tar-filen.

Så lages det et objektlager til med «.1» tilføyd til UUID. Her vil alle endringer i uttrekket ende opp og dette skal bli generasjon 1. Evt. resultatet av konvertering og beriking av innholdet i uttrekket havner her med tilhørende oppdaterte metadata (for eksempel testrapport etc.).

Når mottak anser seg som ferdig med uttrekket, sendes det melding til Bevaring om at prosessering er ferdig og uttrekket kan overføres til bevaring.

Hvis Bevaring aksepterer forespørselen, overføres eierskapet til disse to objektlagrene til bevaring. Mottak mister tilgang til og kontroll over uttrekket. Evt. endringer, som

formatkonvertering eller tilføyelse av metadata må skje ved at det opprettes et nytt objektlager med et løpenummer på slutten som indikerer generasjon.

Ivaretakelse av tilgjengelighet og integritet vil være helt essensielt for å kunne vurdere om man kan ha tillit til arkivmaterialet. Det vil ikke være like enkelt som det er med en enkelt DIAS-pakke / tar-fil, hvor en enkelt sjekksum vil gi sterke garantier om at innholdet er uendret.

I korte trekk vil ivaretakelse av integritet skje på følgende måte:

- I hver generasjon av arkivpakken finnes det en indeks over alle filer i uttrekket med sjekksummer for hver eneste fil – dias-mets.xml. For andre formater enn DIAS, for eksempel BagIt, finnes tilsvarende filer som også er basert på METS.
- Ved å ta sjekksummen på denne filen og oppbevare den utenfor arkivpakken kan vi enkelt verifisere at denne filen ikke har vært tuklet med. Og siden vi vet at denne filen har sin integritet i orden kan vi bruke innholdet i den til å verifisere resten av arkivet og verifisere sjekksum for samtlige elementer i uttrekket.

Fordeler og ulemper

- En fordel ved å prosessere utpakkede arkiver er at vi effektivt unngår 4.7TB-grensen i objektlagrene som ellers vil være krevende å omgå. Det vil fortsatt være noen utfordringer rundt det å ta imot uttrekk større enn 4.7TB – men hvis uttrekkene oppbevares utpakket, er denne problemstillingen begrenset til opplastningskomponenten i mottak og vi trenger ikke ta hensyn til dette i øvrig infrastruktur.
- En annen fordel er at vi kan, om vi ønsker, la infrastrukturen gjøre tilgangskontroll for oss. Om vi ønsker å gi en person tilgang til deler av et uttrekk så støtter objektlagrene dette.
- En fordel ved utpakking er at det er enkelt og forståelig. En kan hente ut en fil fra et uttrekk uten spesielle verktøy. Det vil fortsatt være hensiktsmessig med verktøy for raskt å hente ut filer fra uttrekkene, men disse vil være enkle.
- En klar fordel er at alt vil gå vesentlig raskere når arkivet er allerede pakket ut. En kan hente en fil på et øyeblikk. For en DIAS-pakke må hele arkivet lastes ut og leses gjennom for å finne en fil.
- Kostnadsbildet vil være forskjellig for arkiver over 4.7TB. For å lagre enkeltfiler over 4.7TB må vi bruke lagringssystemer som koster vesentlig (~71 ganger) mer. For utpakkede arkiv vil vi kunne bruke langt billigere lagring.
- En klar ulempe er at dette bryter med eksisterende praksis. Dermed har vi naturlig nok mindre erfaring rundt detaljene.
- En annen ulempe er at det er ingen (?) andre i sektoren som gjør dette. Dermed blir vi sittende med prosesser og verktøy alene og kan ikke lene oss på predefinerte komponenter i sektoren.
- En ulempe ved utpakkede uttrekk er at integritet blir mer komplisert å ivareta og dokumentere.
- Ved å oppbevare uttrekkene utpakket svekkes avhengigheten mot DIAS-pakker. Om Arkivverket ønsker å støtte lagring av data i andre type «kapsler» som f.eks. BAGIT eller E-ARK, så er det relativt enkelt å implementere.

- For tilgjengeliggjøring vil utpakkede arkiver være vesentlig enklere å forholde seg til. Tilgang kan gis direkte til deler av uttrekkene og kan skje på objekt-nivå og ikke bare mot hele uttrekk.

Det er verdt å merke seg at dette ikke skaper bindinger mot spesifikk infrastruktur. Dette vil kunne kjøres på lokal infrastruktur så vel som sky-basert infrastruktur.

Strategi 2: Prosessering av utpakket arkiv i mottak med pakking før overføring til Bevaring

Som et alternativ til strategi 1 kan en mindre omfattende løsning være at den opprinnelige DIAS-pakken (tar-fil) beholdes, men at mottak lager en inkrementell DIAS-pakke før overføring til Bevaring.

Bevaring vil da få oversendt to (inkrementelle) DIAS-pakker som til sammen utgjør et uttrekk. Mottak vil da jobbe på et utpakket arkiv, men arkivet pakkes før det overføres.